Deliverable 5.1: A consistent characterisation of existing and planned RIs

WORK PACKAGE 5 – Reference-model guided RI design

LEADING BENEFICIARY: UNIVERSITY OF EDINBURGH

Author(s):	Beneficiary/Institution
Malcolm Atkinson	University of Edinburgh
Alex Hardisty	Cardiff University
Rosa Filgueira	University of Edinburgh
Cristina Alexandru	University of Edinburgh
Alex Vermeulen	Lund University
Keith Jeffery	British Geological Survey (BGS)
Thomas Loubrieu	L'Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER)
Leonardo Candela	Consiglio Nazionale delle Ricerche (CNR)
Barbara Magagna	Umweltbundesamt GMBH (EAA)
Paul Martin	Universiteit van Amsterdam (UvA)
Yin Chen	EGI
Margareta Hellström	Lund University

Accepted by: Paola Grosso (WP 5 leader)

Deliverable type: REPORT and Wiki

Dissemination level: PUBLIC

Deliverable due date: 30.April.2016/M12

Actual Date of Submission: 30.April.2016/M1

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

ABSTRACT

The preceding FP7 funded ENVRI project did an analysis of the characteristics and requirements of environmental research infrastructures (RIs) by comparing some of these with a common reference model. The outcomes proved to be helpful for the understanding of strengths and weaknesses in the outline and planned developments of the RIs. The current ENVRIplus project has a more ambitious programme and it was felt that the analysis should be updated and expanded.

This report refreshes and revises the information about the *Environmental* Research Infrastructures (RIs), primarily those engaged in ENVRIplus, and available technologies in order to clarify requirements, identify issues and highlight opportunities. The main subjects in this report were selected by the RIs themselves. Nevertheless, the team involved in this product noticed additional common requirements of priority and it was decided to also take these up in supporting the RI developments. All findings and recommendations will be used within the ENVRIplus project to inform the subsequent work. Research developing the required information has helped develop a vital channel of communication between computing specialists with application specialists and strategists. The report is divided into three main parts:

1. the results of systematic requirements gathering (Section 2, page 19 onwards),
2. the integration of a broad technology review (Section 3, page 59 onwards), and
3. an assessment of their quality and their characterisation (Section 4, page 126 onwards)., including implications shaping future actions (Section 5, page 187 onwards).

A collation of possible impacts on the ENVRIplus project and on its participating organisations is presented in Section 5, Page 187. As this is a long document, forming a compendium of work, *a map to help readers find the parts that interest them is provided – Figure 1 on page 15.*

This deliverable document is meant for two purposes:

First of all, it is a description for the stakeholders, as an effective route by which to pass the new information collected to the user communities. The aim is to develop and share an agreed viewpoint on the Research Infrastructure researcher-user requirements, the RI asset offerings and the available technology now and in the near and further future. The document is primarily for the RIs participating on ENVRIplus and their communities, but it should also be helpful to other RIs delivering similar services in any scientific or application domain.

A second important factor is that it is a contribution to an ENVRIplus project review.

This work is undertaken as a Task 5.1 in Work package (WP) 5, which itself is part of a closely related group of work packages forming Theme 2. This theme is concerned with the design, development and implementation of e-Infrastructure, methods, services and tools, that will help RIs more easily manage and fully exploit their data. This report should help Theme 2 integrate and steer its work to meet the priorities of the Research Infrastructures.

Project internal reviewer(s):

Project internal reviewer(s):	Beneficiary/Institution
Jean-Daniel Paris	Commissariat a l'energie atomique et aux energies alternatives (CEA)
Wouter Los	Universiteit van Amsterdam (UvA)

Document history:

Date	Version
11.4.2016	Draft for comments
26.4.2016	Corrected version
27.4.2016	Accepted by Paola Grosso

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the editors (Author Malcolm Atkinson Malcolm.Atkinson@ed.ac.uk, Alex Hardisty HardistyAR@cardiff.ac.uk, Rosa Filgueira rosa.filgueira@ed.ac.uk, or one of the authors listed above.)

TERMINOLOGY

A complete project glossary is provided online here: 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 coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilization between RIs, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environmental understanding and decision-making 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 harmonization and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonize policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance trans-disciplinary 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 inter-RI (European and Global) 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.

Blank page

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

Executive summary

This document is a compendium of work bringing together work from multiple viewpoints in an attempt to capture and understand the needs of the environmental research infrastructures and possible approaches to meeting those needs. As such it is a long and complex document. To help readers navigate to the parts that interest them a map is provided – see Figure 1 on page 15.

The preceding FP7 funded ENVRI project did an analysis of the characteristics and requirements of environmental research infrastructures (RIs) by comparing some of these with a common reference model. The outcomes proved to be helpful for the understanding of strengths and weaknesses in the outline and planned developments of the RIs. The current ENVRIplus project has a more ambitious programme and it was felt that the analysis should be updated and expanded. This work was undertaken by Task 5.1 described in the DoW as follows:

> *Re-analyse the status of involved RIs in ENVRI[PLUS] along the dimensions of data, users, software services and resources in order to update the requirement study performed in the early phase of ENVRI. Together with interoperability requirements (based on use-cases in WP6-8) and the review of data and computing infrastructure such as EGI, Helix Nebula and EUDAT such analysis will point to: (a) commonalities between RIs; (b) differences between RIs; (c) interoperability between RIs; and (d) the state-of-the-art of RI technologies. The characterisation of RIs under a common documentation method which may employ vocabulary defined in existing ENVRI RM allows comparison and discussion leading to best practice and consistent development plans for RI improvement and also RI interoperation. This task will take actions to:*
>
> *a) Update requirements from all involved RIs;*
>
> *b) Define common documentation methods for describing the current status of RIs; this should include any data management issues that affect the RI internally, or affect interoperation.*
>
> *c) Perform a consistent characterisation of existing and planned RIs, and their user requirements (within their principal community and in interoperation with other RIs);*
>
> *d) Review the state-of-the-art of technologies provided by data and computing infrastructures;*
>
> *e) Recommend suitable design and engineering approaches for common operations between RI projects by maximally reusing existing industrial standards and existing tools.*

The information collected from and expressed by the Research Infrastructures (RIs) participating in ENVRIplus shows, after analysis, that there are common issues and technological opportunities that were anticipated when ENVRIplus was proposed. Nevertheless, the team involved in this report noticed additional common requirements of priority and it was decided to also take these up in supporting the RI developments. The primary examples are (see Section 2 on page 19 onwards for complete coverage and details):

- The need to achieve data harmonisation, i.e., consistency of representation, interpretation and access, both within and between RIs.
- The need to learn from one another and pool efforts in order to accelerate and harmonise delivery of data services and working practices that support well each stage of the scientific data lifecycle from data acquisition to delivery of actionable derived information.
- Help with facing the challenge of sustainably delivering data services immediately to meet current RI priorities while taking into account longer-term issues and technology trends.

However, care must be taken not to overestimate the pervasiveness of these similarities; for example:

- Differences in maturity lead to substantially different priorities, e.g., many RIs currently face setting up internal collaborative support for the early stages of data acquisition, whereas a long-established RI, such as EuroARGO, prioritises improved access to existing data products, and
- Differences in the internal diversity and prevailing collaborative arrangements between RIs, e.g., EPOS incorporates more than 600 independent organisations with different priorities in a broad spectrum of geosciences and practices, whereas EISCAT-3D has a comparatively small number of participants all focused on studying the upper atmosphere,

Such differences lead to significant differences in working practices and related requirements. However, there is near universal agreement that the key performance indicator used by RIs is researcher productivity. Hence, ENVRIplus focuses on removing inconsistencies and impediments from researchers working environment as data wrangling can consume large proportions of a researcher's time. It only rarely requires insight from domain experts, and so can be eliminated by appropriate automation.

Prior infrastructure investments, particularly in the more mature RIs have to be considered. These are not just the capital investment in equipment, software and services. They are also the training and development of working practices that become manifest in cultures and collaborative arrangements that have widespread, often global, community support as well as long-term and substantial value.

The requirements gathering was organised in terms of:

- general information gathering (see Section 2.2 starting on page 24) and

the primary topics of Theme 2[1], which are:

- *Identification and citation* (Section 2.3.1 page 46),
- *Curation* (Section 2.3.2 page 47),
- *Cataloguing* (Section 2.3.3 page 48),
- *Processing* (Section 2.3.4 page 49),
- *Provenance* (Section 2.3.5 page 52),
- *Optimisation* (Section 2.3.6 page 53) and
- *Community support* (Section 2.3.7 page 57).

Each of these topics contains information from RIs and then an analysis that collects, collates and interprets the gathered information. The majority of the details are held in a wiki[2]. The achievements and limitations of the requirements gathering are assessed in Section 4.1 page 126 onwards. To a large extent the gathered requirements match the expectations when ENVRIplus was planned. Examples of the kinds of extra requirement that emerged are given in Section 4.1.9 pages 135 onwards. These are predominantly about simplified packaging and early exemplars of functionalities that are already being addressed.

The technology review updates the understanding of the technologies that are pertinent to Theme 2. It will inform future work in Theme 2 and help those steering RIs make technical decisions. The review was conducted over a relatively short period. Thus we drew on existing knowledge and understanding within the project and updated our assessment of technology primarily by considering authoritative or active information resources, such as relevant groups in standardisation organisations, e.g., Research Data Alliance (RDA)[3], Open Geospatial Consortium (OGC)[4] and Worldwide Web Consortium (W3C)[5]. We collaborated with EUDAT to share

[1] These topics are the result of three-years analysis in the predecessor ENVRI project and a formalisation of the distributed architectural structure, which can be found in the wiki space, http://envri.eu/rm.
[2] https://wiki.envri.eu/display/EC/ENVRI+RI+Requirements
[3] http://rd-alliance.org/
[4] http://www.opengeospatial.org/

technology review information and with RIs using the technologies. The technology review was organised in terms of the six pillars underpinning Theme 2 work plan – see Figure 4 on page 60 and four cross-cutting aspects of technology that need to influence every pillar:

1. Introduction and explanation of the *technology review methodology* used – Section 3.1 on pages 59 onwards. This introduces the technology reporting structure. Each topic was led by a topic leader who identified the topic, the information sources and referenced material. They developed an analysis pertinent to the two-to-four-year horizon that is relevant within the ENVRIplus project's lifetime. They also developed a longer-term view by assessing the trends in their area. This is relevant for those making strategic investment and planning decisions in RIs.

Review topics – the six pillars of technological focus

2. The review of technology pertaining to *data identification and citation* – Section 3.2 pages 62 onwards. Here the challenge of minting and using reliable references to data is analysed and potential solutions are concisely compared. Making these mechanisms precise and widely adopted is key to data curation, cataloguing, provenance and optimised workflow processing. It is also key to proper acknowledgement of data creators and data-publishing institutions.
3. The review of technology pertaining to data *curation* – Section 3.3 pages 73 onwards. Preserving all relevant artefacts in today's data-driven science is challenging. To do this in ways that encourage interchange and cross-domain use of data is even more challenging. However, such activities are essential to achieve quality and reliability in the results produced and to successfully address today's intellectual and societal challenges. Potential strategies for building on existing platforms and standards, while accommodating diversity are carefully explored.
4. The review of technology pertaining to *cataloguing* – Section 3.4 pages 76 onwards. With today's rapidly growing diversity and wealth of data, finding data quickly and interpreting it correctly is crucial. The catalogues underpin this capability, which is used intensively by researchers and the software they employ. The review draws on deep experience and existing global campaigns to lay out the options for the ENVRIplus communities.
5. The review of technology pertaining to *processing* – Section 3.5 pages 82 onwards. Processing, data storage and data transport underpin every other activity in the data lifecycle from data acquisition or production to curation and publishing. It is a mature field of great diversity with well supported general solutions and a growing number of specialised technologies to deliver capabilities specific to certain categories of data and algorithms. A map of this complex field is presented that includes continuity of many established tools and working practices, but also reviews the emerging capabilities of data-intensive platforms and workflows needed for scientific and data-management methods.
6. The review of technology pertaining to *provenance* – Section 0 pages 89 onwards. The use of provenance is essential to achieve and validate high-quality research. It supports the automated collection of records about how data was produced, the inputs, the processes applied and who organised that production. It can be selectively applied and users can add their own annotations. It is required for curation and for review of conclusions but it may also be used for diagnostics, help with data management and rapid reruns. There is a deep discussion and analysis of the options for its implementation and candidate standards.
7. The review of technology pertaining to *optimising* – Section 2.3.6, pages 53 onwards. Optimisation has to be considered for every aspect of the data lifecycle and for the majority of working practices that have moved into production. It is motivated either by

[5] http://www.w3.org/

aspects of the user experience proving problematic or by resource consumption becoming unacceptable. The engineering approaches needed to address such issues are identified. The handling of these issues are much helped by ensuring that the platforms and subsystems have the right levers for optimisation engineering. It is clear that one-off solutions are rarely affordable, particularly in the long term. Consequently, metadata-driven automated optimisations are the appropriate strategy.

Review topics – the four cross-cutting aspects of technology

8. The review of technology pertaining to *system architectures* – Section 0 pages 102 onwards. This explores the options for combining the many functions and capabilities in the typical distributed context. A recommended strategy has a logically unified core, that manages such things as definitive catalogues, with many external resources and data sources attached. Standards pertinent to this organisation, example systems and current RDA discussions are recommended. The issues to be faced in e-Infrastructure architecture are illustrated by considering routes to interoperability. This exposes the critical importance of careful decision making during the early stages of e-Infrastructure design and development.
9. The review of technology pertaining to *semantic linking* – Section 0 pages 107 onwards. The precise descriptions of data and services will never use the same terminology everywhere. Furthermore, the chosen terms, structures, notations and vocabularies all evolve as the environmental and Earth sciences develop and as the growing wealth of digital devices and of data exposes new information. The strategy for not just coping, but for making the best use of this wealth of data, depends on automation built on formalised logical descriptions.
10. The review of technology pertaining to *reference model* – Section 3.10 pages 110 onwards. Reference models enable the many organisations engaged in building or revising an e-Infrastructure to describe the crucial structures and agreements so that the system eventually fits together well and performs as intended. This is particularly relevant for the many RIs and groups of RIs that are going to become dependent on their e-Infrastructures; it will aid the design, planning and implementation phases. The reference model provides a vocabulary and conceptual framework for many of today's pressing e-Infrastructure decisions. Its development and wider use will pay off substantially for the RIs, from their strategists and technical teams building their e-Infrastructure to the researchers and other users who expect it to be consistent, perform well and contain the automation they need.
11. The review of technology pertaining to the supply and use of compute, storage and network resources – Section 3.11 pages 115 onwards. Everything that all of those handling data does, everything that the teams building and maintaining e-Infrastructure do, and everything that the user communities and citizen scientists do, depends on the strength, power and availability of this underlying layer. A crucial contribution is the many layers of platform software and subsystems that the e-Infrastructures depend on. Carefully chosen, these supporting resources can provide much of the sophisticated system engineering needed, greatly reduce the system and software maintenance burdens for RIs, and deliver support to developers and users.

The locations of these technology review topics are summarised in Table 1.

TABLE 1: TECHNOLOGY REVIEW TOPICS

Technology topic	Section/Page
Data identification and citation	3.2 / 62
Data curation	3.3 / 73
Cataloguing	3.4 / 76
Processing	3.5 / 82
Provenance	0 / 89
Optimising	3.7 / 98

Architectures	0 / 102
Semantic linking	0 / 107
Reference model	3.10 / 110
Compute, storage and network resources	3.11 / 115

The technology reviews listed above are assessed and analysed to identify their scope and implications in corresponding sections of Section 4.2 pages 136 onwards, namely Sections 4.2.5 to 4.2.10 for the specific topics and Sections 4.2.11 to 4.2.14 for the aspects of technology that apply to all of the subsystem pillars. These assessments are preceded by four strategic considerations that should shape the current R&D reported in those sections, but should also influence the long-term planning of the RIs, of collaborating computational-resource providing organisations and the funders of e-Infrastructures:

1. The critical importance of skills in collaborating effectively between roles and between disciplines. Crossing these intellectual and cultural barriers is essential to make RIs and e-Infrastructures successful and to address today's societal challenges. It is equally important in business and government. The ENVRIplus community should be leaders in developing and valuing these boundary-crossing behaviours because the future of Europe's economy and societal well-being depends on that capacity – Section 4.2.1.
2. The integration of the mathematical models that have described natural phenomena very successfully since Newton's time are complemented by the powerful statistical methods exploiting our new wealth of data that took off at the start of this millennium—the Fourth paradigm. How to harness both of these approaches together remains a challenge in many disciplines. Individual solutions are developing in many disciplines, but principle and both intellectual and technical frameworks are needed to nurture this alliance – Section 4.2.2.
3. Almost all RIs and all the campaigns to address societal challenges depend on building effective alliances between many autonomous organisations and establishing rules and practices for sharing their independently owned data. These organisations often need to meet priorities of their funders, e.g., governments. They continually improve their offered services and they are often involved in many such consortia. We name these alliances to present a consistent and convenient view of independently owned data, as '*data-intensive federations'*. The governance, principles and technologies will all benefit from developing a shared kernel that can be used by many data-intensive federations – Section 4.2.3.
4. The handling and exploitation of data depends on growing bodies of software. The interfaces and data transformations delivering a holistic view depends on a similar complex assembly of software. The working practices and scientific methods, and the convenient interfaces that make them accessible takes more software. The science and the scientists depend as much on their software as on their instruments. The software strategy for RIs and for other e-Infrastructure builders must take careful account of the often overlooked question of sustainability; ways of doing this are introduced – Section 4.2.4.

To support the characterisation of the outcomes and implications five categories are introduced in Section 4.3. These are:

1. **Building on Task 5.1 results**: How do we take forward and develop the collected information and judgements?
2. **Raising the abstraction level in the universe of discourse**: How do we make best use of the reference model to improve the applicability of statements and highlight commonalities?
3. **Awareness raising and training**: we all need to improve our understanding, knowledge and skills to work across boundaries, address ever greater scales of data and activity, and deal with the complexities of the natural and artificial systems?

4. **Usability and take up**: investing in advanced e-Infrastructures and encoding sophisticated scientific and data-management methods will be to no avail if the practitioners do not exploit the new capabilities; how do we keep them engaged and show the potential?
5. **Shared subsystems and sustainability**: How do we minimise the parts which have to be maintained by a small group of RIs, and how do we ensure that those parts are maintainable?

Using these categories, Section 5.1 lists twenty five suggested actions that ENVRIplus should consider. This is not intended to be an exhaustive list, and others should add to it and refine it. It is particularly important that those addressing use cases ensure that their requirements are inserted in future versions of this list. Its initial content will be used as the basis for discussion during the 2016 Spring ENVRI week[6]. These actions will ensure that ENVRIplus and the RIs hear about this comprehensive body of work through extracts highlighting particular issues. They will also build on the material developed here to make it a living resource for the project. T***hink tanks*** may be formed to pool intellectual effort, gather sufficient breadth of experts and to ensure viewpoints are balanced. A competition might be run that will select from proposals for think tanks those that will best serve the environmental cluster communities.

The more strategic and wider implications are summarised in Section 5.2 starting on page 191, with backward references to the details that lead to them. Again, these are indicative, and more thought should be given to the population of this list of eight items and to the exact form of response each item warrants. The current items are:

1. **Improving interdisciplinary collaboration**: How should we invest in the skills and capabilities of our communities, so that we become more expert at collaborating across discipline, role and organisational boundaries?
2. **Leading the formation of a global environmental sounding board**: The context for thinking together and planning from the viewpoint of the environmental cluster but on a global scale needs developing.
3. **Combining both statistical and numerical methods**: Expertise in using numerical models is well established, the use of statistical methods is advancing rapidly, but how well developed are the methods for making them work together?
4. **Sharing computationally expensive results**: When large investments have been made to compute a result that has to be represented by large volumes of data, how best do we share the benefits and encourage others to take advantage of the results?
5. **Data-Intensive Federation support**: Many environmental research activities depend on dynamically sharing data from autonomous organisations; a framework that enables this for many federations, each able to use it to tailor their own, would yield significant benefits.
6. **Software sustainability**: The whole of the data lifecycle and all data-driven science depends on software. Software is expensive and difficult to maintain. The RIs will expect it to work for many years. This will require careful choices of software and sufficient resources to meet the remaining software-maintenance costs.
7. **Promoting ICT harmonisation**: Harmonisation yields benefits when dealing with border-crossing collaborative research and inter-disciplinary data sharing. But its greatest benefit is to help with the software sustainability challenge by pooling effort and by creating sufficiently consistent and extensive demand that ICT vendors become interested in co-development and supply.
8. **How should decisions be made?** The items above and the competing pressures for improvement mean that decisions need to be well informed, to balance long-term and immediate issues, and need authority that leads to their proper adoption. The

[6]ENVRI week Spring 2016: http://www.envriplus.eu/2016/02/25/2nd-envri-week/

mechanisms by which current ICT decisions are made will probably benefit from review and revision.

This document is a compendium of many individual investigations, searches, researches, discussions, analyses and judgements; as such it is indigestible taken as a whole. We therefore present a map to help readers navigate to the parts that interest them – see Figure 1.

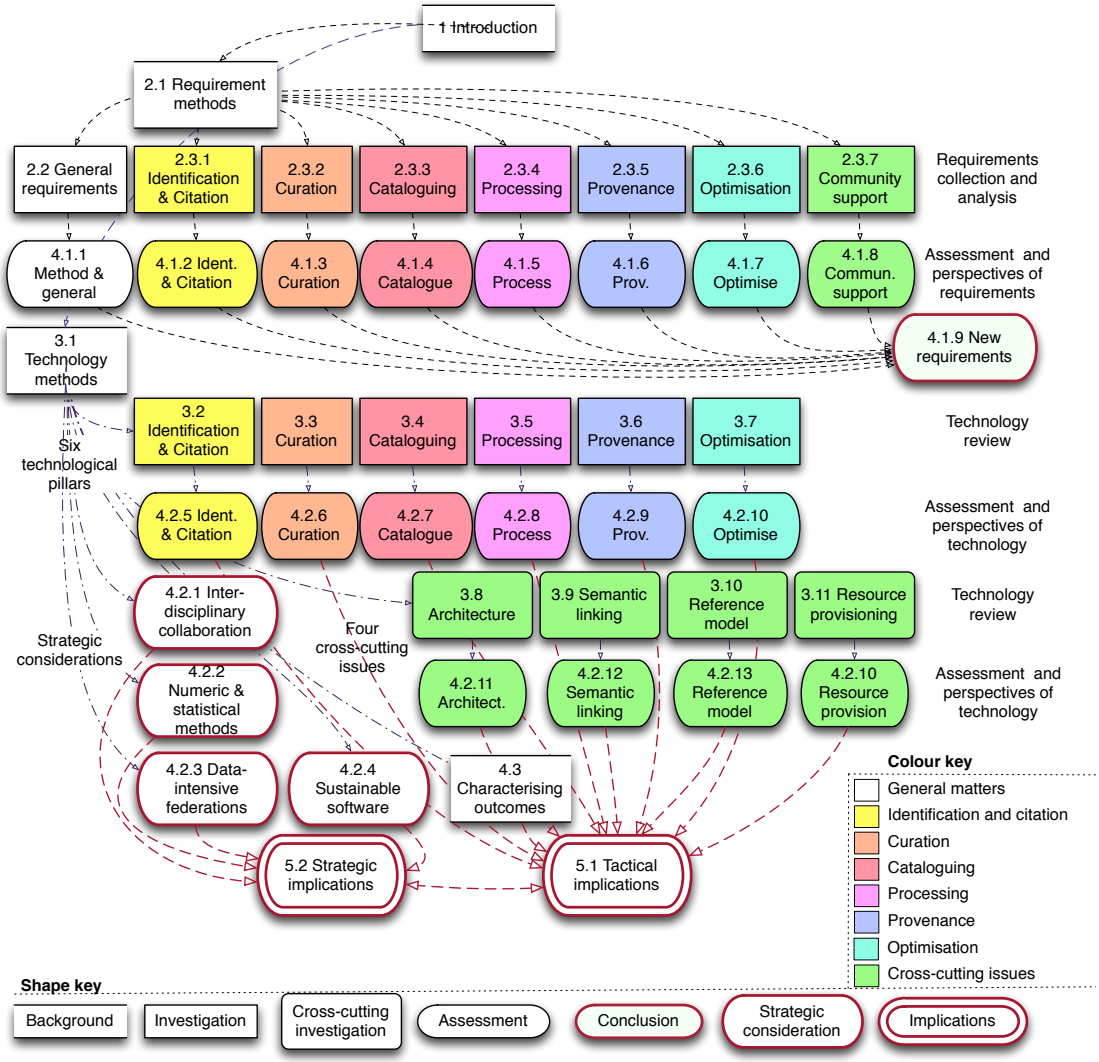

FIGURE 1: MAP OF THIS DOCUMENT SHOWING TOPICS AND TREATMENTS OF THE INVESTIGATIONS.

Those concerned with a particular topic, e.g., Curation, can follow the relevant column and colour scheme, e.g., orange, possibly also taking into account relevant general matters and cross-cutting issues. Similarly, those interested just in requirements can follow the corresponding rows, as can those just interested in technology reviews. Those concerned with implications are invited to read those sections with red borders. Those concerned with strategy and long-term planning should focus on the topics with double red boarders. If you start reading with a specific viewpoint, there are sufficient cross-references in the running text to lead you to relevant material.

We hope readers will enjoy the contents; as we have enjoyed assembling it. Of course, had we had more time we could have investigated much further, every topic certainly warrants it! Similarly, more judgement and analysis could have been invested to sharpen our conclusions. However, this moment is an appropriate one from the point of view of ENVRIplus and from the

point of view of completeness and depth, to share this with you. We will greatly appreciate feedback and comments to improve future versions and to shape future plans and judgements.

1 Introduction

Data has been crucial in science since observations and experiments began; Archimedes must have measured the displaced water for his "Eureka" moment. The advent of digital instruments and the intensive use of computers has transformed the ways in which we study and interpret phenomena; Jim Gray coined the term "*Fourth Paradigm*" for this transition in 2008 after working intensively with astrophysicists exploiting sky surveys [Szalay 2008]. As Szalay points out, this has transformed every branch of science; their wealth of data doubling every year offers tremendous opportunities for developing new understanding but it also poses immense challenges in how to handle and exploit that wealth of data well. All of the sciences concerned with the Earth and the environments it offers are experiencing this transition. Many of the Research Infrastructures in ENVRIplus are engaged in generating and exploiting that data. The project's aim is to help them do this as well as possible; helping them optimise their working practices and the platforms that support their data pipelines from distributed data generators (instruments, sensors, observers) to storage, use, presentation and application. This advances from the previous model where individual researchers assembled data and analysed it as a one off step. The new data scales, diversity, and complexity mean that such one-off approaches are no longer feasible or supportable [Burns 2014].

Today's societal challenges such as hazard mitigation and sustainable resource provision require new interdisciplinary approaches pooling resources, insights, data, methods and models. It is a challenge for the collaborating environmental RIs to be leading in supporting researchers in this challenging scientific field. Thus in the ENVRIplus context the data-driven science opportunities and challenges are compounded and also crucial for each individual RI in the upcoming years. Although the RIs have to shape their own ICT strategy – which will be addressed in the next paragraphs – this report is a first step in supporting them in approaches to common problems. There are, therefore, immense potential benefits from shared solutions that ENVRIplus hopes to encourage in reducing barriers, thereby facilitating boundary crossing. Developing a common understanding of requirements is a first step.

This will need to build on strategies for globally sharing data. The international sharing of meteorological data commenced in 1873[7]. The advent of networks of digital observation systems and the multiple methods of computationally deriving data poses new data sharing challenges. These were first explored for *curated computationally produced digital* data in 1972 by the X-ray crystallographers. They set about openly sharing their data about the structures of biologically significant molecules, such as haemoglobin[8]. They mandated that any publications reporting new structures had to be matched with a PDB deposited data set. The stages by which this evolved over the first 40 years to meet new needs and to exploit the advances in ICT are given in [Berman 2008]. Today PDBe, the European support for PDB, is just one of 24 life-science curated reference data sources the environmental research infrastructure ELIXIR[9] supports. Its community includes world leading centres at the frontier of data-driven science. Consequently, it has a very advanced e-Infrastructure and sophisticated strategies for developing it further. ENVRIplus is more likely to learn from such research e-Infrastructures than influence them. There are many others in the environmental research infrastructure cluster that have deep histories and are similarly advanced.

Collaborative sharing of reference data archives, with much improved technology, is now widespread in all sciences, and certainly key in environmental sciences, where global consortia

[7] From Wikipedia: The **International Meteorological Organization** (1873–1951) was the first organization formed with the purpose of exchanging weather information among the countries of the world. It was born from the realization that weather systems move across country boundaries; and that knowledge of pressure, temperature, precipitations, etc. upstream and downstream is needed for weather forecasting. It was superseded by the World Meteorological Organization.

[8] Initially data was shared by posting to registered laboratories a new magnetic-tape master each month, with a few authorised to approve additions to the master.

[9] https://www.elixir-europe.org/

are capturing and sharing data about many aspects of the Earth and its biosphere. Given the scale and maturity of many environmental research infrastructures, ENVRIplus needs to focus on finding ways of improving their interaction, e.g., by sharing methods and solutions, and enhancing the opportunities for combing data from multiple RIs. It may be instructive, given this goal of sharing methods, solutions and data among large investment research campaigns to review a strategy that worked well for digital astronomy. Innovation in such a context was pioneered by astrophysicists for sharing many significantly different sky surveys. They call their scientific gateways that give access to the collection of data produced by one sky survey a "*Virtual Observatory*" (VO). They recognised the significant advantage from all of these VOs offering consistent services for both human interaction and *computational interaction*. This meant the careful definition and verification of *globally* adopted standards. But that had to avoid the undesirable effects of lock-in to poor standards and the chaotic effects late agreement on newly needed standards as each instrument and observing campaign introduced new data, and as each advance in data analytics required new elements in their catalogues. Astrophysicists therefore took matters into their own hands and formed the International Virtual Observatory Alliance (IVOA)[10]. This speedily judges new requirements, encourages researcher-led proposals, and verifies global adoption of agreements, typically through six-monthly cycles of catalogue re-builds.

Ernst Mayr pointed out that biological systems are more complex than physical systems [Mayr 2004][11]. This makes the development and adoption of effective, relevant and widely adopted standards much more important for environmental and Earth sciences and many of those consulted echoed this sentiment. However, it also makes the task more challenging and that challenge is exacerbated by the connection with societal challenges and economic factors that mean many additional viewpoints need to be considered—the INSPIRE directive is one example [EU Parliament 2007]. We can envisage an International Virtual Earth and Environment Alliance (IVEEA) to take on this mantle. It is doubtful whether this can be grown in the context of existing organisations. Once an organisation such as IVEEA exists, it would take responsibility for a long-term and detailed campaign of requirements gathering and analysis as a necessary precursor to agreeing and adopting standards. Such an initiative is foreseen already by ESFRI [ESFRI 2016] as a recommendation for the long-term but we suggest this needs to be accelerated. Such a body could also complement the Belmont Forum; the world collective of major and emerging funders of global environmental change research. The ENVRIplus communities might consider the value of such a body and decide to nurture its creation.

The last decade has seen the emergence of *data science*. This has emerged as four factors have combined:

1. The rapid increase in volumes of collected data stored within one regime, e.g., Google;
2. The rapid increase in the affordable power to conduct statistical analyses over very large volumes of data;
3. Substantial advances in the machine learning methods that have become feasible because of the above two factors; and
4. Significant successes using these techniques in finance, business, science and medicine.

ENVRIplus is committed to enable the RIs and their research communities to fully exploit data-science advances. This poses both intellectual and technical challenges. As these are propagating through the Environmental and Earth sciences contemporaneously with ENVRIplus [Aston 2016], they are a perturbing factor that should be considered as we gather and analyse requirements.

[10] http://www.ivoa.net/

[11] Some of the environmental research infrastructures deal with physical systems, but here they are complex, and as exemplified by solid Earth and climate sciences, have to deal with the complexities that come from a deep history and many interacting systems. Mayr was jousting at particle physics, where the previous history of a particle does not normally affect its behaviour.

The requirements gathering began from the start of the ENVRIplus project and continued as a number of parallel dialogues, with oversight by the *topic leaders* and coordination by the *task leaders*. Please see below for definitions of terms. The results were collected and refined in the ENVRI Community wiki[12], which will be referenced frequently throughout this report. It should be consulted for detail and for up-to-date information as the wiki will be active after this report is completed. These requirements were then reviewed and summarised by the topic leaders. The state of that material when this report was completed led to the summary information in the sections below. We first present the methods used. We then present each of the topic areas around which requirements gathering was focused. For each of these topic areas there is an initial summary that digests and assesses the overall information gathered. Then within each topic area we briefly review the information per RI that was engaged in the process. We conclude with a short summary for that topic that identifies and quantifies common factors and enumerates any exceptions.

2 Review of existing RIs: their requirements

2.1 Requirement gathering methods and completion status

Task 5.1 aims to re-analyse the status of involved Research Infrastructures[13] (RI)s in ENVRIplus along the dimensions of data, users, software services and resources in order to update the requirement study performed in the early phase of ENVRI, the precursor to ENVRIplus, describing the commonalities, differences and interoperability between RIs and reviewing the state-of-the-art of RI technologies.

The requirements study used the following workflow, conducted in parallel by individuals from each RI and also employed by the project, pairs and small groups:

1. *Committing* to focus on a topic and in many cases a context.
2. *Starting* a new page to record their findings, and delimiting their scope.
3. Working progressively to *refine and record their understanding* of their chosen focus. Recording progress on their page and linking to uploaded files and cited references for details.
4. Indicating their *progress* by revising their status.
5. *Reporting* to the weekly Theme 2 meeting when appropriate.
6. *Contributing* to the wiki and deliverable report.
7. Engaging in *handover* to subsequent tasks and RIs.

At the end of this independent and concurrent requirements gathering the topic leaders (see below) reviewed, collated and analysed these focused descriptions to develop an integrated overall report on requirements in their area and discussed these with the wider team. These are summarised later in this report from Section 2.3.1 page 46 onwards.

The first step of the requirements study was to define a common method for describing all aspects of the Information and Communication Technologies (ICT) that are needed to provide the facilities and capabilities required by researchers using environmental Research Infrastructures (RIs). This led us to group the requirements under seven common topics:

[12] https://wiki.envri.eu/display/EC/ENVRI+RI+Requirements

[13] A Research Infrastructure (RI) is an organisation and technological infrastructure to enable a community of researchers to pursue a particular, domain-specific, research goal that requires significant sustained resources and expertise. Many of the *environmental* RIs in ENVRIplus are endorsed by the European Strategic Forum for Research Infrastructures (**ESFRI**), http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri, in their 2016 road map [ESFRI 2016]. The technology involved often includes instruments and observing systems, and extensive, distributed, digital-data transport, transformation and management.

1. *Identification and citation:* Mechanisms to provide durable references to data objects and collections of data objects.
2. *Curation:* Processes to assure the availability and quality of data over the long term.
3. *Cataloguing:* Catalogues are built to accelerate access to data subsets that can be delimited by queries over a catalogue. They may optimise work by containing frequently required data derivatives, so that many scientists use these instead of re-computing them—this, of course, needs to be supported by appropriate APIs, metadata and steps in formalised methods, which appear under *Processing* and *Optimisation*. They also advertise the data resources that a particular service or collection has so scientists and software know when to use the underlying facility. Either the same or similar catalogues may collate and yield access to formalisations of methods and working practices, often encoded as services, scripts, workflows or other forms of software.
4. *Processing:* This includes every computational transformation of data including but not restricted to the following examples: processing and selection of raw data close to instruments, including signal processing, analysis of data for quality assurance (QA) purposes or to derive results used elsewhere, simulation runs with subsequent comparison with observations, etc.
5. *Provenance:* This is concerned with recording information about how data, code and working practices were created and were transformed to their current form. It not only records such historical information, it also works as a foundation for many tools that help researchers organise and evaluate their research.
6. *Optimisation*: Optimisation transforms data handling and computational processes so that they achieve the same effects from the viewpoint of domain scientists, curators and other practitioners. The optimisation may address any cost function a community chooses, e.g., energy used, financial charges or response time, or some combination of these.
7. *Community support:* Community support addresses all aspects of the use of resources and the relationships with resource providers.

Table 2 identifies who is responsible for leading the requirements gathering for each of the seven specialist topics. Information was also gathered for general requirements, that was also a background and context for these specialist topics. That was led by Rosa Filgueira.

TABLE 2: LEADER OF EACH SPECIALISED REQUIREMENTS TOPIC

Topic	Topic Leader	Organisation
Identification and citation	Alex Vermeulen	ICOS (LU)
Curation	Keith Jeffery	BGS
Cataloguing	Thomas Loubrieu	IFREMER
Processing	Leonardo Candela	CNR
Provenance	Barbara Magagna	EAA
Optimisation	Paul Martin	UvA
Community Support	Yin Chen	EGI

To coordinate concurrent requirements gathering, the ActiveCollab tool[14] was used. All the information requested by each topic leader was collected in a single document[15]. That document describes the information to be gathered for each topic, including generic information, such as

[14] The shared information management and coordination framework provided for all of ENVRIplus by the project coordination team.
[15] https://envriplus.manageprojects.com/s/notebook/m69o8P0wwr3eT/page/82

the size and maturity of each RI that pertains to all of the topics. The generic material is gathered in the wiki with the title: *Generic requirements and background*.

Once the topics and generic requirements were defined, the next step was to design a communication strategy (Figure 2 below). This strategy involved three roles:

1. *Topic leaders* who partitioned and shaped investigations at the start to identify the information they would like gathered. At the end of the requirements gathering period they reviewed, integrated and interpreted the gathered information about their topic and produced a summary and analysis taking into account their own knowledge and their work package (WP) commitments. They had opportunity to interact with the go-betweens throughout the requirements gathering process.
2. *RI representatives* gathered information about their RIs and communicated it to their *ICT-RI go-between.* This often triggered follow-up discussions with others.
3. *ICT-RI go-betweens*, focused on one (or a few) RIs and gathered information from their RIs, recorded it in the wiki or ActiveCollab page, and fed it to the relevant topic leaders. They also arranged follow-up discussions with one or more topic leaders, when these were requested.

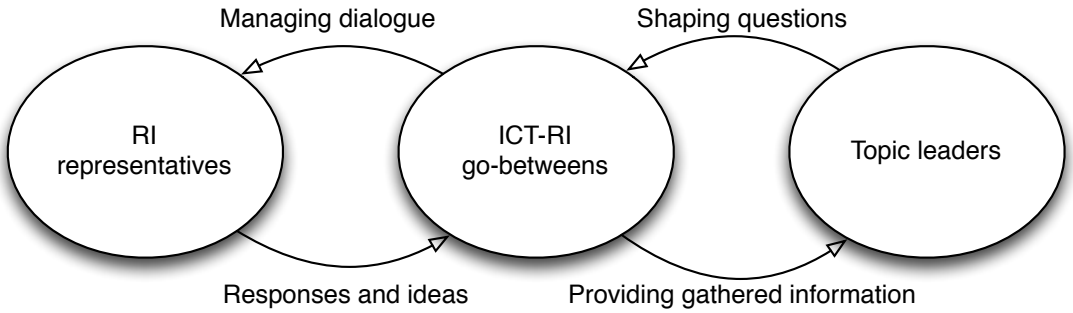

FIGURE 2: THREE ROLES ENGAGED IN REQUIREMENTS GATHERING

The role of a *topic leader* is defined in the joint ENVRIplus ActiveCollab communication tool[16]. They had to be receptive to input from *ICT-RI go-betweens* and had to partition and delimit their topic to minimise duplication of work by those contributing to their topic.

The role of an *RI representative (RIREP)* was to collect and present to requirement gatherers information about their RI's requirements, including its existing inventory of facilities, its plans as they affect technical choices, their alliances with e-Infrastructure providers and the work of various roles within their RI who need better data facilities. They introduced others from their RI into the requirements gathering process to work directly on specific issues or topics. These have been identified by formal responses in the ActiveCollab[17,18].

The role of an *ICT-RI go-between* (*GB*) was to avoid duplication of effort by an *RIREP* in an RI they are responsible for. Otherwise, an *RIREP* might have had to field overlapping questions from a succession of topic leaders. The GBs were guided by a common set of information requirements[19]. They developed an awareness of the common factors that have to be completed to meet the standard template for requirements reporting[20].

[16] https://envriplus.manageprojects.com/s/notebook/m69o8P0wwr3eT/page/22
[17] https://envriplus.manageprojects.com/s/notebook/m69o8P0wwr3eT/page/19
[18] https://envriplus.manageprojects.com/s/notebook/m69o8P0wwr3eT/page/22
[19] https://envriplus.manageprojects.com/s/notebook/m69o8P0wwr3eT/page/23
[20] https://envriplus.manageprojects.com/s/notebook/m69o8P0wwr3eT/page/83

A common set of actions, time issues, and deadlines for the interactions between *GB* and *RIPEP* were defined in ActiveCollab[21]. Once each *GB* agreed to take the responsible for at most four RIs (Table 3), they identified the *RIPEP* for each of their assigned RIs. Later, *GBs* conducted a sequence of interactions with the *RIREPs* to build an understanding of that RI's requirements, and to develop a written record that they both agreed to. These interactions were always initiated by collecting the "*Generic requirements and background*". In the subsequent interactions, information for each topic was gathered. Then, *GBs* communicated that understanding and record to the relevant *topic leaders*. Each pair *GB-RIREP* kept their collected records in an ActiveCollab Notebook page, one per RI[22]. On some occasions, *GBs* with the *RIREP*, arranged direct communications between others in the RI for a topic, and then delegated the pursuit of more detailed understanding and requirements to them within the ethical framework. Such delegation of direct communication was explicitly consented, initiated and written up.

Three of the research infrastructures do not appear in the tables, namely **EUROFLEETS2**, **JERICO** and **ESONET** because their requirements are covered by **SeaDataNet** and **Euro-ARGO**. More specifically:

- The **EUROFLEETS2** research infrastructure does not have an integrated data management infrastructure; instead, it relies on **SeaDataNet's** network of data centres.
- The **JERICO** research infrastructure does not have an integrated data management infrastructure.
    - It relies on **SeaDataNet** for long-term data preservation and distribution.
    - It relies on Copernicus Marine service for real time data management.
- For the **ESONET** research infrastructure the requirements are not yet fully determined, but **SeaDataNet** contributes to ESONET's data management.

As a consequence, although ENVRIplus represents 20 research infrastructures, only 17 are shown in each table, with **SeaDataNet** representing itself and these other three.

TABLE 3: INDIVIDUALS COMMITTED TO REPRESENT RIS AND GO-BETWEENS

RI	RI Representative	Go-Between	Notebook Page
ACTRIS	Lucia Mona	Rosa Filgueira	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/36
AnaEE	André Chanzy	Paul Martin	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/37
EISCAT-3D	Anders Tjulin	Paul Martin	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/38
ELIXIR	Guy Cochrane, Petra ten Hoopen	Barbara Magagna	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/39
EMBRC	Ilaria Nardello	Cristina Alexandru	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/40
EMSO	Robert Huber	Paul Martin	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/41
EPOS	Daiele Bailo	Rosa Filgueira	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/42
Euro-ARGO	Sylvie Pouliquen	Thierry Carval	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/43
EuroGOOS	Glenn Nolan	Cristina Alexandru	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/57
FixO3	Andree Behnken Robert Huber	Yin Chen	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/58
IAGOS	Damien Boulanger	Yin Chen	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/59
ICOS	Margareta Hellström	Alex Vermeulen	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/60

[21] https://envriplus.manageprojects.com/s/notebook/m69o8P0wwr3eT/page/25
[22] https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/36

RI	RI Representative	Go-Between	Notebook Page
INTERACT	Morten Rasch	Barbara Magagna	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/61
IS-ENES2	Sylvie Joussaum	Yin Chen	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/62
LTER	Johannes Peterseil	Barbara Magagna	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/65
SeaDataNet	Michele Fichaut	Thomas Loubrieu	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/66
SIOS	Vito Vitale	Yin Chen	https://envriplus.manageprojects.com/s/notebook/UjPxIJI2mtB3M/page/67

All *GBs* ensured that the ethical procedures were implemented. In particular, ensuring that those involved received the information sheet and signed the consent form[23], and that informed consent was given to cover all of the requirements gathering discussions. If sound recordings were used, they protected their privacy and arranged that they were deleted once used. They ensured that participants agreed to the written record before it was passed on to others to use.

Each *topic leader* integrated and summarised the initial information gathered by *GBs*, raising issues needing clarification if necessary, and produced an integrated overview, summary and collation of the material for their topic. These appear in the wiki and a snapshot is summarised in this report.

The *requirements coordinator* integrated the *topic leaders'* results and developed an executive summary and integrating overview, asking for clarifications when necessary.

The collected information varies by topic and by RI. Possible reasons for this are discussed in Section 4.1. The current status is recorded in Table 4, where a tick indicates the information was gathered, recorded and agreed by the stakeholders. The crosses indicate requirements investigations that have not been completed. Various reasons led to this: the topic was not relevant at this time in that RI, the topic was known to have been covered by another RI with the same requirements, or the relevant experts were unable to allocate sufficient time to reach completion. In other cases, an infrastructure is too complex to be described in all its facets, thus only a few use cases are provided not offering a comprehensive view of a RI on the topics (e.g., ELIXIR).

Requirements topics:

0. Generic requirements and background
1. Identification and citation
2. Curation
3. Cataloguing
4. Processing
5. Provenance
6. Optimization
7. Community Support.

TABLE 4: REQUIREMENT GATHERING PROGRESS

RI	Generic req.	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
ACTRIS	✓	✓	✓	✓	✓	×	×	✓
AnaEE	✓	×	×	×	×	×	×	×
EISCAT-3D	✓	×	×	✓	×	×	×	×
ELIXIR	✓	×	×	×	×	×	×	×

[23] https://envriplus.manageprojects.com/s/notebook/m69o8P0wwr3eT/page/80

RI	Generic req.	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
EMBRC	✓	✓	✓	✓	✓	×	×	✓
EMSO	✓	×	×	×	×	×	×	×
EPOS	✓	✓	✓	✓	✓	✓	×	✓
Euro-ARGO	✓	✓	✓	✓	✓	✓	✓	✓
EuroGOOS	✓	✓	✓	×	×	×	×	✓
FixO3	✓	×	×	×	×	×	×	×
IAGOS	✓	✓	✓	✓	✓	✓	×	×
ICOS	✓	✓	✓	✓	✓	✓	✓	✓
INTERACT	✓	×	×	×	×	×	×	✓
IS-ENES2	✓	✓	✓	✓	✓	✓	✓	✓
LTER	✓	✓	×	✓	×	✓	×	✓
SeaDataNet	✓	✓	✓	✓	✓	✓	✓	✓
SIOS	✓	✓	✓	✓	×	×	×	✓

In order to organise the RI requirements analysis, space in the ENVRI Community Wiki was utilised[24]. A top-level page for recording requirements was created[25], and each *GB* imported all of their material gathered into a dedicated wiki page per topic and generic information, and per RI (e.g., for ACTRIS[26]). The contents and organisation of the Wiki space is explained on the 'Getting started' page[27].

2.2 Gathered generic information

This concerns all of the information that is not related to a specific topic, e.g., the role and characteristics of each RI and quantifications that may be indicative of scale and diversity factors.

2.2.1 Summary of generic information

ENVRIplus brings together Environmental and Earth System RIs, projects, networks and technical specialists with the common ambition to create a holistic, coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. ENVRIplus gathers all domains of Earth system science – Atmospheric, Marine, Biosphere/Ecosystem and solid-Earth science to work together, to capitalise on the progress made in various disciplines and strengthen interoperability amongst RIs and domains.

Table 5 gives an overview of the RIs that have participated in the Task 5.1. These RIs are typically composed of distributed entities (data generators, data processors, data sharers) and thus federations of often diverse autonomous organisations. These organisations have established roles, cultures, working practices and resources. The organisations' roles must remain unperturbed, as they are their primary business. RIs and organisations have internal diversity that may be relevant. They may need to incrementally engage with their federations. Organisations are often engaged in many federations. They would then benefit from using the same framework for each federation. Federating for multi-domain science is one of the goals of ENVRIplus.

[24] The ENVRI Community wiki, http://wiki.envri.eu/ is part of the ENVRI Community platform, http://www.envri.eu/. It is the collaboration and documentation space where members of the wider ENVRI community, as well as participants in the ENVRIplus or other projects can author or discover information relevant to a wide range of ENVRI RIs stakeholders.
[25] https://wiki.envri.eu/display/EC/ENVRI+RI+Requirements
[26] https://wiki.envri.eu/display/EC/Identification+and+citation+in+ACTRIS
[27] Requirements review wiki pages, https://wiki.envri.eu/display/EC/Getting+started+(RI+Requirements)

TABLE 5: RIS CONTRIBUTING TO REQUIREMENTS GATHERING

RI	Type of RI[&]	Domain[&]	Current Status[±]	Data lifecycle[+]	ESFRI 2016 Roadmap*
ACTRIS	Distributed	Atmospheric	Entry	Production to publishing	✓
AnaEE	Distributed	Biosphere, Ecosystem	Preparatory	Curation to processing	✓
EISCAT-3D	Single RI, multi-site	Atmospheric	Construction	Production to publishing	✓
ELIXIR	Distributed	Biosphere, Ecosystem	Operational, ELIXIR CA 2013	Acquisition to publishing	✓
EMBRC	Distributed	Marine, Biosphere, Ecosystem	Construction, Operational	Production to publishing	✓
EMSO	Single RI, multi-site	Marine, Multi-Domain	Operational, ERIC	Acquisition to publishing	✓
EPOS	Distributed	Solid Earth	Implementation	Acquisition to publishing	✓
Euro-ARGO	Distributed	Marine	Operation, ERIC	Production to publishing	✓
EuroGOOS	Distributed	Marine	Operational	Production to publishing	•
FixO3	Distributed	Marine	Implementation	Acquisition to publishing	•
IAGOS	Distributed	Atmospheric	Operational, AISBL	Acquisition to processing	✓
ICOS	Distributed	Atmospheric, Marine, Ecosystem	Operational, ERIC	Acquisition to publishing	✓
INTERACT	Distributed	Biosphere, Ecosystem	Operational	Acquisition to publishing	•
IS-ENES2	Virtual	Multi-domain Earth's climate system	Integrated	Acquisition to publishing	•
LTER	Distributed	Biosphere, Ecosystem	Operational	Production to publishing	•
SeaDataNet	Virtual	Marine	Operational	Acquisition to publishing	•
SIOS	Distributed	All	Interim	Publishing	✓

NOTES

& Derived from general information
± According to ESFRI roadmap 2016, some stages overlap.
+ Data lifecycle as identified by ENVRIplus
* (✓)Included, (•) mentioned, (×) not included

The information of columns two and three is derived from the generic information provided by each RI. The current status of the RI indicated in column four is aligned with their status on the ESFRI Roadmap. The status is defined as the stage of the RI in the ESFRI lifecycle. The ESFRI lifecycle defines seven phases: (1) ESFRI Roadmap entry, (2) Preparation, (3) Interim, (4) Implementation, (5) Construction, (6) Operation start, and (7) Legal status (ERIC, AISBL, other). The ESFRI lifecycle is based on the chronology of events, including year of first appearance and year of re-application to the Roadmap, years of preparation phase (funded at national level or by EC FP), years of construction phase, year of start of delivery of some scientific services and

expected start of full operation. The legal status is indicated when established[28]. For RIs not listed in the ESFRI roadmap 2016, the status is derived from the requirements analysis. The description of the data life cycle coverage on column five is provided in the next paragraphs. The last column indicates whether the RI is included in the ESFRI roadmap 2016 (✓) in one of their stages or just mentioned (•).

The data lifecycle shown in column 5 reflects the stages of data handling from its production or acquisition to its final presentation, as defined by the ENVRI reference model (ENVRI RM)[29]. Some RIs include observation networks of scientists and/or instruments producing data (e.g., ACTRIS, EISCAT-3D, EMSO), while others provide advanced processing services (e.g., AnaEE, IAGOS, IS-ENES2). The details of the data lifecycle for each RI are presented in Table 6. The data lifecycle is shown in Figure 3. This matches the developments in the reference model that were part of the current task and have been published in the ENVRIplus Newsletter to raise awareness and to gain interaction with the RIs [Hardisty 2016]. The reference model is being refined and developed as the analysis of requirements and solutions takes place, and as the reference model practitioners are engaged in agile use-case development teams – see Section 3.10.

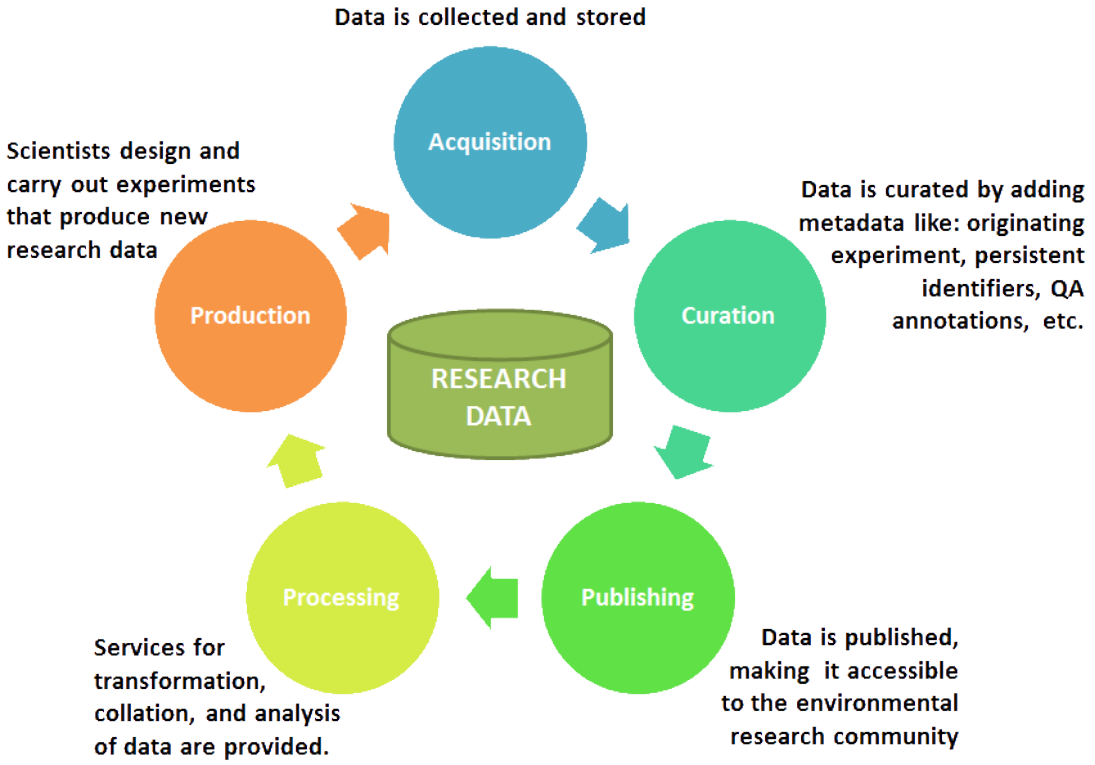

FIGURE 3: STAGES IN THE DATA LIFECYCLE

Table 6 shows the stages, their relationship to other data stages, a definition of the state and the activities that can support the transition of data to that state. The names in square brackets indicate synonyms used to describe the state.

TABLE 6 STAGES OF DATA LIFECYCLE

Stage	Data state	Definition	Supporting activities

[28] ESFRI (2016) STRATEGY REPORT ON RESEARCH INFRASTRUCTURES, ROADMAP 2016. Online: https://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/esfri_roadmap_2016.pdf
[29] http://envri.eu/rm

Production	Produced [Raw]	Data generated by experimental process, observation or automatic recording of events.	Setting up monitoring devices or networks of such devices Providing recording tools for individuals
Acquisition	Acquired [Registered]	Data stored in digital form.	Activities to store data in digital form (digitalization), regardless of the lifespan assigned to collected data.
Curation	Curated [Annotated, QA assessed, reviewed, mapped]	Additional data created to facilitate identification and retrieval	Activities designed to preserve, link, and identify data; such as: quality assessment, annotation, digital identification (DOI)
Publishing	Published	Additional data created to facilitate access	Activities designed to make data accessible to other parties
Processing	Processed	Additional data created from further processing	Activities designed to derive new data products, including information and knowledge.

In many cases, there is a much more complex pattern as successive uses repeat such cycles. For example, seismological observations are recorded and analysed in near real-time to detect earthquake events, and to alert responders if the magnitude and location warrant such actions. The accumulated, and quality controlled traces from seismometers are archived and curated as globally agreed and accessible data. These are supplemented by other deployments, such as the US seismic array and the responsive deployments after a major earthquake to obtain data from the aftershocks. These data are then correlated to identify subsurface phenomena, such as changes in the seismic wave velocity, normally due to change in temperature; or compared with wave-propagation simulations based on Earth models. Results from that misfit analysis can be back propagated to refine the Earth model. Using data from many earthquakes and seismometers, the Earth models can eventually reflect phenomena in the mantle, such as thermal plumes that manifest themselves in the lithosphere as chains of islands, such as the Hawaii archipelago [French 2015]. These Earth models can then be compared with the fluid dynamics models of mantle convection, to refine those models. Clearly data representing successive models depends on many stages, each of which treats the results from the previous stage, proceeds through a number of data-driven or model-driven scientific methods, and delivers results worthy of archiving and curation.

The generic aspects of each RI were collected first, as they show the high-level commonalities, differences and potential interoperability between RIs. For that purpose, the *GBs* asked a series of general questions to set the scene for subsequent discussions. These are available at ActiveCollab[30]. They covered the following areas of interest:

- The basic purpose of their RI, here including some representative use cases, data types and the data lifecycle, user and stakeholder responsibilities;
- High-level questions spanning the ENVRIplus main topics:
  - Data lifecycle
  - Data and services offered
  - Data standards and software used
  - Data management plan
  - Data security and access
  - Non-functional constraints
  - Optimisation plans
  - Interactions with other RIs
- What objectives and services their RI are expecting from ENVRIplus.

[30] https://envriplus.manageprojects.com/s/notebook/m69o8P0wwr3eT/page/82

Table 7 shows the wiki page for each RI generic requirements report, the authors (*GB* and *RIREPs*) of these reports, the date range in which the interactions between *GBs* and *RIREPs* were performed, and the volume of information recorded.

TABLE 7: GENERAL REQUIREMENTS AND BACKGROUND

RI	Authors	Wiki Page	Date	Volume
ACTRIS	GB: Rosa Filgueira RIPEPs: Lucia Mona, Markus Fiebig	https://wiki.envri.eu/display/EC/General+requirements+of+ACTRIS	July - November 2015	6 Pages
AnaEE	GB: Paul Martin RIPEPs: Abad Chabbi, André Chanzy, Christian Pichot	https://wiki.envri.eu/display/EC/General+requirements+of+AnaEE	September – November 2015	4 Pages
EISCAT-3D	GB: Paul Martin RIREPs: Ingemar Häggström, Anders Tjulin	https://wiki.envri.eu/display/EC/General+requirements+of+EISCAT-3D	September - October 2015	3 Pages
ELIXIR	GB: Barbara Magagna RIREP: Petra ten Hoopen	https://wiki.envri.eu/display/EC/General+requirements+of+ELIXIR	September 2015	3 Pages
EMBRC	GB: Cristina A. Alexandru *RIREP*: Nicolas Pade	https://wiki.envri.eu/display/EC/General+requirements+for+EMBRC	September - October 2015	8 Pages
EMSO	*GBs*: Paul Martin, Yin Chen *RIPEPs*: Robert Huber, Andree Behnken	https://wiki.envri.eu/display/EC/General+requirements+of+EMSO	August – September 2015	4 Pages
EPOS	*GB*: Rosa Filgueira *RIREP*: Daniele Bailo	https://wiki.envri.eu/display/EC/General+requirements+for+EPO	September - November 2015	6 Pages
Euro-ARGO	*GB*: Thierry Carval *RiIREP*: Sylvie Poulique	https://wiki.envri.eu/display/EC/General+requirements+for+EPOS	September - October 2015	7 Pages
EuroGOOS	*GB:* Cristina A. Alexandru *RIREPs*: Glenn Nolan, Julien Mader, *et al*.	https://wiki.envri.eu/display/EC/General+requirements+for+EuroGOOS	July - December 2015	6 Pages
FixO3	*GB:* Yin Chen, Paul Martin *RIREPs:* Andree Behnken, Robert Huber	https://wiki.envri.eu/display/EC/General+requirements+for+FixO3	September 2015	3 Pages

RI	Authors	Wiki Page	Date	Volume
IAGOS	*GB:* Yin Chen *RIREP:* Damien Boularnger	https://wiki.envri.eu/display/EC/General+requirements+for+IAGOS	November – December 2015	4 Pages
ICOS	*GB:* Alex Vermeulen *RIREP:* Margareta Hellström	https://wiki.envri.eu/display/EC/General+requirements+for+ICOS	September – December 2015	7 Pages
INTERACT	*GB:* Barbara Magagna *RIREP:* Morten Rasch	https://wiki.envri.eu/display/EC/General+requirements+for+INTERACT	October 2015	3 Pages
IS-ENES2	*GB:* Yin Chen *RIREPs:* Sylvie Joussaume, Francesca Guglielmo	https://wiki.envri.eu/display/EC/General+requirements+for+IS-ENES2	October – November 2015	6 Pages
LTER	*GB:* Barbara Magagna *RIREP:* Johannes Peterseil	https://wiki.envri.eu/display/EC/General+requirements+for+LTER	September 2015	3 Pages
SeaDataNet	*GB:* Thomas Loubrieu *RIREP:* Michele Fichaut	https://wiki.envri.eu/display/EC/General+requirements+for+SEADATANET	November 2015	8 Pages
SIOS	*GB:* Yin Chen *RIREPs:* Jon B. Orbek, Angelo Viola, Vito Vitale	https://wiki.envri.eu/display/EC/General+requirements+for+SIOS	October-December 2015	3 Pages

The following subsections summarise **each RI's purpose** and **what objectives and services those RIs expect from their participation in ENVRIplus**. They provide links to the complete generic requirements reports. Finally, the generic information analysis subsection compares the rest of information recorded (e.g., standards, software, optimization plans, etc.) across the RIs, pointing out the generic commonalities, differences and potential interoperability between RIs.

2.2.2 Generic information about ACTRIS

ACTRIS (Aerosols, Clouds, and Trace gases Research Infrastructure) addresses the scope of integrating state-of-the-art European ground-based stations for long-term observations of aerosols, clouds and short-lived gases[31].

The overall goal of ACTRIS is to provide scientists and other user groups with free and open access to high-quality data about atmospheric aerosols, clouds, and trace gases from coordinated long-term observations, complemented with access to innovative and mature data products, together with tools for quality assurance, data analysis and research.

ACTRIS is composed of observing stations, exploratory platforms, instrument calibration centres, and a data centre with three data repositories (also called topic databases): near surface data (EUSAAR), aerosol profiles (EARLINET) and cloud profiles (CLOUDNET). Currently, ACTRIS is developing a new database with satellite data linked to ACTRIS ground based data.

[31] A complete report on ACTRIS generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+of+ACTRIS.

ACTRIS would like, through their participation in ENVRIplus, to improve their interoperability so as to make their data as accessible and understandable as possible to others:

- to understand which are the best practices when researchers need to discover data,
- to link with others RIs, because there are many points in common (technologically and scientifically), and
- to improve through the experience of other RIs.

ACTRIS expects that ENVRIplus will provide technology/advice for:

- planning and managing the activity of sensors,
- developing understanding of how instruments work in extreme conditions, and
- improving the capabilities of small sensors.

2.2.3 Generic information about AnaEE

AnaEE (Analysis and Experimentation on Ecosystems) focuses on providing innovative and integrated experimentation services for ecosystem research. It will strongly support scientists in their analysis, assessment and forecasting of the impact of climate and other global changes on the services that ecosystems provide to society. AnaEE will support European scientists and policymakers to develop solutions to the challenges of food security and environmental sustainability, with the aim of stimulating the growth of a vibrant bioeconomy[32].

It is the intention of AnaEE to provide excellent platforms with clear accessibility conditions and service descriptions, and a clear offering to researchers. The gathering of information in a common portal should help with this. Experiences gathered from the construction and operation of other platforms would be helpful to shape this development.

Within the context of ENVRIplus, AnaEE is particularly interested in participating in the work on identification and citation and on cataloguing, as these are of fairly immediate concern to their infrastructure. Consequentially, it would be useful to synchronise their approach with other RIs. Processing is of some interest as well, in particular the interoperability between models and data, and the quality control of data produced by platforms.

2.2.4 Generic information about EISCAT-3D

EISCAT-3D is a research infrastructure that will use a new generation of phased array radars to study the Earth's middle atmosphere, ionospheric incoherent scatter and objects in space, contributing to near-Earth space environment research. It aims at establishing a system of distributed phased array radars. The system will enable comprehensive three-dimensional observations of ionospheric parameters and atmospheric dynamics above Northern Fenno-Scandinavia, which is an important location for research on coupling between space and the polar atmosphere[33].

EISCAT-3D will produce about 2 petabytes of data each year and aims at using standard systems for:

- data storage and cataloguing
- user authentication
- identification and citation of datasets

EISCAT-3D expects ENVRIplus to help them:

[32] A complete report on AnaEE generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+of+AnaEE

[33] A complete report on EISCAT-3D generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+of+EISCAT-3D

- ensure that the tools that they select are well documented and open, in order to minimise the risk of lock in to proprietary systems,
- define workflows for data, and
- ensure interoperability with other RIs and instruments via virtual observatories.

2.2.5 Generic information about ELIXIR

ELIXIR is a European infrastructure for biological information that unites Europe's leading life-science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information[34].

ELIXIR will provide the facilities necessary for life-science researchers — from bench biologists to chemo-informaticians — to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built.

By participating in ENVRIplus, ELIXIR would like to establish a closer collaboration with environmental Research Infrastructures (RIs) and improve their access to life science data. An enhanced interaction, a better insight into data structures and relevant data standards widely adopted across environmental RIs can facilitate an effective evaluation of areas of collaboration for development of new tools, services and training. Ultimately, this can lead to better interoperability and discoverability of environmental and life science data by users across atmospheric, marine, solid earth and biosphere domains.

2.2.6 Generic information about EMBRC

EMBRC (European Marine Biological Resource Centre) is a distributed European RI which is set up to become the major RI for marine biological research, covering everything from basic biology, marine model organisms, biomedical applications, biotechnological applications, environmental data, ecology, etc. Having successfully completed a 3-year Preparatory phase (2011-2014), it is now in its Implementation phase (2014-2016), and operation is planned to start in 2016-2017[35].

The main purpose of EMBRC is to promote marine biological science and the application of marine experimental models in mainstream research by providing the facilities (lab space), equipment (e.g., electron microscopes, real time PCR machines, crystallography, lab equipment, equipment for accessing the environments such as research vessels, scientific divers, ROVs, etc.), expertise and biological resources that are necessary for carrying out biological research

In what concerns data, the role of EMBRC is to generate and make it available. It does not usually do any analysis of those data, unless it is contracted to do so. Data is usually generated through sensors in site in the sea or samples that are collected and then measured in the lab.

EMBRC would like to achieve several objectives through participation to ENVRIplus:

- Establishing collaborations with the environmental community, which would benefit from their environmental and ecological data.
- Developing and learning about new standards and best practices in terms of standards.
- Developing new standards within INSPIRE [8], which can be used for other datasets.
- Exploring new data workflows, which make use of marine biological and ecological data.
- Networking with other RIs.

[34] A complete report on ELIXIR generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+ELIXIR
[35] A complete report on EMBRC generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+EMBRC

2.2.7 Generic information about EMSO

EMSO (the European multidisciplinary seafloor & water column observatory) is a large-scale European Research Infrastructure in the field of environmental sciences for integrating data gathered from a range of ocean observatories. It tries to ensure open access to those data for academic researchers[36].

EMSO is based on a European-scale distributed research infrastructure of seafloor observatories with the basic scientific objective of long-term monitoring, mainly in real-time, of environmental processes related to the interaction between the geosphere, biosphere, and hydrosphere, including natural hazards. It is presently composed of several deep-seafloor observatories, which will be deployed on specific sites around European waters, reaching from the Arctic to the Black Sea passing through the Mediterranean Sea, thus forming a widely distributed pan-European infrastructure.

A goal of EMSO is to harmonise data curation and access, while averting the tendency for individual institutions to revert to idiosyncratic working practices after any particular harmonisation project has finished.

There is a notable overlap between EMSO and FixO3 data (i.e., some FixO3 data is provided within the EMSO infrastructure).

EMSO would like to obtain with the help of ENVRIplus better mechanisms for ensuring harmonisation of datasets across their distributed networks. Heterogeneous data formats increase the effort that researchers must invest to cross discipline boundaries and to compose data from multiple sources. Improved search is also desirable; currently expert knowledge is required, for example to be able to easily discover data stored in the MyOcean environment.

Furthermore, EMSO is investigating collaborations with data processing infrastructures such as EGI for providing resources for infrastructure-side data processing.

2.2.8 Generic information about EPOS

EPOS is a long-term plan for the integration of Research Infrastructures for Solid Earth Science in Europe. Its main aim is to integrate communities to make scientific discovery in the domain of solid earth science. EPOS integrates the existing (and future) advanced European facilities into a single, distributed, sustainable infrastructure (EPOS Core Services) taking full advantage of new e-science opportunities[37].

EPOS will allow the Earth Science community to make a significant step forward by developing new concepts and tools for accurate, durable, and sustainable answers to societal questions concerning geo-hazards and those geodynamic phenomena (including geo-resources) relevant to the environment and human welfare.

EPOS would need advice from ENVRIplus to improve the Interoperable AAAI system (federated & distributed), taking already existing software and make it available and scalable across communities.

[36] A complete report on EMSO generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+of+EMSO
[37] A complete report on EPOS generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+EPOS

2.2.9 Generic information about Euro-ARGO

The objectives of the Euro-ARGO are to optimise, sustain and improve the European contributions to ARGO and to provide a world-class service to the research (ocean and climate) and operational oceanography (Copernicus Marine Service) communities[38].

Euro-ARGO also aims at preparing the next phase of ARGO with an extension to deeper depths, biogeochemical parameters and observations of the Polar Regions.

The Euro-ARGO research infrastructure comprises a central facility and distributed national facilities. On May 2014, the EC awarded European legal status (ERIC) to the central facility. Euro-ARGO aims at developing a capacity to procure and deploy and monitor 250 floats per year and ensure that all the data can be processed and delivered to users (both in real-time and delayed-mode).

Euro-ARGO would like ENVRIplus to design and pioneer access to and use of a cloud infrastructure with services close to European research data to deliver data subscription services. Users would provide their criteria: time, spatial, parameter, data mode, update period for delivery (daily, monthly, yearly, near real time):

- The relevant data are pushed from the RI to the ENVRI cloud
- The data may be converted/transformed on the ENVRI computation grid
- The cloud account of the user is updated regularly with the new data provided above
- An accounting of data provision and data delivery is performed.

2.2.10 Generic information about EuroGOOS

EuroGOOS (European Global Ocean Observing System) is an international Not-for-Profit organisation. It promotes operational oceanography, i.e., the real time use of oceanographic information, and develops strategies, priorities and standards, which would enable its evolution at a European level. EuroGOOS is not an RI *per se*, but it has many members (40 institutes from 19 countries) who contribute to an RI for ocean observing[39].

EuroGOOS strives to improve the coordination between their different member research institutes. Another important role of EuroGOOS is that of facilitating access to data for their community.

Through participation to ENVRIplus, EuroGOOS would value:

- Learning about other European RIs and getting inspiration from them for deciding on the general objectives and services that they could provide at European level
- From a technological perspective, getting recommendations about the design of their common data system, including formats or data platforms and data treatments.
- Getting inspiration from RIs about ways to distribute the data to end users using applications which are more focused in this respect.

[38] A complete report on Euro-ARGO generic requirements can be found at http://envriplus.manageprojects.com/projects/requirements/notebooks/470/pages/43/comments/294/attachments/342/download

[39] A complete report on EuroGOOS generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+EuroGOOS

2.2.11 Generic information about FixO3

FixO3 (Fixed Open Ocean Observatory network) is an I3 research project that integrates oceanographic data gathered from a number of ocean observatories and provides open access to that data to academic researchers[40].

FixO3 seeks to integrate European open ocean fixed-point observatories and to improve access to these key installations for the broader community. These will provide multidisciplinary observations in all parts of the oceans from the air-sea interface to the deep seafloor. The FixO3 network will provide free and open access to *in situ* fixed-point data of the highest quality. It will provide a strong integrated framework of open ocean facilities in the Atlantic from the Arctic to the Antarctic and throughout the Mediterranean, enabling an integrated, regional and multidisciplinary approach to understand natural and anthropogenic change in the ocean.

Like EMSO, FixO3 requires from ENVRIplus better mechanisms for ensuring harmonisation of datasets across their distributed networks. Heterogeneous data formats make life difficult for researchers. Improved search is also desirable; currently expert knowledge is required, for example to be able to easily discover data stored in the MyOcean environment.

2.2.12 Generic information about IAGOS

The In-service Aircraft for a Global Observing System (IAGOS) is a European research infrastructure which implements and operates a global observation system for atmospheric composition by deploying autonomous instruments aboard a fleet of commercial passenger aircraft. It conducts long-term observations of atmospheric composition, aerosol and cloud particles on a global scale[41].

IAGOS provides freely accessible data for users in science and policy including air quality forecasting, verification of CO_2 emissions and Kyoto monitoring, numerical weather prediction, and validation of satellite products.

IAGOS expects through its participation in ENVRIplus to:

- Improve data discovery
- Metadata standardisation
- Interoperability
- Citation and DOI management

It also expects ENVRIplus to provide services for, citation, cataloguing and provenance.

2.2.13 Generic information about ICOS

The **Integrated Carbon Observation System (ICOS)** Research Infrastructure provides the long-term observations required to understand the present state and predict future behaviour of the **global carbon cycle and greenhouse gas emissions and concentrations[42].**

The objectives of ICOS are to provide effective access to a single and coherent data set to facilitate research into multi-scale analysis of greenhouse gas emissions, sinks and the processes that determine them, and to provide information, which is profound for research and for the understanding of regional budgets of greenhouse gas sources and sinks, their human and natural drivers, and the controlling mechanisms.

[40] Complete report on FixO3 generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+FixO3
[41] A complete report on IAGOS generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+IAGOS
[42] A complete report on ICOS generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+ICOS

ICOS expects ENVRIplus to provide access to tools and services in the fields of:

- Metadata curation (including "recipes" for cataloguing and storage)
- Data object identification and citation
- Collection and handling of provenance information

2.2.14 Generic information about INTERACT

The **International Network for Terrestrial Research and Monitoring in the Arctic (INTERACT) is a circumarctic network of 76 terrestrial field stations** in northern Europe, Russia, USA, Canada, Greenland, Iceland, the Faroe Islands and Scotland. INTERACT's main objective is to build capacity for identifying, understanding, predicting and responding to diverse environmental changes throughout the wide environmental and land-use envelopes of the Arctic. Together, the INTERACT stations host many thousands of scientists from around the world working in multiple disciplines, and INTERACT collaborates with many research consortia and international research and monitoring networks[43].

INTERACT is keen on working on homogenisation with other infrastructures. The most important bilateral benefits of NordGIS (the INTERACT geographical metadata information system[44]) versus ENVRIplus are the broad European standards exposed to NordGIS, as well as the grass-root requirements exposed to ENVRIplus.

INTERACT is open for new interactive solutions, and recognises that standards on how to turn primary data into data products suitable for OPEN dissemination need to be adopted.

2.2.15 Generic information about IS-ENES2

The **European Network for Earth System Modelling (IS-ENES2) is the second phase of the I3 infrastructure project for the European Network for Earth System Modelling (ENES)**. ENES gathers the community working on climate modelling. IS-ENES runs a distributed, federated data infrastructure based on a few (3-4) main data centres and various associated smaller ones[45].

IS-ENES encompasses climate models and their environment tools, model data and the interface of the climate modelling community with high-performance computing, in particular the European RI PRACE.

The requirements information provided to ENVRIplus refers to the climate-modelling community, to two data-dissemination systems (ESGF for project run time; LTA as long-term archiving), to CMIP5 as climate modelling data project 2010-2015 and CMIP6 2016-2021

By participating in ENVRIplus IS-ENES2 expects to obtain a better understanding of interdisciplinary use cases and end-user requirements, as well as advice for data catalogues to compare their model data with other data (e.g., observations).

2.2.16 Generic information about LTER

Long-Term Ecosystem Research (LTER) is an essential component of worldwide efforts to better understand ecosystems. This comprises their structure, functions, and long-term response to environmental, societal and economic drivers. LTER contributes to the knowledge

[43] A complete report on INTERACT generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+INTERACT
[44] http://www.nordgis.org/
[45] A complete report on IS-ENES2 generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+IS-ENES2

base informing policy and to the development of management options in response to the Grand Challenges under Global Change[46].

From the beginning (around 2003) the design of LTER-Europe has focused on the integration of natural sciences and ecosystem research approaches, including the human dimension. LTER-Europe was heavily involved in conceptualising socio-ecological research (LTSER). As well as LTER Sites, LTER-Europe features LTSER Platforms, acting as test infrastructures for a new generation of ecosystem research across European environmental and socio-economic gradients.

LTER Europe aims at providing information on ecosystem functioning and processes as well as related drivers and pressures for a whole ecosystem (e.g., a watershed). This information is very diverse in its technical formats (sensor Information, aerial photographs, field recordings, pictures, etc.). The purpose of the RI is to focus on harmonised methodologies and data products.

Due to the fragmented character of LTER Europe harmonised data documentation, real-time availability of data as well as harmonisation of data and data flows are the overarching goals for the forthcoming years. Currently, LTER Europe is developing a Data Integration Portal (DIP, e.g. including a time series viewer) and is working on the integration of common data repositories into their workflow system (including metadata documentation with LTER Europe DEIMS[47]). Therefore, based on the common reference model, ENVRIplus can provide development advice on those matters, which would be appreciated by LTER.

2.2.17 Generic information about SeaDataNet

SeaDataNet is a Pan-European infrastructure for ocean & marine data management, which provides on-line integrated databases of standardised quality. It develops an efficient distributed Marine Data Management Infrastructure for managing large and diverse data sets deriving from *in situ* and remote observation of the seas and oceans[48].

The on-line access to *in situ* data, metadata and products is provided through a unique portal interconnecting the interoperable node platforms constituted by the SeaDataNet data centres.

SeaDataNet would like to enhance the cross-community expertise on observation networks, requirements support and data management expertise by participating in ENVRIplus. More specifically, SeaDataNet would like technology support for cross-community (ocean, solid earth and atmosphere) visibility of information provided by SeaDataNet (platforms, metadata, datasets, vocabulary services), as well as expertise on interoperability services and standards.

2.2.18 Generic information about SIOS

SIOS, Svalbard Integrated Earth Observing System, is an integral Earth Observing System built on existing infrastructure in order to better understand the on-going and future climate changes in the Arctic[49].

Currently, SIOS is building a distributed data management system called SIOS Knowledge Centre, to develop methods for how observational networks are to be designed and implemented. The centre will lay the foundation for better-coordinated services for the international research community with respect to access to infrastructure, data and knowledge management, sharing of data, logistics, training and education.

[46] A complete report on LTER generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+LTER
[47] See http://data.lter-europe.net/deims/
[48] A complete report on SeaDataNet generic requirements can be found at http://envriplus.manageprojects.com/projects/requirements/notebooks/470/pages/66
[49] A complete report on SIOS generic requirements can be found at http://wiki.envri.eu/display/EC/General+requirements+for+SIOS

2.2.19 Analysis of Generic Information

The following tables summarise the information gathered from RIs, and allow for a parallel consideration of the replies collected on the generic questions from each RI. Each table presents our findings on one of the topics covered by the generic questions:

- The data lifecycle (Table 8)
- Data and services offered (Table 9)
- Data standards and software used (Table 10)
- Data management (Table 11)
- Data security and access (Table 12)
- Non-functional constraints (Table 13)
- Optimisation plans/ issues/ challenges (Table 14)
- Interactions with other RIs and initiatives (Table 15)

TABLE 8: SUMMARY OF THE DATA LIFECYCLE OF THE DIFFERENT RIS

RI	Data lifecycle[50]
ACTRIS	Data from stations [production] are transferred to a computational resource [acquisition] to perform first data quality assurance (QA) [curation] and store it afterwards to one of their topic-databases. Through the ACTRIS portal users can visualise and gain access to data separately [publishing]. A combination of DOI and code station is used for identification and citation purposes.
AnaEE	Data is distributed with different facilities belonging to different institutions. A portal with the ability to identify [curation] all data held in AnaEE is planned. Data centres are provided at a national level [publishing]. A European data-modelling centre is foreseen for backup purposes [processing].
EISCAT-3D	The EISCAT-3D operations centre collects [acquisition] data from the radar sites [production] and keeps the full data set (up to 20 PB) for three months [processing], after which all high-level data and 1% of the low-level data are archived at two redundant archives (the data centres) [curation]. Data access with authentication will be via an API and Web portal [publishing].
ELIXIR	ELIXIR connects bioinformatics activities across its national and international nodes into a sustainable European infrastructure for biological research data. ELIXIR research infrastructure provides data, compute, tools, standards and training for life sciences. Core data resources support all stages of data lifecycle [production, acquisition, curation, publication, processing].
EMBRC	Data through sea-sensors or laboratory samples [production and acquisition]. It does not do any analysis on the data, unless it is contracted to do so. It has two main types of data: A) Environmental data, which is mostly provided free of charge in public databases [publishing]. EMBRC acquires the data and submits it in raw form, depending on the project, to these national or international open access databases. B) Molecular data that is generated by the EMBRC or by its users, the scientists from member institutes or the users of EMBRC usually do some work on the data to curate it and, if part of a bigger project, they may perform some annotation and assembly [curation]. As part of the data policy, users who are scientists and generate molecular data will deposit it in an open access database [publishing].

[50] Data lifecycle stages are indicated in square brackets. Refer to Figure 3.

RI	Data lifecycle[50]
EMSO	Most observatories contribute data to the MyOcean/Copernicus Marine Environment Monitoring Service [acquisition]. Some data is also contributed to EMECO (the European Marine Ecosystem Observatory) [acquisition]. Institutions gather data and links to the data are made available online to researchers [acquisition and curation]. Many observatories store their own data independently of any dedicated data infrastructure; each has its own data management, data access services (typically via FTP) [acquisition]. EMSO data may be provided to researchers via different channels [publishing]. Each data domain has different policies, which any unified data infrastructure would have to accommodate. Different data types have different requirements.
EPOS	Each community decides how data is acquired, curated and made available [acquisition, curation, and publishing]. The data is backed up regularly in federated repositories [publishing]. The data is made available by the Integrated Core Services interface (website or portal) [publishing]. Metadata will be available in different formats. The data from Thematic Core Services has to be available reasonably quickly. PIDs are used for identification and citation purposes.
Euro-ARGO	Observations from ARGO floats are transmitted [production] to a Data Assembly Centre (DAC) [acquisition]. The DAC decodes, quality controls, and distributes the data [curation, publishing]. Once a month a DOI is attached to the ARGO dataset [curation]. On ARGO GDAC, the list of all ARGO data, metadata and technical files is continuously updated [publishing].
EuroGOOS	Data from sea-sensors [production] with an acquisition system is transferred to the user ashore [acquisition]. Satellite information comes through a receiving station [production], either from the satellite producers or from an agency. Forecast data comes from national monitoring programmes [production]. The data are collected [acquisition], catalogued and quality assured [curation] in data centres from different national research institutes. They make it available through web portals and discovery tools [publishing], and share data and information amongst themselves.
FixO3	Most observatories contribute data to the MyOcean/Copernicus Marine Environment Monitoring Service [acquisition]. Some data is also contributed to EMECO [acquisition]. Institutions gather data [acquisition] and links to the data [curation] are made available online to researchers [publishing]. Many observatories store their own data independently of any dedicated data infrastructure; each has its own data management and data access services (typically via FTP). FixO3 has no plans for infrastructure-side data processing.
IAGOS	Raw data is automatically transferred into the reception server [acquisition], and then validated automatically or manually [curation]. Validated and calibrated data is stored in a centralised database, from where, end-users access it via a web-based data portal [publishing].
ICOS	Three data types are stored: raw sensor data collected at the measurement stations [acquisition]; 2) aggregated and quality-controlled observational data produced by expert centres based on the sensor data [curation]; and 3) “elaborated” data produced by researchers external to ICOS, but based on ICOS observational data [curation]. All relevant data will be accessible through the Carbon Portal (CP) [publishing]. The CP will provide a "one-stop shop" for all ICOS data products.
INTERACT	The main information provided are the metadata regarding research, monitoring, and other activities at the stations. Monitoring data is so far not accessible to the public. In most cases, principal investigators own the research data. 80% of the information is kept at the station level.

RI	Data lifecycle[50]
IS-ENES2	Data is generated by climate modelling groups. Data is post-processed according to the standards and agreements of the inter-comparison project. Data is ingested at IS-ENES/ESGF data nodes [acquisition] and quality-controlled [curation]. Data is published to the IS-ENES/ESGF data infrastructure [publishing]. Publication makes metadata available and searchable and data accessible via IS-ENES portals (as well as via APIs) [publishing]. Important data products are replicated to dedicated long-term archival centres. Additional quality checks are run as a pre-requisite for DOI assignment and availability for DOI-based data citation.
LTER	Data acquisition and Quality control is done by the single sites and usually stored locally. DEIMS (data discovery Portal) provides a central repository of metadata on research sites, data sets and persons [curation]. Furthermore, it also provides a possibility to upload and share data files from basic and regular sites [publishing].
SeaDataNet	Large and diverse sets of data deriving from in situ and remote observation of the seas and oceans [acquisition]. The research lab or National Ocean Data Centre (NODC) provides quality controlled data in a delayed mode and curates the data in homogeneous files [curation]. Data are made available to users through a central portal [publishing], from which requests are re-directed to the NODC. When data access is restricted, requests are controlled by the data managers.
SIOS	Data is made available from each data management system in each organisation. Data is accessed through a data portal [publishing]. Users can access different observation streams from different organisations. Each organisation manages its own data. In future users will be able to access integrated data sets and services.

TABLE 9: SUMMARY OF THE DATA AND SERVICES OFFERED BY THE DIFFERENT RIS

RI	Data and services offered
ACTRIS	Data: Free and open access to all data and data products.
	Software for: quality assurance (QA) and data analysis.
	Instrumentation: TNA to different calibration centres and laboratories.
	Expertise: Calibration centres offer training and specific advice to users.
	Training: Training of operators and users in the field of atmospheric science.
AnaEE	**Data**: Data and data products are open.
	Services to: Exploitation of that data, and analytical and modelling services. Facilities to forecast the impact of global changes and feed into public policy.
EISCAT-3D	Data: Access to raw and analysed data is restricted according to the statutes of EISCAT, with an embargo time for the associate carrying out an experiment. Quick look overview data is open for non-commercial purposes.
	Software for: Reducing raw data into physical parameters. Visualisation of low-level data.
	Training: Courses on the use of their radar systems.
ELIXIR	Data and services covering all stages of data lifecycle.
EMBRC	Data: People may share data on a personal basis.
	Software: for population analysis of genetic and environmental data.
	Instrumentation: Number of buoys that are connected to various labs. It can also provide detectors and lab equipment.
	Expertise: in taxonomy and specific model organisms.
	Literature: libraries with grey literature at several stations.
EMSO	Services: data provision and the physical access necessary to run experiments.
	Software: for reformatting data not in the desired formats.
	Instrumentation: Facilities for ocean science academics to make requests for usage time on observatories. Technically access to deployed resources is limited to academia rather than industry.
EPOS	Data: Most of the data is available for any registered users.
	Software: for building their own systems and for analysing data.

RI	Data and services offered
	Instrumentation: Policies for regulating the TNA.
	Literature: Technical reports public through different project websites.
	Data: All data are public.
	Software: for ARGO floats data management.
	Expertise: It can be solicited to provide advice on various topics.
Euro-ARGO EuroGOOS	Data: Facilitates data access between its member institutes.
	Computational facilities: It does not have its own ships or platforms for HPC, but all of its member institutes do.
	Expertise: Marine domain, and understanding of end users and customers.
FixO3	Data: Working towards open access to all datasets.
	Instrumentation: TNA. Technically access to deployed resources is limited to academia rather than industry.
	Training: on the use of marine data infrastructures to acquire data.
IAGOS	Data: data open access for research purpose.
ICOS	Data: All data products are free. Aggregated "finalised" data sets via the ICOS Carbon Portal. Other types of data can be obtained via the Thematic Centres or from the PI of the observation stations.
	Computational facilities: Planning to set up facilities to produce elaborated data products based on observations.
	Expertise: various topics.
	Literature: The portal will host a database of all relevant scholarly publications.
INTERACT	Metadata: Metadata about research, monitoring and other station activities.
	Expertise: Best practice of grass-root level environmental monitoring and in-field research.
IS-ENES2	Services: Activities to provide future data near processing functionalities.
	Computing resources: computational facilities as part of the ESGF nodes and portals or IS-ENES portals interfacing with the IS-ENES data infrastructure.
	Expertise: On request, about their running environment.
	Literature: Website with RI information.
LTER	**Metadata**: Metadata on research sites (LTER Sites and LTSER Platforms) are centrally available using the LTER Europe DEIMS Site and dataset registry platform. Metadata on research sites don't have any restrictions in use. This includes information on literature.
	Data: some of the data shared by the different LTER sites are freely available. A common data policy and data sharing agreements will be developed in the upcoming years.
	Semantics: LTER is working on a common controlled vocabulary EnvThes as the basis for MD tagging and data tagging for data discovery and harmonisation.
	Software: Tools can be shared with the scientific community. DEIMS (extended by LTER Europe) can be shared freely
	Services: LTER is working on the implementation and use of data provision services (e.g. OGC services like WFS, WMS, WCS and SOS); metadata shared by using OGC CSW service endpoints (using ISO19115 MD model) and harvesting lists (using EML MD model); for part of the LTER network e.g. OGC SOS data services (e.g. TERENO) are already available.
SeaDataNet	Data: Most of them are freely available (water column). Some (mostly sea bed observation) are restricted but may be made available.
	Software: Software free: NEMO, MIKADO, ODV, DIVA, Oceanotron.
	Processing: Computing resources to host the datasets.
	Expertise in: data management, marine science and standardisation.
RI	**Data and services offered**
SIOS	**Data**: Access to observation streams via the data portal.
	Computing resources: May bring computing resource in at a later stage.

TABLE 10: SUMMARY OF THE DATA STANDARDS AND SOFTWARE USED BY THE DIFFERENT RIS

RI	Data standards and software used
ACTRIS	Data standards: NetCDF. CF 1.5 -Compliant format, NASA-Ames 1001.
	Software used: Linux servers, relational databases.
AnaEE	Data standards: OBOE, SSN ontology, OAI.
	Metadata standard: ISO 19115, compatible with the INSPIRE directive. OpenSearch and PANGAEA.
	Software used: Management tools for metadata.
EISCAT-3D	Data standards: HDF5, and a storage and catalogue system.
	Hardware used: FPGA, cluster computers
	Software: Open to EGI and AARC recommendations. Considering EUDAT services (the B2 family) and DIRAC for cataloguing, dCache, or iRODS for data backend..
ELIXIR	Data standards and software covering all stages of data lifecycle.
EMBRC	Data standards: GBIF.
	Metadata standards: MEDIN, INSPIRE directive.
	Software: Darwin Core.
EMSO	Data standards: NetCDF, ODV and SWE (being encouraged).
	Metadata standards: ISO and an extended version of Dublin-Core. It wants to be able to interoperate with WDS via long-term data archives like PANGAEA.
	Security standards: ISO 27001.
EPOS	Metadata standards: CERIF metadata model, RDF export, OAI-PMH, CKAN and OpenSearch. EPOS is open to EUDAT solutions.
	Software: community software libraries (e.g., dispel4py and Obspy).
Euro-ARGO	Data standards: NetCDF, CF, OpenDAP.
	Software used: Linux VM, Matlab scripts, C++ programs, perl scripts and scientific calculator (Caparmor).
	Hardware used: SGI cluster of 294 calculation nodes, with a total 2352 cores with a 27 teraflops capacity.
EuroGOOS	Hardware used: HPC cluster.
	Software used: Matlab, Fortran, Python, IDL, Fortran (proposed).
	Metadata standards: ISO.
FixO3	Data standards: NetCDF, ODV, OAI, SWE (being encouraged).
	Metadata standards: ISO, an extended version of Dublin-Core, ISO 19139 (being considered), OpenSearch and PANGAEA.
	Software used: Open source data-reformatting software.
IAGOS	Data standards: ASCII, NASA Ames and NetCDF format.
	Metadata standard: ISO 19115 and align with INSPIRE.
	Software used: FLEXPART, PostgreSQL and MongoDB databases, Matlab and open source libraries and tools.
ICOS	Data standards: CSV ASCII, NetCDF. Data can be provided in other formats
	Metadata standards: Text files (spreadsheets).
	Software used: different ICOS components used several software packages. Windows and Microsoft products. Considering Open Source products.
INTERACT	Software used: Java script libraries, PostgreSQL with PostGIS, UMNmapserver engine, apache webserver, Linux server.
IS-ENES2	Data standards: NETCDF-CF, OpenDAP data access protocol, Thredds.
	Metadata standards: ISO 19139 and Federated Solr/Lucene
	Software used: Globus FTP, CMOR, open source community components (security, catalogues, data access services, portal parts etc.). B2FIND, B2DROP are being considered.
	Hardware used: Heterogeneous and locally environments at sites according to site-specific constraints.
LTER	Data standards: Data are not standardised; using EnvEurope data reporting sheet for file-based data exchange. Some data provided as time series using SOS. Wide range of solutions for data storage (file based: CSV, NetCDF, Excel).

RI	Data standards and software used
	Metadata standards: Dataset: EML / ISO19115 / INSPIRE profile; RI documentation: DEIMS Sites MD model; Provenance: Prov-O (being considered).
	Software used Controlled vocabulary: PoolParty / TopBraid ; Data storage: B2SHARE (in testing phase), PostgreSQL, ORACLE, MySQL, Microsoft Access, ftp-repository, local data repositories; GIS: spatial databases, shapefiles, grids; Data files: CSV, Net-CDF, TXT, XLS, proprietary formats and software; Metadata: US-LTER DEIMS, Drupal 6 (migrating to Drupal 7), geonetwork; Data services (in evaluation and testing): geoserver, 52°North SOS Suite, etc.
SeaDataNet	Data standards: ASCII ODV, MEDATLAS, NetCDF and SEGY. Datasets format management: NEMO.
	Metadata standards: NEMO, ISO19115 and ISO19139 series, INSPIRE profile, OAI-PMH, OGC/CSW, OGC/WMS, OGC/WFS, OGC/SWE, OpenDAP.
	Software used: Geonetwork for CSW and ISO191*, 52North SOS and javascript client for SWE, Oceanotron for WMS, OPENDAP, SOS, WFS.
SIOS	

TABLE 11: SUMMARY OF DATA MANAGEMENT FOR THE DIFFERENT RIS

RI	Data management
ACTRIS	Covers all the topics except the optimisation.
AnaEE	Preparatory phase – Under development the data management; Integrated procedure both for data access and modelling is in place in AnaEE-France.
EISCAT-3D	Data management covers all stages of data lifecycle and is defined in the statutes.
ELIXIR	Data management covers all stages of data lifecycle.
EMBRC	Data policy in place.
EMSO	
EPOS	CERIF metadata model for data management and exploitation. At community level, users are free to use any standards as long as the data is accessible and discoverable by the ICS. EPOS does not have a data management plan yet.
Euro-ARGO	The data management procedures applied to ARGO floats, from real-time decoding to delayed mode procedures are described in ARGO data management document.
EuroGOOS	Cataloguing, processing and optimisation mostly.
FixO3	Data access policy defined.
IAGOS	
ICOS	ICOS doesn't have a data management plan but all of the topics are covered in the internal discussions and documentation of the RI.
INTERACT	INTERACT will establish a plan for managing metadata and data in the period 2016 – 2020.
IS-ENES2	CORDEX data management plan, CMIP6 data management preparation documents.
LTER	A common data policy and data management plan is in development as the outcome of the eLTER (H2020) project. Currently data policies and data management plans are defined by the different participating organisations. Core LTER data management functions cover currently the discovery of RI elements. Discovery and access to dataset across the different RI elements is under development.
SeaDataNet	Covers Identification and citation, curation, cataloguing and provenance.
SIOS	

TABLE 12: SUMMARY OF DATA SECURITY AND ACCESS FOR THE DIFFERENT RIS

RI	Data security and access
ACTRIS	Open data access without login. Some communities place restriction with

	password / login. ACTRIS has different timing to publish data based on the type of data. ACTRIS does not have any embargo period.
AnaEE	AnaEE data license attached to the data. Private companies may access platforms at a full cost rate with the possibility of controlling the dissemination of their data. Academic users are charged at marginal cost and then have to disseminate the data according to the AnaEE dissemination rules with academic embargo periods.
EISCAT-3D	Access to data is restricted according to the EISCAT statutes (Blue Book). There is an embargo time for use exclusively by the experimentPI, after which the data are open to all EISCAT members. Quick-look data products are openly accessible for non-commercial purposes. EISCAT-3D does not have security or privacy issues in general, but there is one sensitive issue—the incidental detection of satellites in orbit, not all of which are white-listed for public tracking
ELIXIR	Open access to all publicly available data and secure controlled access to sensitive human data.
EMBRC	Open data access policy. Some timing restrictions depending on the purposes of the originating research. Private sector users retain the IPR of their generated data.
EMSO	General open and free data access policy, but some Copernicus data is password protected. Data tracking retrieval may be implemented.
EPOS	Login and password access with all the existed credentials. EPOS has 85 % of the data open. Only a small amount of data is not open, which is subject to an embargo period (6 months) or paid data.
Euro-ARGO	All ARGO data are public. IFREMER operates the Euro-ARGO data distribution. They follow the security procedures of IFREMER IT infrastructure.
EuroGOOS	Free and open data access. The use of such free data by research institutes by exchange and copyright agreements. Some embargo period for publication periods. EuroGOOS do not have set procedures for security and access. Metrics about the end users can be obtained directly from their IP addresses.
FixO3	General open and free data access policy, but some Copernicus data is password-protected. Single sign-on process before any data is accessed for accounting reasons.
IAGOS	Data is open but registration is needed (password control). It needs to be improved to use (e.g., a certificate-based approach).
ICOS	Single-sign-on system to control and monitor user identification, authorisation and authentication for data and computational resources that require this. Other ICOS components (Thematic Centres) are using systems that are local to their host institutes for these purposes.
INTERACT	Four levels through OPEN public access, PI editorial level, station management level, and level of system management.
IS-ENES2	Single sign on across multiple portals as well as authorisation based on membership of various “projects”. CORDEX data are available for both commercial and research purposes. Some modelling centres restrict their data use to “non-commercial research and educational purposes.”
LTER	Free access to metadata on RI elements and datasets. Data are free if collected in European funded research projects but local restrictions may be applied.
SeaDataNet	A user directory with self-registration provided. Authentication is managed via a Central Authentication Service. Some data are free.
SIOS	

TABLE 13: SUMMARY OF NON-FUNCTIONAL CONSTRAINTS FOR THE DIFFERENT RIS

RI	Non-functional constraints
ACTRIS	Computational environment costs.
AnaEE	
EISCAT-3D	Administrative constraints from funding agencies.
ELIXIR	Rapid exponential data growth and rapid uptake of biomolecular methods.
EMBRC	Maintenance and operational costs.

EMSO	
EPOS	Different non-functional constraints depending on the ICS or TCS layer, like maintenance, capital, and operational costs.
Euro-ARGO	Capital costs, maintenance costs, operational costs, security, computational environment in which your software runs.
EuroGOOS	
FixO3	Difficult to normalise data management costs.
IAGOS	Maintenance costs supported by AERIS.
ICOS	Capital costs, maintenance costs, operational costs, security, privacy.
INTERACT	Will be operated and managed by the INTERACT field-stations themselves, and is hence quite robust.
IS-ENES2	Annual operating cost of the infrastructure is estimated to be of 1560 k€.
LTER	Long-term preservation of data; common data policy; implementation of data services across the RI; maintenance and operation costs; security, privacy.
SeaDataNet	Long-term preservation of data, privacy.
SIOS	

TABLE 14: SUMMARY OF OPTIMISATION PLANS/ ISSUES/ CHALLENGES FOR THE DIFFERENT RIS

RI	Optimisation plans / Issues / Challenges
ACTRIS	Data visualisation, data provision, inter-operability between data centre nodes.
AnaEE	
EISCAT-3D	Workflow definitions. Data access with searching and visualisation. Interoperability with other RIs and instruments via virtual observatories.
ELIXIR	Data interoperability across research domains.
EMBRC	Common standards and workflows. Harmonisation of data between labs. Backup system. Maintenance of software and their integration into a single platform.
EMSO	Data inter-operability across distributed networks and data search.
EPOS	Improve the Interoperable AAI system, taking already existing software and make it available and scalable across communities.
Euro-ARGO	
EuroGOOS	Data assimilation.
FixO3	Harmonisation of data formats and protocols across their distributed networks, as well harmonise data curation and access.
IAGOS	Data processing.
ICOS	Data and Metadata storage.
INTERACT	Moving into the realm of handling actual data concerning 76 active field-stations.
IS-ENES2	Share best practices as fast as new nodes integrate the RI federation. Data near processing. Handling volume and distribution of data: Replication, Versioning. Providing related information for data products (provenance, user comments, usage, detailed scientific descriptions needed for usage).
LTER	Online data documentation, data harmonisation and access to distributed data services.
SeaDataNet	Data policy to involve data providers in the publication of their own datasets.
SIOS	

TABLE 15: SUMMARY OF INTERACTIONS WITH OTHER RIS AND INITIATIVES

RI	Interactions with other RIs and Initiatives
ACTRIS	IAGOS and ICOS (from ENVRIplus); AeroCom (Outside EU).
AnaEE	ICOS, LifeWatch and LTER.

EISCAT-3D	COOP+, DIRAC, EGI, EUDAT, Nordic Tier 1, RDA
ELIXIR	A few examples include EMBRC, LifeWatch and SeaDataNet.
EMBRC	
EMSO	FixO3
EPOS	Might have interactions with other RIs to access some computational services.
Euro-ARGO	
EuroGOOS	RIs for ocean observing from across Europe.
FixO3	EMSO
IAGOS	Interested in collaboration with ACTRIS and ICOS.
ICOS	
INTERACT	EUDAT, CLINF
IS-ENES2	
LTER	EUDAT, ICOS, LifeWatch, EU-BON, GEOBON, AnaEE, ENVRIplus, ILTER, InterAct, TERN (Austrialia), SAEON (South Africa), NEON (US)
SeaDataNet	Eurofleet, EuroARGO, ESONET, FixO3 and JERICO.
SIOS	INTERACT, EMSO, ICOS and GEM.

Table 16 below summarises expectations of the RIs as to what they will gain by participating in ENVRIplus.

TABLE 16: SUMMARY OF RI'S EXPECTATIONS FROM PARTICIPATING IN ENVRIPLUS.

RI	Expectations from ENVRIplus
ACTRIS	Planning and managing the activity of sensors. Developing understanding of how instruments work in extreme conditions. Improving the capabilities of small sensors.
AnaEE	Homogenous approach on Identification and citation and on cataloguing across RIs. Interoperability between models and data. Quality control of data produced by platforms.
EISCAT-3D	Selecting open and well-documented tools. Increased interoperability between domains.
ELIXIR	Establishing a closer collaboration with environmental Research Infrastructures (RIs) and improving their access to life science data. Ultimately, better interoperability and discoverability of environmental and life science data by users across atmospheric, marine, solid earth and biosphere domains.
EMBRC	Establishing collaborations with the environmental community. Developing and learning about new standards and best practices. Developing new standards within INSPIRE, which can be used for other datasets. Exploring new data workflows, which make use of marine biological and ecological data. Networking with other RIs.
EMSO	Ensuring harmonisation of datasets across their distributed networks. Handling heterogeneous data formats. Improving search is also desirable.
EPOS	Improving the Interoperable AAAI system, taking already existing software and make it available and scalable across communities.
Euro-ARGO	Designing and pioneering access to and use of a cloud infrastructure with services close to European research data to deliver data subscription services.
EuroGOOS	Learning about other European RIs to decide on the general objectives and services. Recommendations about the design of their common data systems and data distribution to end-users.
FixO3	Harmonisation of datasets across distributed networks. Heterogeneous data formats to enhance cross-community collaboration. Improved search is also desirable.
IAGOS	Improving data discovery. Metadata standardisation. Interoperability. Citation and DOI management.
ICOS	Metadata curation, including “recipes” for cataloguing and storage. Data object identification and citation. Collection and handling of provenance information.

RI	Expectations from ENVRIplus
INTERACT	Recommendations about how to turn primary data into data products need to be adopted. Metadata and data standardisation at all levels. Homogenisation with other RIs.
IS-ENES2	Better understanding of interdisciplinary use cases and end-user requirements, as well as advice for data catalogues to compare their model data with other data.
LTER	Support on data curation and data object identification (especially on the aspect of dynamic data series and identification of results from data queries (e.g. data services); technical support on optimisation of data flows and annotation (e.g. integrating of a data repository, data integration portal).
SeaDataNet	Enhancing the cross-community expertise on observation networks, requirements support and data management expertise. Technology support for cross-community visibility of information provided by SeaDataNet, as well as expertise on interoperability services and standards.
SIOS	

2.3 Gathered specific topic information

Each of the topics into which requirements gathering has been partitioned is presented below by the relevant topic leader, see Table 2. They introduce their topic and then analyse the requirements information gathered. That primary information, updated after this report was produced, can be found in the ENVRI Community Wiki[51].

2.3.1 Identification and Citation Analysis

Introduction

Identification of data (and associated metadata) throughout all stages of processing is really central in any RI. This can be ensured by allocating unique and persistent digital identifiers (PIDs) to data objects throughout the data processing life cycle. The PIDs allow unambiguous references be made to data during curation, cataloguing and support provenance tracking. They are also a necessary requirement for correct citation (and hence attribution) of the data by end users, as this is only possible when persistent identifiers exist and are applied in the attribution.

Environmental research infrastructures are often built on a large number of distributed observational or experimental sites, run by hundreds of scientists and technicians, financially supported and administrated by a large number of institutions. If this data is shared under an open access policy it becomes therefore very important to acknowledge the data sources and their providers. There is also a strong need for common data citation tracking systems that allow data providers to identify downstream usage of their data so as to prove their importance and show the impact to stakeholders and the public.

Identification

The survey found a large diversity between RIs regarding their practices. Most are applying file-based storage for their data, rather than database technologies, which suggests that it should be relatively straightforward to assign PIDs to a majority of the RI data objects. A profound gap in knowledge about what persistent and unique identifiers are, what they can be used for, and best practices regarding their use, emerged. Most identifier systems used are based on handles (DOIs from DataCite most common, followed by ePIC PIDs), but some RIs rely on formalized file names. While a majority see a strong need for assigning PIDs to their "finalized" data (individual files and/or databases), few apply this to raw data, and even fewer to intermediate data – indicating PIDs are not used in workflow administration. Also, metadata objects are seldom assigned PIDs.

[51] https://wiki.envri.eu/display/EC/ENVRI+RI+Requirements

Costs for maintaining PIDs are typically not treated explicitly. Assignment of PIDs to other forms of data, such as continuous time series, is discussed in Sections 3.2.

Citation

NOTE: RIs were asked to characterise their "designated user community" needs, but most responded with RI-centric requirements. This may be because there was not sufficient opportunity to directly communicate with users. Normally, their highest priority is to improve their productivity, in this case by having as much of the data identification and citation automated – see Sections 3.2 and 4.2.5.

Currently, users refer to data sets in publications using DOIs if available, and otherwise provide information about producer, year, report number etc. either in the article text or in the References section. A majority of RIs feel it is absolutely necessary to allow unambiguous references to be made to specified subsets of datasets, preferably in the citation, while few find the ability to create and later cite collections of individual datasets is important. Ensuring that credit for producing (and to a lesser extent curating) scientific data sets is "properly assigned" is a common theme for all RIs – not the least because funding agencies and other stakeholders require such performance indicators, but also because individual PIs want and need recognition of their work. Connected to this, most RIs have strategies for collecting usage statistics for their data products, i.e., through bibliometric searches (quasi-automated or manual) from scientific literature, but thus often rely on publishers indexing also data object DOIs.

Conclusion

The use of persistent and unique identifiers for both data and metadata objects throughout the entire data life cycle needs to be encouraged, e.g., by providing training and best-use cases. There is strong support for promoting "credit" to data collectors, through standards of data citation supporting adding specific sub-setting information to a basic (DOI-based) reference. Demonstrating that this can be done easily and effectively, and that data providers can trust that such citations will be made, will be a priority, as it will lead to adoption and improvement of citation practices.

2.3.2 Curation Analysis

Curation, cataloguing and provenance are closely related and all three topics have metadata element requirements that overlap considerably with one another. Hence, they are often considered together.

At present there is available information based on the questionnaires used by the go-betweens for 7 RIs.

Curation of Datasets

Briefly, the responses range from 'no curation or plans' to detailed information on metadata formats used. None referred to a data management plan although it is known to be an essential component within EPOS. Many RIs have elements of a DMP in place in their statutes, but these may not be formulated as a DMP yet.

Only one RI mentioned OAIS (the ISO/IEC 14721 standard for curation although it is not much used and when it is the implementations are very varied since it is really an overview architecture rather than a metadata standard).

With regard to the metadata standards used or required by the RIs:

- Several use ISO19115/INSPIRE but this does not really provide much curation information.
- One uses CERIF, which does provide curation information.
- One uses Dublin Core, which does not provide curation information.

Curation of Software

None mentioned metadata covering software and its curation except EPOS (using CERIF). A few use Git to manage software. Most have no curation of software nor plans for this.

Curation of Resources used (computers, equipment, detectors)

None mentioned metadata for curation of information on these assets.

Curation of User information

None mentioned metadata for curation of user information although it is known that EPOS uses CERIF for this purpose (and will use the metadata for driving AAAI and collaborative working).

Conclusion

Possibly due to the early stage of some RIs, or due to interacting with RIREPs who were not informed about curation (it is often dealt with by a small group of specialists) the requirements for curation were not made explicit, for example, none of the RIs (who responded) has appropriate metadata and processes for curation. It is known that EPOS has plans in place and there are indications of such planning for some of the others. Since curation often underpins validation of the quality of scientific decisions and since environmental sciences observe phenomena that do not repeat in exactly the same form, the profile of curation needs raising. This should be attempted by awareness raising programmes, beginning with discussions during ENVRIWeek spring 2016. If it transpires that there is a need then a best practice guide should be developed on curation, provenance and cataloguing, which should be offered to all RIs.

2.3.3 Cataloguing Analysis

Regarding the possible items to be managed in catalogues, the RIs have shown interest in:

- **Observation systems and lab equipment**: most RIs manage equipment which requires management (scheduling, maintenance, monitoring, ...) and some of them are managing or would like to manage this with an information system. Some are already using a standardised approach (OGC/SWE, SSN).
- **Data processing procedures and systems, software:** a very few or none mentioned an interest to support this in a catalogue. We observe, however, that this may be necessary as part of the provision for provenance and as an aid for those developing or formalising new methods.
- **Observation events**: not explicitly mentioned as a requirement most of time. Again, this need may emerge when provenance is considered.
- **Physical samples**: mentioned by a few especially in bio-diversity field.
- **Processing activities:** not explicitly mentioned.
- **Data products or results**: widely mentioned as being done by existing systems (EBAS, EARLINET, CLOUDNET, CKAN, MAdrigal, DEIMS). Widely standardised (ISO/IEC 191XX). Compliance is sometimes required with the INSPIRE directive; support for this in the shared common subsystems would prove beneficial. Once with WIS.
- **Publications**: widely mentioned. However very few manage the publications on their own. Links for provenance between publications and datasets are quite commonly required.
- **Persons and organisations:** not explicitly mentioned. However, this is reference information, which is required for the other described items (datasets, observation systems, etc.) and for provenance (contact points).
- **Research objects or features of interest**: mentioned once as feature of interest (airports for IAGOS).

As a consequence, the following three categories of catalogues are cited in the requirements collection:

- **Reference catalogues**, which are not developed by ENVRIplus or within RIs but are pre-existing infrastructures containing reference information to be used. They can also be considered as gazetteer, thesaurus or directories. Among them we consider catalogues for:
  - people and organisations[52],
  - publications,
  - research objects, features of interest.
- **Federated catalogues**, which are pre-existing and partly harmonised in an RI but could be federated by ENVRIplus. Among them we consider:
  - data products or results
  - results, observation systems and lab equipment. It would be helpful to promote the management of metadata to improve provenance.
  - physical samples
  - data processing procedures, systems and software components metadata management should be promoted to improve provenance
- **Finally, activity records, observation events, processing activities, usages logs can be considered**. They should be provided by RIs and harmonised at the ENVRIplus level to link together the catalogues and fulfil the provenance requirements. The tracking of usage of datasets in scientific papers is widely mentioned by RIs. These activity records need to be harmonised in ENVRIplus.

2.3.4 Processing Analysis

Data Processing or Analytics is an extensive domain including any activity or process that performs a series of actions on dataset(s) to distil information [Bordawekar 2014]. It is particularly important in scientific domains especially with the advent of the 4th Paradigm and the availability of "big data" [Hey 2009]. It may be applicable at any stage in the data life cycle from QA and event recognition close to data acquisition to transformations and visualisations to suit decision makers as results are presented. Data analytics methods draw on multiple disciplines including statistics, quantitative analysis, data mining, and machine learning. Very often these methods require compute-intensive infrastructures to produce their results in a suitable time, because of the data to be processed (e.g., huge in volume or heterogeneity) and/or because of the complexity of the algorithm/model to be elaborated/projected. Moreover, these methods being devised to analyse dataset(s) and produce other "data"/information (than can be considered a dataset) are strongly characterised by the "typologies" of their inputs and outputs. In some data-intensive cases, the data handling (access, transport, IO and preparation) can be a critical factor in achieving results within acceptable costs.

In fact, when analysing the needs of Research Infrastructures involved in ENVRIplus we focused on collecting four major aspects that characterise each RI's data processing needs:

- **Input**, i.e., what are the characteristics of the dataset(s) to be processed? This includes dataset(s) typologies, volume, velocity, variety/heterogeneity, and access methods;
- **Analytics**, i.e., what are the characteristics of the processing tasks to be enacted? This includes computing needs quantification, implementation aspects including programming languages, standards and re-use potential;
- **Output**, i.e., what are the characteristics of the products resulting from the processing? This includes typologies, volume, variety, variety/heterogeneity, and availability practices;

[52] For the purposes of an ERIC, and RI may need a formal list of investigators for quality control. Shared support for such lists may prove helpful.

- **Statistics**, i.e., what are the scientific motivations leading to the identification of the specific data processing envisaged by a community. This includes aspects related to data collection and hypothesis generation.

Each of these are summarised below.

Input

As largely expected, RIs' needs with respect to dataset(s) to be processed are quite diverse because of the diversity in the datasets that they deal with. Dataset(s) and related practices are diverse both across RIs and within the same RI. For instance, in EPOS there are many communities each having its specific typologies of data and methodologies (e.g., FTP) and formats (e.g., NetCDF, text) for making them available. Time series and tabular data are two very commonly reported types of dataset to be processed yet they are quite abstract. In what concerns "volume", dataset(s) vary from a few KBs to GBs and TBs. In the large majority of cases dataset(s) are made available as files while few infrastructures have plans to make or are making their data available through OGC services, e.g., ACTRIS.

The need to homogenise and promote state-of-the-art practices for data description, discovery and access is of paramount importance to provide RIs with a data processing environment that makes it possible to easily analyse dataset(s) across the boundaries of RI domains.

Analytics

When moving to the pure processing part, it emerged that RIs are at diverse levels of development and that there is a large heterogeneity. For instance, the programming languages currently in use by the RIs range from Python, Matlab and R to C, C++, Java, and Fortran. The processing platforms range from the 3 Linux servers in the case of ACTRIS to HPC approaches exploited in EPOS. No major issues emerged with respect to licences. Software in use or produced tends to be open source and freely available. In the majority of cases there is almost no shared or organised approach to make available the data processing tools systematically both within the RI and outside the RI. One possibility suggested by some RIs is to rely on OGC/WPS for publishing data processing facilities.

Some care needs to be taken balancing the benefits of common solutions with the need to support a wide range of working practices well – we return to this in Section 4.2. The platform should be "open" and "flexible" enough to allow (a) scientists to easily plug-in and experiment with their algorithms and methods without bothering with the computing platform, (b) service managers to configure the platform to exploit diverse computing infrastructures, (c) third-party service providers to programmatically invoke the analytics methods, and (d) to support scientists executing existing analytic tasks eventually customising/tuning some parameters without requiring them to install any technology or software.

Output

In essence, we can observe that the same variety characterising the input is there for the output also. I this case, however, it is less well understood that there is a need to make these data available in a systematic way, including information on the entire process leading to the resulting data. In the case of EMBRC it was reported that the results of a processing task are to be made available via a paper while for EPOS it was reported that the dataset(s) are to be published via a shared catalogue describing them by relying on the CERIF metadata format.

In many cases, but by no means all, output resulting from a data processing task should be "published" to be compliant with Open Science practices. A data processing platform capable of satisfying the needs of scientists involved in RIs should offer an easy to use approach for having access to the datasets that result from a data processing task together. As far as possible it should automatically supply the entire set of metadata characterising the task, e.g., through the provenance framework. This would enable scientists to properly interpret the results and reduce the effort needed to prepare for curation. In cases where aspects of the information are

sensitive, could jeopardise privacy, or have applications that require a period of confidentiality, the appropriate protection should be provided.

Statistical

Only a minority of the RIs within ENVRIplus responded to the statistics questions within the processing requirements gathering. We know from the ENVRI project that LifeWatch had the support of a wide range of statistical investigations, not just biodiversity, as part of its mission. Unsurprisingly given the diversity of the component RIs, there were a variety of different attitudes to the statistical aspects of data collection and analysis. One RI (IS-ENES-2) felt that data analysis (as opposed to collection) was not their primary mission whereas for others (e.g., within EMBRC researchers at the University of St Andrews) reaching conclusions from data is very much their primary purpose.

As environmental data collection is the primary aim of many of the RIs it appears that day-to-day consideration of potential hypotheses underlying data collection is not undertaken. Hypothesis generation and testing is for scientific users of the data and could take many forms. However, some RIs (e.g., LTER and ICOS) stressed that general hypotheses were considered when the data collection programmes and instruments were being designed especially if the data fed into specific projects. Hypotheses could be generated after the fact by users after data collection and indeed this would be norm if data collection is primary a service to the wider scientific community.

RIs can be collecting multiple streams of data often as time series, thus there is the potential to undertake multivariate analysis of the data. Again unsurprisingly given the diversity of science missions, there was no consistency in approaches. Data could be continuous and discrete, be bounded by its very nature or have bounds enforced after collection. Data sets are potentially very voluminous; total data sets with billions of sample points might be generated. Most analysers will be engaging in formal testing of hypotheses rather than data mining although the latter was not necessarily ruled out. Many RIs had or are going to implement outlier or anomaly detection on their data.

Again unsurprisingly given the potential uses for the data, a variety of statistical methods can be undertaken. RIs did not feel restricted to working solely within either a frequentist or Bayesian framework. Much of the data collected takes the form of time series.

The current mission of ENVRIplus will address the aspects of data collection, preparation and integration that should provide a context for such statistical approaches. The integration of tools and statistical methods, and their mapping onto platforms, should be supported in an appropriate virtual research environment or science gateway. This requires collaborative R&D building on experience from the EU project Biodiversity Virtual e-Laboratory (BioVeL)[53]. This would fully integrate statistical analysis tools with the data handling, and map the processing tasks automatically to appropriate data-intensive subsystems and computational resources. The sustainable path, which would also promote international exchanges of environmental-data analysis methods, would benefit from collaboration with organisations such as the NSF-funded Science Gateway Institute[54]. This environmental-analytical virtual e-Laboratory kit is a good example of a candidate common subsystem, where the balance of a core used by many RI communities with tailoring to support specialised working practices would need careful investigation. Providing such an integrated combination of data lifecycle support with easily activated and steered analysis and visualisation tools will improve researcher productivity by removing many hurdles they have to get over today. This will accelerate discovery and evidence production, but it will also boost those who take those results and present them to decision makers. This will interact with the arrangements for federation support –see Section 4.2.3.

[53] https://www.biovel.eu/
[54] http://sciencegateways.org/ with relevant publications at http://iwsg-life.org/site/iwsglife/publications

2.3.5 Provenance Analysis

In order to correctly use and reuse and interpret data within a research infrastructure, and cross research infrastructures the data's evolutionary history must be known in detail. This is especially crucial in environmental sciences in order to understand changes through history from billions years ago up to recent and current (up to picoseconds) history. The required combinations span time scales, span regional scales, span species scales and a wide range of observing and sampling strategies. This inevitably requires many data pipelines, each based on their own research and observation practices. As biological and environmental systems are intricately intertwined, these then need to be brought together. Hence, the criticality of provenance to validate the quality of the ultimate products.

This history covers all the steps of the data lifecycle:

- **data production and acquisition**: detailed information about scientific question and investigation design, observation or measurement methods, measurement devices and so forth is needed,
- **data curation**: exact description of QA measurements (flagging and annotation of data), metadata to assist with correct (future) interpretation and data replication,
- **data publication**: which data were accessed, which data are not accessible (the selection of data can strongly influence any further results of data processing), which query was carried out and when,
- **data processing**: which method was used for further processing (aggregation of data, transformation, modelling)
- **data interpretation**: scientific knowledge drawn out of data plus the theories behind.

It is important to point out, that knowing the evolutionary history of data – and at very different time scales – is important for any use and reuse of data: use and reuse within institutes (reuse some years after the investigation was made, reuse by other persons within institutes), use and reuse within Research Infrastructure and cross Research Infrastructures.

Inter alia provenance can help to avoid undetected duplication (production or storage) of datasets.

In order to have information on those steps, their description has to be tracked as the so called "data provenance" and made available to data users.

The requirements questionnaire with focus on provenance intended to collect whether provenance was already considered in each RI's data lifecycle and if so which system is in use. If this was as yet not implemented, the next set of questions is grouped about the RI's possible interest in provenance tracking: which type of information should be tracked, which standards to rely on and finally which sort of support is expected from ENVRIplus.

Most RIs already consider provenance data as essential and are interested in using a provenance recording system. Among all of the nine RIs who gave feedback about provenance only two already had a data provenance recording system embedded in their data processing workflows. EPOS uses the dispel4py workflow engine in VERCE, which is based on and is able to export to PROV-O whereas in future it is planned to use the CERIF data model and ontology instead. IS-ENES2 instead does not specify which software solution is applied but mentions: the use of community tools to manage what has been collected from where, and what is the overall transfer status to generate provenance log files in workflows. Some, such as SeaDataNet and Euro-ARGO, interpret provenance as information gathered via metadata about the lineage data with tools like Geonetwork based on metadata standards like ISO19139, but the information gathered is not sufficient to reproduce the data as the steps of processing are not documented in enough detail. Other RIs, such as ICOS and LTER, are already providing some provenance information about observation and measurement methods used within the metadata files but are aware that a real tracking tool still needs to be implemented. IAGOS is using the versioning

system GIT for code but not for the data itself. A versioning system can only be seen as a part of the provenance information sought.

On which information is considered to be important, the answers range from versioning of data to the generation of data and modification of the data as well as on who, how and why data is used. So there seems to be two interpretations about what provenance should comprise: should it enable the community to follow the data 'back in time' and see all the steps that happened from raw data collection, via quality control and aggregation to a useful product, or should it enable the data provider as a means of tracking the usage of the data, including information about users in order to understand the relevance of the data and how to improve their services? These two roles for metadata may be served by the same provenance collecting system. The provenance data is then interpreted via different tools or services.

Regarding the controlled vocabularies used for the descriptions of the steps for data provenance, some RIs already use research specific reference tables and thesauri like EnvThes and SeaDataNet common vocabularies.

There is a big interest among the RIs to get clear recommendations from ENVRIplus about the information range provenance should provide. This includes drawing an explicit line between metadata describing the 'dataset' and provenance information. Also it should be defined clearly whether usage tracking should be part of provenance.

It is considered as being very important to get support on automated tracking solutions and or provenance management APIs to be applied in the specific e-science environments. Although there are some thesauri already in use there is a demand for getting a good overview of the existing vocabularies and ontologies that are ready to use or that need to be slightly adapted for specific purposes.

There is a strong relationship between the task of *identification* of data and the *provenance* task as there must be a direct link between the data and its lineage that can be followed by the interested user. Provenance tracking is also an important feature for optimisation. The connections with curation and cataloguing is evident which also becomes clear in the IC_2 Provenance implementation case[55] which aims amongst others at defining a minimum information set that has to be tracked, finding a conceptual model for provenance which conforms to the needed information, maps existing models to the common model and finds a repository to store the provenance information.

2.3.6 Optimisation Analysis

Introduction

Environmental science now relies on the acquisition of great quantities of data from a range of sources. The data might be consolidated into a few very large datasets, or dispersed across many smaller datasets; it may be ingested in batch or accumulated over a prolonged period. Although efforts are underway to store data in common data stores, to use this wealth of data fast and effectively, it is important that the data is both optimally distributed across a research infrastructure's data stores, and carefully characterised to permit easy retrieval based on a range of parameters. It is also important that experiments conducted on the data can be easily compartmentalised so that individual processing tasks can be parallelised and executed close to the data itself, so as to optimise use of resources and provide swift results for investigators.

We are concerned here with the gathering and scrutiny of requirements for optimisation. More pragmatically, we are concerned with how we might develop generically applicable methods by which to optimise the research output of environmental science research infrastructures, based on the needs and ambitions of the infrastructures surveyed.

[55] https://wiki.envri.eu/display/EC/Use+Cases

Perhaps more so than the other topics, optimisation requirements are driven by the specific requirements of those other topics, particularly processing, since the intention is to address specific technical challenges in need of refined solutions, although implemented in a way that can be generalised to more than one infrastructure. For each part of an infrastructure in need for improvement, we must consider:

- What does it mean for this part to be optimal?
- How is optimality measured—do relevant metrics already exist as standard?
- How is optimality achieved—is it simply a matter of more resources, better machines, or is there need for a fundamental rethink of approach?
- What can and cannot be sacrificed for the sake of 'optimality'? For example, it may be undesirable to sacrifice ease-of-use for a modest increase in the speed at which experiments can be executed.

More specifically, we want to focus on certain practical and broadly universal technical concerns:

- What bottlenecks exist in the functionality of (for example) storage, data management subsystems, e.g., file systems or databases, access and delivery of data, data processing, and workflow management?
- What are the current peak volumes for data access, storage and delivery for parts of the infrastructure?
- What is the (computational) complexity of different data-processing workflows?
- What are the specific quality (of service, of experience) requirements for data handling, especially for real-time data handling?

Overview and summary of optimisation requirements

Many optimisation problems, whether explicitly identified as such by RIs, or implicit in the requirements for other topics, can be reduced down to ones of *data placement*, often in relation to specific services, resources or actors.

- Is the data needed by researchers available from a location such that they can be easily identified, retrieved and analysed, in whole or in part?
- Is it feasible to perform analysis on data without substantial additional preparation, and if not, what is the overhead in time and effort required to prepare the data for processing?

This latter question in particular relates to the notion of *data staging*, whereby data is placed and prepared for processing on some computational service (whether that is provided on a researcher's desktop, within an HPC cluster or on a web server), which in turn concerns the further question of whether data should be brought to where they can be best computed, or instead computing tasks be brought to where the data currently reside. Given the large size of many RI's primary datasets, bringing computation to data is appealing, but the complexity of various analyses also often requires supercomputing-level resources, which require the data be staged at a computing facility such as are brokered in Europe by consortia such as PRACE. Data placement is reliant however on data accessibility, which is not simply based on the existence of data in an accessible location, but is also based on the metadata associated with the core data that allows it to be correctly interpreted; it is based on the availability of services that understand that metadata and can so interact (and transport) the data with a minimum of manual configuration or direction.

Reductionism aside, the key performance indicator used by most RIs is researcher productivity. Can researchers use the RI to efficiently locate the data they need? Do they have access to all the support available for processing the data and conducting their experiments? Can they replicate the cited results of their peers using the facilities provided? This raises yet another question: how does the service provided to researchers translate to requirements on data placement and infrastructure availability?

This is key to intelligent placement of data—the existence of constraints that guide (semi-) autonomous services by conferring an understanding of the fundamental underlying context in which data placement occurs. The programming of infrastructure in order to support certain task workflows is a part of this.

We can now consider how optimisation of data movement and processing links with the other topics of the Data for Science theme based on the information acquired from RIs so far.

Relationship with processing

The distribution of computation is a major concern for the optimisation of computational infrastructure for environmental science. Processing can be initiated at the request of users, or can be part of the standard regime for data preparation and analysis embarked on as part of the 'data pipeline' that runs through most environmental science research infrastructures. Given a dataset, an investigator can retrieve the data within to process on their own compute resources (ranging from a laptop or desktop to a private compute cluster), transfer the data onto a dedicated resource (such as a supercomputer or cluster for which they have leased time and capacity, Cloud infrastructure provisioned for the purpose, or for smaller tasks simply invoke a web service), or direct processing of the data on-site (generally only possible where the investigator has authority to use the site in question, and generally limited to standard analyses that are part of the afore-mentioned data pipeline). Each of these options confers a (possibly zero) cost for data movement, data preparation, and process configuration. Given constraints on compute capacity, network bandwidth, and quality of service, the most pertinent question in the sphere of optimisation is simply, given the sum of all activities engaged in by the research community at large, *where should the data be processed*?

It should be noted that the *outputs* of data processing are as much of concern as the inputs, especially if the curation of experimental results is considered within the scope of a given research infrastructure, and fold back into the domain of data curation.

Relationship with provenance

Good provenance is fundamental to optimisation—in order to be able to anticipate how data will be used by the community, and what infrastructure elements should be able conscripted to provide access to and processing capability over those data, it is necessary to understand as much about the data as possible. Thus provenance data is a key element of knowledge-augmented infrastructure, and provenance-recording services are a major source of the knowledge that needs to be disseminated throughout the infrastructure in order to realise this ideal. Provenance is required to answer *who*, *what*, *where*, *when*, *why* and *how* regarding the origins of data, and the role of an optimised RI is to infer the answers for each of those things as they regard the present and future use of those data. Ensuring that these questions can be asked and answered becomes more challenging the greater the heterogeneity of the data being handled by the RI, and so potential for runtime optimisation in particular will depend on the solutions for optimisation provided by the provenance task (T8.3) in ENVRIplus.

As far as optimisation-serving provenance in and of itself is concerned, the management of provenance data streams during data processing is the most likely area of focus. Preserving the link between data and their provenance metadata is also important, particularly in cases where those metadata are *not* packaged with their corresponding datasets.

Relationship with curation

Streamlining the acquisition of data from data providers is important to many RIs, both to maximise the range and timeliness of datasets then made available to researchers, and to increase data security (by ensuring that it is properly curated with minimal delay, reducing the risk of data corruption or loss) is important.

In general, the principal concerns of curation are ensuring the accessibility and availability of research assets (especially, but not exclusively, data). High availability and long-term durability require effective replication procedures across multiple sites. It would be expedient to minimise

the cost of synchronising replicas and to anticipate where user demand (for retrieval) is likely to be so as to minimise network congestion.

Relationship with cataloguing

Data catalogues are expected to be the main vector by which data is identified and requested by users, regardless of where that data is ultimately accessed from and taken for processing and analysis. As such, the optimisation of both querying and data retrieval is of concern.

Relationship with identification and citation

With regard to identification and citation, it is necessary to ensure availability of identification services, and it is necessary to direct users to the best replicas of a given dataset that would ensure the most effective use of the underlying network.

Optimisation methodology

Optimisation of infrastructure is dependent on insight into the requirements and objectives of the set of research interactions that the infrastructure exists to support. This insight is provided by human experts, but in a variety of different contexts:

- Concerning the immediate context, the investigator engaging in an interaction can directly configure the system based on their own experience and knowledge of the infrastructure.
- Concerning the design context, the creator of a service or process can embed their own understanding in how the infrastructure operates.
- Alternatively, experts can encode their expertise as knowledge stored within the system, which can then be accessed and applied by autonomous systems embedded within the infrastructure.

In the first case, it is certainly possible and appropriate to provide a certain degree of configurability with data processing services, but with the caveat that casual users should not be confronted with too much fine detail. In the second case, engineers and designers should absolutely apply their knowledge of the system to create effective solutions, but should also consider the general applicability of their modifications and the resources needed to realise optimal performance in specific circumstances. It is the third case however that is of most interest in the context of interoperable architectures for environmental infrastructure solutions. The ability to assert domain-specific information explicitly in a generic architecture and thus allow the system to reconfigure itself based on current circumstances is potentially very powerful.

One of the goals of ENVRIplus is to provide an abstraction layer over a number of individual research infrastructures and a number of shared services that they interact with. The purpose of this is to identify and benefit from sharing substantial parts of the e-Infrastructure – see Section 4.2.4 for an explanation of the benefits. To achieve this, every level of the system needs to be well enough described to support automated management and optimisation – see also Section 0 for additional benefits from such descriptions. As developing and delivering these e-Infrastructures has to be collaborative to be sustainable – see Section 4.2.4 – that development of sufficient descriptions of appropriate detail and quality remains a challenge that may take political as well as technical effort. These aspects of optimisation significantly affect the productivity of those building and running e-Infrastructures. They may also reduce operational costs or accelerate the rate at which results and analyses are returned. This last improvement also addresses the highest priority for most RIs, and that is *improving the productivity and success of their researchers*. This of course has to be met by effective automation that reduces their chores and distracting data wrangling. It has to be met by improved usability and easier to understand systems. Making that progress depends on the productivity of the development work. A key step towards this is effective pooling of effort and alliances.

2.3.7 Community Support Analysis

We define Community Support as a subsystem concerned with managing, controlling and tracking users' activities within an RI and with supporting all users to conduct their roles in their communities. It includes many miscellaneous aspects of RI operations, including for example (non-exhaustively) authentication, authorisation and accounting, the use of virtual organisations, training and helpdesk activities.

The questions we asked RI communities focused on 3 aspects: 1) functional requirements, 2) non-functional requirements (e.g., privacy, licensing and performance), and 3) training.

Functional requirements

The following is a summary of the main functional requirements expressed by the RIs (not all apply to all RIs):

- **Data Portal**: a data portal was frequently requested by RIs. Many RIs already have their own data portal, and some of the others are in the process of developing one. Data portals provide (a single point of) access to the system and data products both for humans and machines (via APIs). The following functionalities are commonly requested:
  - Access Control: AAI (Authentication and Authorisation Infrastructure) management is requested by many RIs. For example, IS-ENES2 currently uses OAuth2, OpenID, SAML and X.509 for AAI management.
  - Discovery of services and data facilities: metadata-based discovery mechanisms are commonly used.
- **Accounting**: the tracking of user activities, which is useful for analysing the impact of the RI, is commonly requested. For example, EMBRC records where users are going, what facilities they are using, and the number of requests. The EMBRC head office will in the future provide a system to analyse resource DOIs, metrics for the number of yearly publications and impact factor, and questionnaires submitted by users about their experience with their services. LTER plans to track the provenance of the data, as well as its usage (e.g., download or access to data and data services). DEIMS[56], for example is planning that statistics about users will be implemented, mainly to allow for a better planning of provided services. Features will be implemented by exploiting EUDAT services, e.g., provenance support of B2SHARE to track data usage. Google analytics is currently used to track the usage of the DEIMS interface.
- **Issue tracker**: ACTRIS has recently introduced an issue tracker to link data users and providers, and to follow up on feedback on datasets in response to individual requests.
  - **Community software**: EPOS is in the process of deciding which private software to use and how to integrate it in the data portal. In LTER, the R statistical software and different models (e.g., VSD+ dynamic soil model, LandscapeDNDC regional scale process model for simulating biosphere-atmosphere-hydrosphere exchanges, etc.) are provided.
- **Wiki**: a wiki is often used to organise community information, and as a blackboard for collaborative work for community members (e.g., to add names and responsibilities to a list of tasks to be done). Sometimes, it is also used to keep track of the progress on a task, both for strategic and IT purposes. FAQ pages (and other material targeting a more general audience, or outreach materials for educational institutes) are a special type of wiki page describing more technical aspects of data handling and data products, and also a system for collecting user feedback.
- **Mailing lists, twitter & Forums** are intended to facilitate communication to and from groups of community members. Forums and mailing lists can be interlinked so that any message in the mailing list is redirected to the forum and vice-versa.

[56] https://data.lter-europe.net/deims/

- **Files and image repositories** represent shared spaces where members and stakeholders can upload/download and exchange files. They are also a fundamental tool for storing and categorising images and other outreach material.
- **Shared calendars** keep track and disseminate relevant events for community members.
- **Tools to organise meetings, events and conferences** should handle all the aspects of a conference/meeting: programme, user registration, deadlines, document submission, dissemination of relevant material etc. Tools like Indico are currently popular.
- **Website**: The purpose of the website is to disseminate community relevant information to all stakeholders. The website should not contain reserved material but only publicly accessible material (e.g., documents and presentations for external or internal stakeholders, images for press review). The website should also include news and interactions from social networks. The website should be simple enough to allow almost anyone with basic IT skill to add pages, articles, images. A simple CMS (content management system) is the most reasonable solution (e.g., Wordpress, Joomla).
- **Teleconferencing tools**: Communication with all stakeholders (internal and external) is also carried on through teleconferencing. For this purpose, good quality tools (screen sharing, multi-user, document exchange, private chat, etc.) are needed. Popular tools include Adobe Connect, Web Ex, GoToMeeting, Google Hangout and Skype.
- **Helpdesk & Technical support**: For example, the data products that ICOS produces are complex and often require experience of, and detailed knowledge about, the underlying methods and science to be used in an optimal way. Technical support must be available to solve any problem. The ICOS Thematic Centres (for Atmosphere, Ecosystems and Ocean) are ready to provide information and guidance for data users. If needed, requests for information may also be forwarded to the individual observation stations. The mission of ICOS also comprises a responsibility to support producers of derived products (typically research groups performing advanced modelling of greenhouse gas budgets) by providing custom-formatted "data packages".

As final remark, at the moment, it is difficult to find a pre-existing software package with the aforementioned features. On the other hand, it would be better to re-use tools that community members are familiar with, or are already offered by other (e-) infrastructures. The best approaches could be to provide a toolkit available for RIs, e.g., DataOne[57], or to manually build an internal environment with a single sign-on, which gives access to a bundle of tools (but this second option would need strong efforts in community uptake, appropriation and maintenance).

Non-functional requirements

The non-functional requirements of the RIs that were most frequently referred to were:

- **Performance**: RIs need robust, fast-reacting systems, which offer security and privacy. Moreover, they need good performance for high data volumes.
- **Data policy and licensing constraints**: The data produced by some communities has licensing constraints that restrict access to a certain group of users. For example, while ICOS will not require its users to register in order to use the data portal or to access and download data, it plans to offer an enhanced usage experience to registered users. This will include automatic notifications of updates of already downloaded datasets, access to additional tools, and the possibility to save personalised searches and favourites in a workspace associated with a user's profile. Everyone who wishes to download ICOS data products must also acknowledge the ICOS data policy and data licensing agreement[58] (registered users may do so once, while others must repeat this step every time.)

Training

Training activities within ENVRIplus communities can be categorised as follows:

[57] https://www.dataone.org/investigator-toolkit
[58] ICOS data policy: http://www.socat.info/upload/ICOS_data_policy.pdf

- **No training plan**: The majority of ENVRIplus RI communities do not have a common training plan at the moment.
- **No community-wide training activities**: For example:
    - *Within SIOS*, many organisations have their own training activities. Training is provided to students or scientists. For example, The University Centre in Svalbard (UNIS) has its own high-quality-training programme on Arctic field security (i.e., how to operate safely in an extreme cold climate and in accordance with environmental regulations) for students and scientists.
    - *Within ACTRIS*, each community has its own set of customised training plans. Courses and documentation are made available online, for example for training on how to use the data products. Their preferred methods for delivering training are through the community website or through targeted sessions during community specific workshops. ACTRIS also considers organising webinars.
    - *ICOS* does not have a common training plan at the moment. The Carbon Portal organises occasional training events, e.g., on Alfresco DMS (the Document Management System used by ICOS RI). The different Thematic Centres periodically organise training for their respective staff and in some cases also for data providers (station PIs). ICOS also (co-)organises and/or participates in summer schools and workshops aimed at graduate students and post-docs in the relevant fields of greenhouse gas observational techniques and data evaluation. Representatives of ICOS have participated in training events organised by EUDAT, e.g., on PID usage and data storage technology. The method of delivering training through one- or two-day face-to-face workshops concentrated on a given topic and with a focus on hands-on activities is probably the most effective. This should also be backed up by webinars (including recordings from the workshops) and written materials.
- **A community training plan is under development**: A number of communities are in the process of developing a community training plan. For example:
    - LTER plans the development of a community-training plan. Within LTER Europe[59], the Expert Panel on Information Management is used to exchange information on a personal level and to guide developments such as DEIMS to cater for user needs. LTER Europe also provides dissemination and training activities to selected user groups. Training activities will enhance the quality of the data provided, by applying standardised data quality control procedures for defined data sets.
    - For EPOS, training is part of its communication plan.
- **An advanced system is in place for training activities:**
    - Within IS-ENES2, workshops are organised from time to time. Also, communities communicate about the availability of training courses and workshops organised by HPC centres or European projects (PRACE, EGI, etc.).
    - Within EMBRC, a Training web portal is provided, offering support to training organisers to advertise and organise courses.

3 Review of technologies

3.1 Technology review methods

Task 5.1 is also involved in performing a review of the state-of-the-art technologies provided by data and computational infrastructures. The technology review has two important purposes:

1) *Informing the other tasks in Theme 2*, including the six pillars supporting the data lifecycle, the three cross-cutting topics to make them work together and the provision of

[59] http://www.lter-europe.net/

computational resources on which the envisaged services and systems will run (see Figure 4) for their contribution to ENVRIplus and their relationships. This will ensure that those working on data-infrastructure tasks in ENVRIplus work packages or use cases will have access to up-to-date and relevant information. They would still be well advised to refresh this information with a close focus on the work they are undertaking. Inevitably, the technology review is a broader analysis than they will require and, as technology in this context evolves rapidly, an update is always wise.

2) *Advising the RIs in ENVRIplus* when they decide to implement or upgrade their e-Infrastructures. Again, the information gathered here and in the corresponding wiki[60], has long-standing value as a review of the issues to be considered in each context and a current set of entry points to sources of information. This should be revisited and re-analysed focussing on the specific technological issues an RI or group of RIs are considering. This will also refresh the information as the available solutions may have changed dramatically.

These technology review results are publicly available and publicly updateable to contribute the information to others addressing similar issues and to act as a virtual-whiteboard where those with good solutions can contribute evidence of their value.

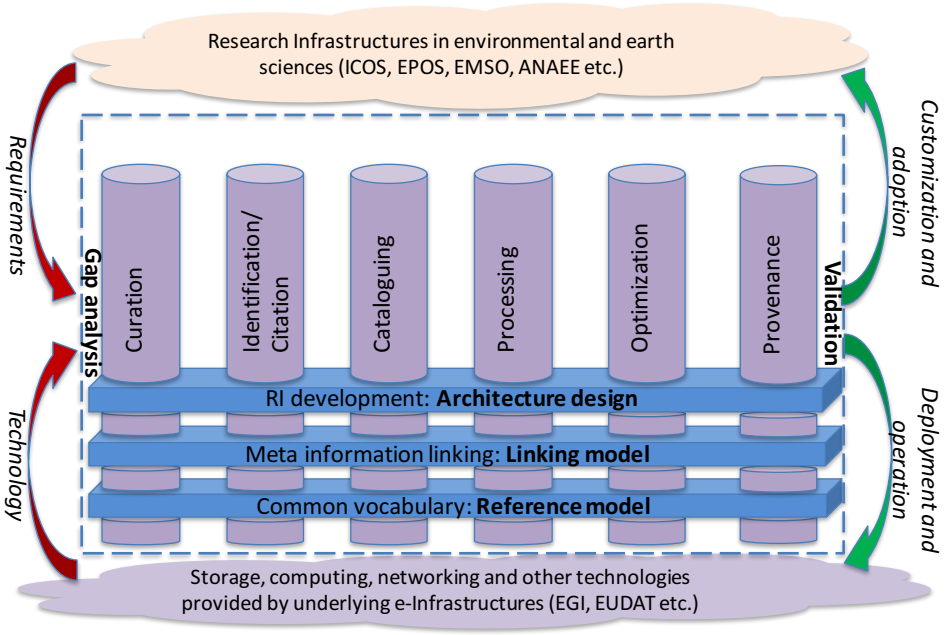

FIGURE 4: SIX PILLARS AND CROSSCUTTING MECHANISMS TO MAKE THEM WORK TOGETHER[61]

A start in the direction of considering and discussing current relevant technological trends was made by ENVRIplus through the organisation of the IT4RIs workshop, in conjunction with the IEEE e-Science Conference 2016, Munich, in September 2015[62]. They contain key inputs and initial insights from an international cohort of experts.

In planning the survey of candidate technologies, it was agreed to partition them to match the six pillars of Theme 2 and three crosscutting topics. For each area the open issues should be

[60] https://wiki.envri.eu/display/EC/ENVRI+Technology+Review
[61] Z. Zhao, The theme of data for science, presentation in the 1st ENVRIPLUS week meeting. It will also appear in the chapter of Computational Challenges in Global Environmental Research Infrastructures in the book of Terrestrial Ecosystem Research Infrastructures: Challenges, New developments and Perspectives.
[62] The papers and programme can be found at http://escience2015.mnm-team.org/?page_id=319

clarified and candidate solutions should be investigated and evaluated. The following is the full list of topics:

1. Data identification and citation (Section 3.2)
2. Data curation including quality control (Section 3.3)
3. Data, Method, Process and Resource cataloguing (Section 3.4)
4. Data Processing including transformations for data integration (Section 3.5)
5. Optimisation of data-handling to reduce human effort or resource use (Section 3.7)
6. Provenance including tools and mechanisms exploiting provenance (Section 0)
7. Architecture including mechanisms to meet non-functional requirements (Section 0)
8. Semantic linking models (Section 0)
9. ENVRI reference model with extensions to ontologies (Section 3.10)
10. Provision of compute, storage and network resources / platforms (Section 3.11).

For each topic a team was formed to identify issues, identify relevant information sources, investigate these as far as time permitted, and to develop, discuss and refine the material to be made available via the wiki and as a snapshot in this report. Each team has a *leader* and at least one *independent member* to ensure breadth. Table 17 outlines the contributors for each topic.

TABLE 17: CONTRIBUTORS TO THE TECHNOLOGY REVIEW PER TOPIC

Topic	Topic Leader	Independent Member(s)
Identification and citation	Margareta Hellström	Alex Vermeulen and Ari Asmi
Curation	Keith Jeffery	Data Curation Centre and RDA metadata group
Cataloguing	Thomas Loubrieu	Gergely Sipos, Alex Hardisty and Malcolm Atkinson
Processing	Leonardo Candela	Rosa Filgueira
Optimisation	Paul Martin	Zhiming Zhao
Provenance	Barbara Magagna	Malcolm Atkinson, Margareta Hellström and Alessandro Spinuso
Architecture	Keith Jeffery	Malcolm Atkinson, Alex Hardisty
Linking model	Paul Martin	Zhiming Zhao
Reference model	Alex Hardisty	Keith Jeffery, Marcus Stocker and Abraham Nieva
Provision of compute, storage and networking	Yin Chen	Damien Lecarpentier

As for requirements gathering, a wiki space was specifically created for the purpose of the technology review[63]. A page of candidate technologies[64] was developed where members of the teams outlined items that need to be considered, or are being considered and reviewed as part of the technology review. The items refer to general areas that need to be covered, specific technologies, specific examples of implementations of those technologies, or specific examples of the application of those implementations. The page proposes a structure for reporting the items. It had the purpose of recording progress and avoiding duplication.

[63] Technology review wiki pages, https://wiki.envri.eu/display/EC/Getting+started+(Technology+Review)
[64] https://wiki.envri.eu/display/EC/Candidate+technologies+for+review

The wiki also includes pages for each of the areas of investigation, where the teams entered their reviews, e.g., for curation[65].

Each of the technology review sections adopted a similar structure that was developed to aid readers.

1. An introduction to the topic.
2. A summary of the primary sources used to guide follow-up investigations.
3. A short-term analysis of the available choices—what should we do today? The analysis tried to consider the next two to five years, i.e., the duration of ENVRIplus's direct actions.
4. A longer-term vision attempted to assess the ways in which the topic will evolve. Inevitably these are not detailed and cannot be used without further thought, but such visions may stimulate useful long-term planning.
5. Existing relationships within the ENVRIplus project and its members that may be informed by and inform this technology review topic.
6. Known open issues and their implications for this technology topic.

The aspect regarding existing relationships (item 5 above) included work packages and tasks within those WP. It also included current use cases. These are intended to focus on a well-defined target issue, in order to develop deeper understanding and if possible devise implementation strategies and exemplar prototypes to validate those solution strategies and communicate the potential value of investing along these lines. They will involve agile methodologies in most cases. They are organised by WP9 and their current status can be found in the relevant wiki space[66].

It was not always possible to cover all of these aspects for each technology topic, however, a very high standard was achieved. The overall achievements of the technology review and some individual aspects are assessed in Section 4.2 page 136 onwards.

3.2 Identification and citation technologies

Margareta Hellström and Alex Vermeulen, ICOS RI and Lund University

3.2.1 Introduction, context and scope

General comment

It is important to keep in mind that there are many different actors involved in *data identification and citation* as there are in *all of the technology review topics* that follow: data producers (RIs, agencies, individuals); data centres (community repositories, university libraries, global or regional data centres); publishers (specialised on data, or with a traditional focus); and data users (diverse ecosystem, from scientists, experts to stakeholders and members of the public). Technologies should reflect needs and requirements for all of these. Here the focus is on RIs that typically involve all of those viewpoints. Time constants for changing old practices and habits can be very long, especially if they are embedded in established cultures or when capital investment is required.

For these reasons, updating, or implementing totally new, technology alone does not improve "*usage performance*"[67], as the behaviour of the "*designated scientific community*" will influence the discoverability and ease of reuse of research data. Scientific traditions and previous investments into soft- or hardware can lead to large time constants for change. Adopting new database technology quickly could, on paper, provide large benefits (to the data providers) like

[65] https://wiki.envri.eu/display/EC/Curation
[66] https://wiki.envri.eu/display/EC/Use+Cases
[67] The working practices actually adopted by the practitioners in all of the roles involved with data or the work that created it or that it is used for.

lower costs and easier administration and curation, but may *de facto* be unacceptably lowering overall productivity for significant parts of the user community over a long period of time while the transition is achieved.

Unequivocal identification of resources and objects underlies all aspects of today's research data management. The ability to assign persistent and unique identifiers (PIDs) to digital objects and resources, and to simultaneously store specific metadata (url, originator, type, date, size, checksum, etc.) in the PID registry database, provides an indispensable tool towards ensuring reproducibility of research [Duerr 2011], [Stehouwer 2014], [Almas 2015]. Not only do PIDs enable us to make precise references in reports and literature, but it also facilitates recording of object provenance including explicit relationships between connected objects (data and metadata; parent and child; predecessor and successor), as well as unambiguous descriptions of all aspects and components of workflows [Moreau 2008], [Tilmes 2010]. A pervasive adoption of persistent identifiers in research is expected to contribute significantly to scientific reproducibility and efficient re-use of research data, by increasing the overall efficiency of the research process and by enhancing the interoperability between RIs, ICT service providers and users [Almas 2015].

Background - Identification

A number of approaches have been applied to solve the questions of how to unambiguously identify digital research data objects [Duerr 2011]. Traditionally, researchers have relied on their own internal identifier systems, such as encoding identification information into filenames and file catalogue structures, but this is neither comprehensible to others, nor sustainable over time and space [Stehouwer 2014]. Instead, data object identifiers should be unique "labels", registered in a central database that contains relevant basic metadata about the object, including a pointer to the location where the object can be found as well as basic information about the object itself. (Exactly which metadata should be registered, and in which formats, is a topic under discussion, see e.g., [Weigel 2015].) Environmental observational data pose a special challenge in that they are not reproducible, which means that also fixity information (checksums or even "content fingerprints") should be tied to the identifier [Socha 2013].

Duerr *et al*. [Duerr 2011] provide a comprehensive summary of the pros and cons of different identifier schemes, and also assess nine persistent identifier technologies and systems. Based on a combination of technical value, user value and archive value, DOIs (Digital Object Identifiers provided by DataCite) scored highest for overall functionality, followed by general handles (as provided by e.g., CNRI and DONA) and ARKs (Archive Resource Keys). DOIs have the advantage of being well-known to the scientific community via their use for scholarly publications, and this has contributed to their successful application to e.g., geoscience data sets over the last decade [Klump 2015]. General Handle PIDs have up to now mostly been used to enable referencing of data objects in the pre-publication steps of the research data life cycle [Schwardmann 2015]. They could however in principle equally well be applied to finalised "publishable" data.

Persistent identifiers systems are also available for other research-related resources than digital data & metadata, articles and reports—it is now possible to register many other objects, including physical samples (IGSN), software, workflow processing methods— and of course also people and organisations (ORCID, ISNI). In the expanding "open data world", PIDs are an essential tool for establishing clear links between all entities involved in or connected with any given research project (Dobbs 2014).

Background - Citation

The FORCE11 Data Citation Principles [Martone 2014] state that in analogy to articles, reports and other written scholarly work, also data should be considered as legitimate, citable products of research. (Although there is currently a discussion as to whether data sets are truly "published" if they haven't undergone a standardised quality control or peer-review, see e.g., [Parsons 2010].) Thus, any claims in scholarly literature that rely on data must include a

corresponding citation, giving credit and legal attribution to the data producers, as well as facilitating the identification of, access to and verification of the used data (subsets).

Data citation methods must be flexible, which implies some variability in standards and practices across different scientific communities [Martone 2014]. However, to support interoperability and facilitate interpretation, the citation should preferably contain a number of metadata elements that make the data set discoverable, including author, title, publisher, publication date, resource type, edition, version, feature name and location. Especially important, the data citation should include a persistent method of identification that is globally unique and contains the resource location as well as (links to) all other pertinent information that makes it human and machine actionable. In some (sensitive) cases, it may also be desirable to add fixity information such as a checksum or even a "content fingerprint" in the actual citation text [Socha 2013].

Finding standards for citing subsets of potentially very large and complex data sets poses a special problem, as outlined by Huber *et al*. [Huber 2013], as e.g., granularity, formats and parameter names can differ widely across disciplines. Another very important issue concerns how to unambiguously refer to the state and contents of a dynamic data set that may be variable with time, e.g., because new data are being added (open-ended time series) or corrections introduced (applying new calibrations or evaluation algorithms) [Rauber 2015], [Rauber 2016]. Both these topics are of special importance for environmental research today.

Finally, a number of surveys have indicated that the perceived lack of proper attribution of data is a major reason for the hesitancy felt by many researchers to share their data openly [Uhlir 2012], [Socha 2013], [Gallagher 2015]. This attitude also extends to allowing their data to be incorporated into larger data collections, as it is often not possible to perform micro-attribution – i.e., to trace back the provenance of an extracted subset (that was actually used in an analysis) to the individual provider – through the currently used data citation practices.

3.2.2 Sources of state of the art technology information used

- Web sites of
  - *CNRI (Corporation for National Research Initiatives)* https://www.cnri.reston.va.us/
  - *CrossRef* http://crossref.org/
  - *DataCite* https://www.datacite.org/
  - *DataONE* https://www.dataone.org/
  - *DDI (Data Documentation Initiative) Alliance* http://www.ddialliance.org/
  - *DONA (Digital Object Numbering Authority)* https://dona.net/
  - *ePIC (European Persistent Identifier Consortium)* http://www.pidconsortium.eu/
  - *EUDAT* http://eudat.eu/
  - *euroCRIS* http://eurocris.org/ - responsible for CERIF development
  - *ICSU-CODATA (ICSU Committee on Data for Science and Technology)* http://www.codata.org/
  - *IGSN (International Geo Sample Number)* http://www.igsn.org/
  - *ISNI (International Standard Name Identifier)* http://www.isni.org/
  - *MDC (Making Data Count project)* - http://mdc.lagotto.io
  - *OKNF (Open Knowledge Foundation)* https://okfn.org/
  - *OpenAIRE* https://www.openaire.eu/
  - *ORCID* http://orcid.org/
  - *PANGAEA* http://www.pangaea.de/ - Data Publisher for Earth & Environmental Science
  - *RDA (Research Data Alliance)* https://rd-alliance.org/ and the web pages of its many active interest and working groups, including:
    - Bibliometrics Working Group (active)
    - Data Citation Working Group (finished)
    - Data Fabric Interest Group (active)
    - Data Publishing Interest Group (active; in collaboration with ICSU World Data System)

    - Data Type Registries Working Group (finished phase 1, starting phase 2)
    - Metadata Interest Group (active)
    - PID Information Types Working Group (finished)
    - Persistent Identifiers Interest Group (active)
    - Research Data Collections Working Group (active)
  - *Taverna (workflow management system)* http://www.taverna.org.uk/
  - *Thomson-Reuters* http://innovation.thomsonreuters.com
  - *THOR (and its precursor ODIN) project* http://project-thor.eu/
  - *W3C (World Wide Web Consortium)* https://www.w3.org
  - *Wf4ever project* http://wf4ever.github.io/ro/
- Webinars organised by RDA, OpenAIRE, THOR.
- Proceedings from some recent conferences (IEEE etc.),
- Articles in scientific literature (see bibliography)
- Discussions with colleagues and experts, from ENVRIplus partners and other organisations

3.2.3 Two-to-five year analysis of state of the art and trends

As evident from the large number of on-going initiatives for applying identifiers to, and subsequently providing linkages between, all components of research – from individual observation values to the people making them – it is a very difficult task to even try to envisage how the data-intensive research landscape will look in a few years from now.

Here, we list some of the issues and ideas that are being worked on now, and which we feel will continue to be of importance in the coming years:

A. A majority of (starting-up) RIs adapt data curation strategies that are fully capable of handling dynamic data (both versioned static files and truly dynamic databases), centred around persistent identifiers for both data & metadata objects and queries.
B. Standards for unambiguous referencing of subsets of data sets (in citations and in workflow contexts) will become widely adopted by scientists and publishers alike, enabling both efficient (human and machine) extraction of "slices" of data as well as detailed (micro)attribution of the producers of the data subset.
C. More complex data objects will become common, including data collections, "research objects" containing both data and related metadata, and other (virtual) aggregates of research information from a multitude of sources. This will require new strategies for content management and identification at both producer and user level.
D. Systems for allocating persistent identifiers will become more user-friendly, e.g., by development of APIs and human-oriented UIs that are common to all major identifier registries. This will have profound positive impacts on the administration and reproducibility of scientific workflows.
E. To enable efficient automation of data discovery and processing, it will become common to store an enhanced set of metadata about the objects directly in the PID registries' data bases, e.g., related to fixity, versioning, basic provenance and citation.
F. The current trend to implement an ever tighter automated information exchange between publishers, data repositories and data producers will continue, and become the norm in many fields including Environmental and Earth Sciences.
G. More effective usage tracking and analysis systems that harvest citation information not only from academic literature but from a wide range of sources will be developed.

Individual ENVRIplus RIs are engaged in a number of the above-mentioned developments through the activities outlined in the Description of Work of several work packages in Themes 1 and 2.

There is also active participation, by individual ENVRIplus RIs, in projects such as EUDAT2020 or as use cases in RDA groups. However, the relatively short lifetimes, and limited number of members, of this type of project often has several negative consequences. Firstly, there may not

be enough diversity within the use cases to encourage the development of broad solutions that cover the needs and requirements of a wider range of communities. Secondly, the knowledge and experience gained through such work often ends up benefiting only a small number of RIs – if there is any long-lasting application at all!

- ENVRIplus could therefore make a difference by setting up a platform for informing practitioners about on-going initiatives (especially those that involve ENVRIplus members, but not as part of ENVRIplus itself), collection of RI use cases for passing on to the technology developers, and finally promoting the dissemination, implementation and uptake of effective examples.

3.2.4 Details underpinning the above analysis

In this section, we present more background for the 7 topics (A-G) listed above. For each topic, some specific examples of relevant technologies are listed, together with a brief narrative discussion and suggestions for further reading – either links to the bibliography or to organisations whose web site addresses are listed under 4.2.2.

A. A majority of (starting-up) RIs adapt data curation strategies that are fully capable of handling dynamic data (both versioned static files and truly dynamic databases), centred around persistent identifiers for both data & metadata objects and queries.

- Main technology needs: versionable databases to support "time machine" retrieval of large datasets (also sensor data) that are dynamic.

Sources: [Rauber 2016] and personal communications with A. Asmi, 2016.

There exist already today several different technical database solutions that support versioning of database records—both SQL and NoSQL-based. Both approaches have advantages and disadvantages, but with optimised and well-planned schemas for storing all transactions and their associated timestamps, it is possible to achieve "time machine"-like extraction of data (and metadata) as they existed at any given time, without significant losses in performance – at least for moderately-sized databases. But challenges remain, e.g., for databases required to store long time series of high-frequency sensor data. For data stored as flat files, it is mainly the metadata that must be stored in a database supporting versioning database, to allow identification of what file(s) represent the "current state" of the data at a given point in time.

- Connections to cataloguing and maintenance of provenance records, supporting automated metadata extraction and production for machine-actionable workflows.

Sources: [Tilmes 2010], [Duerr 2011] (see example in the article supplement!)] + on-going work in RDA Metadata Interest Group, RDA Research Data Provenance Interest Group and EUDAT2020 (Work Package 8).

In order for data-driven research to be reproducible, it is an absolute requirement that not only all analysis steps be described in detail, including the software and algorithms used, but that the input data that were processed are unambiguously defined. Ideally, this is achieved by minting a persistent identifier for the data set as the basis for the citation, and then adding details about the date when the data was extracted, the exact parameters of the subset selection (if used), version number (if applicable) and some kind of fixity information, like a checksum or content fingerprint. Optimally, at least one of 1) the citation itself; 2) the PID record metadata and/or 3) the resource locator associated with the PID, will provide all this information in a machine-actionable format, thus allowing workflow engines to check the validity and applicability of the data of interest.

Currently, a majority of the ENVRIplus RIs – and their intended user communities – haven't yet started to implement the outlined practices in a consistent manner. As a consequence, the reproducibility of research based on data from these RIs could be called into question. What is

needed to change this situation, are good examples and demonstrators that can be easily adopted by the RIs (without much investment in time and software). Such best practices need to be developed in cooperation across the Work Packages of Theme 2.

B. Standards for unambiguous referencing of subsets of data sets (in citations and in workflow contexts) will become widely adopted by scientists and publishers alike, allowing both efficient (human and machine) extraction of "slices" of data as well as detailed (micro)attribution of the producers of the data subset.

- Query-centric citations for data, allowing for both unambiguous and less storage resource-intensive handling of dynamic data sets

Sources: [Duerr 2011], [Huber 2013], [Rauber 2016]

Data sets from research may undergo changes in time, e.g., as a result of improvements in algorithms driving a re-processing of observational data, errors having been discovered necessitating a new analysis, or because the data sets are open-ended and thus being updated as new values become available. Unless great care is taken, this dynamic aspect of data sets can cause problems with reproducibility of studies undertaken based on the state of the data set at a given point in time. The RDA working group on Data Citation has therefore produced a set of recommendations (in 14 steps) for implementing a query-based method that provides persistently identifiable links to (subsets of) dynamic data sets. The WG have presented a few examples of how these recommendations can be implemented in practice, but there is a great need for continued work towards sustainable and practical solutions that can easily be adopted by RIs with different types of data storage systems.

C. More complex data objects will become common, including data collections, "research objects" containing both data and related metadata, and other (virtual) aggregates of research information from a multitude of sources. This will require new strategies for content management and identification at both producer and user level.

- systems for cataloguing and handling more complex collections, both of data sets and metadata (c.f. "research objects").

Sources: OKFN, wf4Ever, the RDA Data Collections WG (just starting) + RDA Data Type Registries WG (concluded with recommendations).

The increasing complexity of research data and metadata objects adds more challenges. Firstly, in contrast to printed scholarly records like articles or books, data objects are often in some sense "dynamic" – updates due to re-analysis or discovered errors, or new data are collected and should be appended. The content can also be very complex, with thousands of individual parameters stored in a single data set. Furthermore, there is a growing trend to create collections of research-related items that have some common theme or characteristic.

In the simplest form, collections can consist of lists of individual data objects that belong together, such as 365 daily observations from a given year. Similarly, it may be desirable to combine data and associated metadata into packages, or to create even more complex "research objects" that may also contain annotations, related articles and reports, etc. Collections can be defined by the original data producers, but may also be collated by the users of the data – and may thus contain information from a large variety of sources and types. This diversity is prompting work on providing tools for organising and managing collections, e.g., using APIs that are able to gather identity information about collection items (through their PIDs), as well as minting new PIDs for the collections themselves.

There is also a need for sustainable registries for data type definitions that can be applied to "tag" content in a way that is useful and accessible both to humans and for machine-actionable workflows. However, the use of data types varies greatly between different user communities, making it a difficult task to coordinate both the registration of definitions as well as a sustainable

operation of the required registries, especially if these are set up and operated by RIs. Here more work is needed in collaboration with a number of RIs each with differing data-set structures and catalogue organisations, in order to provide clear recipes for data typing.

D. Systems for allocating persistent identifiers will become more user-friendly, e.g., by development of APIs and human-oriented UIs that are common to all major identifier registries. This will have profound positive impacts on the administration and reproducibility of scientific workflows.

- Adoption of a common API for PID minting, applicable across registries and methods.

Sources: [Duerr 2011], [Socha 2012], [Klump 2015] + work by the RDA PID Information Types WG (concluded) and the RDA PID Interest Group (starting now).

Although a number of systems for persistent identification of e.g., scientific publications have been available for over a decade, relatively few researchers are consistently applying these systems to their research data. There is, at the same time, a pressing need to encourage data producers to mint PIDs for any (digital) items belonging in the research data lifecycle that should be "referable" – including also raw data and datasets produced during analysis, and not just finalised and "published" data sets. Surveys have indicated that the reasons for the slow adoption rate include a lack of knowledge about the existing opportunities, confusion over their relative differences and merits, and difficulties related to the identifier minting process (especially when it needs to be performed on a large scale, as often the case for data). The latter problem is to a large extent due to the large variety in design and functionality of PID registry user interfaces and APIs, and there are now several initiatives looking into how the registration and maintenance of PID records can be streamlined and simplified. However, the proposed inclusive user and programmatic interfaces will need extensive testing by a wide range of different user communities. There are also institutional issues, concern over intellectual property rights may inhibit the adoption of working practices or the delegation of authority to allocate PIDs.

E. To enable efficient automation of data discovery and processing, it will become common to store an enhanced set of metadata about the objects directly in the PID registries' databases, e.g., related to fixity, versioning, basic provenance and citation.

- Handle registries also need to become federated, and allow users to add community- or project-specific metadata to the handle records (see recommendations of the RDA WG on PID information types), including those required for identity and fixity verification.

Sources: RDA PID Information Types WG (final), new RDA Data Collections WG + presentations from the ePIC & DataCite PID workshop in Paris, 2015[68].

Mainly motivated by a desire to speed up and facilitate the automation of data discovery and processing, there are calls for the centralised handle (and other PID system) registries to also allow data producers and curators to store more types of metadata about the objects directly in the registries' data bases. Examples include information related to data content type(s), fixity, versioning, basic provenance and citation. This would speed up data processing since the requesting agent (e.g., a workflow process) would be able to collect all basic metadata via just one call to the PID registry, instead of needing to first call the registry and then follow the resource locator pointer to e.g., a landing page (which data would need to be harvested and interpreted).

Some PID management organisations, such as DataCite (and the DOI foundation) already support a relatively broad range of metadata fields, but other registries are more restrictive. The

[68] See http://blog.datacite.org/recap

technology for storing the metadata is already in place, but database systems would need to be upgraded to allow for more PID information types. Also, registry servers' capacity to handle the expected large increase in lookup query requests must be upgraded. Optimal performance will require the PID information types themselves to be defined and registered in a persistent way, e.g., using a data type registry.

F. The current trend to implement an ever tighter automated information exchange between publishers, data repositories and data producers will continue, and become the norm in many fields including Environmental and Earth Sciences.

- Expanding the application of persistent unique identifiers for people and institutions in research data object management, including metadata and PID registry records.

Sources: ORCID and DataCite, THOR web site and webinar series.

Driven by demands from large scientific communities (e.g., biochemistry, biomedicine and high energy physics), publishers and funding agencies, there is a strong movement towards labelling "everything" and "everyone" with PIDs to allow unambiguous (and exhaustive) linking between entities. Currently it quite common for individual researchers to register e.g., an ORCID identity, and subsequently use this to link to articles in their academic publications record. This could be equally well applied to (published) research data, for example by entering ORCID IDs in the relevant "author" metadata fields of the DataCite DOI registry record, and allowing this information to be harvested by CrossRef or similar services.

Connected with this is a growing trend to implement tighter information exchange (primarily links to content) between publishers, data repositories and data producers. There are several on-going initiatives looking into how to optimise and automate this, including the THOR project (operated by CERN), which involves amongst others OpenAIRE, ORCID, DataCite and Pangea. It is expected that the outcomes of these efforts will set the norm.

However, to be fully inclusive and consistent (from a data curation and cataloguing point of view), this practice should be extended to all relevant "personnel categories" involved in the research data life cycle, including technicians collecting data, data processing staff, curators, etc. – not just principal investigators and researchers, This would allow both a complete record of activities for individuals (suitable for inclusion in a CV), but conversely can also be seen as an important source of provenance information for linked data sets.

G. More effective usage tracking and analysis systems that harvest citation information not only from academic literature but from a wide range of sources will be developed.

- Discovering and accounting for (micro)attribution of credit to data producers and others involved in the processing & management of data objects – especially in the context of "complex" data objects

Sources: [Uhlir 2012], [Socha 2012], [Huber 2013] + RDA Research Data Collections Interest Group

There is strong encouragement from policy makers and funding agencies for researchers to share their data, preferably under open-access policies, and most scientists are also very interested in using data produced by others for their own work. However, studies show that there is still widespread hesitancy to share data, mainly because of fears that the data producer will not receive proper acknowledgement and credit for the original work.

These apprehensions become stronger when discussing more "complex" data containers – how to give "proper" credit if only parts of an aggregated data set, or a collection of data sets, were actually used in later scientific works? Indeed, many scientists deem it inappropriate or misleading to attribute "collective" credit to everyone who contributed to a collection.

Proposed solutions, now under investigation by various projects focus on two approaches: 1) making the attribution information supplied together with data sets both more detailed and easier to interpret for end users; and 2) providing means for data centres and RIs to extract usage statistics for collection members based on harvested bibliometric information available for the collections. The first of these could be achieved by e.g., labelling every individual datum with a code indicating the producer, or minting PIDs (DOIs) for the smallest relevant subsets of data, e.g., from a given researcher, group or measurement facility. Based on such information, a data end user can provide detailed provenance about data sets used (at least in article text). The second approach may combine tracing downloads and other access events at the data centre or repository level with bibliometry, with the aim to produce usage statistics at regular intervals or on demand (from a data producer). However, handling each file's records individually would quickly become cumbersome, so methods of reliably identifying groups of files should be considered.

- Organisation of (RI-operated) metadata systems that will allow fast and flexible bibliometric data mining and impact analysis.

Sources: [Socha 2012], ePIC and DataCite PID workshop (Paris, 2015)[68], Make Data Count project, CrossRef, OpenAIRE, THOR.

By analysing information about the usage of research data, e.g., through collecting citations and references from a variety of (academic) sources, it is possible to extract interesting knowledge of e.g., what (subsets of) data sets are of interest, who has been accessing the data and how, and in what way they have been used and for what purpose.

Traditionally, this data usage mining is performed based on searching through citation indices or by full-text searches of academic literature (applying the same methods as for articles, e.g., CrossRef, Scopus, Web of Science), sometimes also augmented by counting downloads or searches for data at repositories and data portals. However, up till recently, citations of data sets were not routinely indexed by many publishers and indices, and such services are still not comprehensively available across all science fields. At least partly, this is due to limits in the design of citation record databases and the insufficient capacity of lookup services. Here, updated technologies and increased use of, e.g., semantic web-based databases, should bring large improvements.

However, it is important to cover also non-traditional media and content types. Such "altmetric sources" include Mendeley, CiteULike and ScienceSeeker, as well as Facebook and Twitter. Indeed, while references to research data (rather than research output) in social media may not be very common in Earth Science yet, it may become more prevalent, e.g., where inferences from digital-media activity complement direct observations in poorly instrumented regions. (There are already examples from e.g., astronomy.) Data are in any case already being referred to in many other forms of non-peer-reviewed science-related content, such as Wikipedia articles, Reddit posts, and blogs. Since authors using these "alternative" information outlets are less likely to use PIDs or other standard citation formats, it is a great challenge to bibliometry mining systems to identify and properly attribute such references.

- Discovery and sharing, especially of data contained in "complex data objects", may be enhanced by the use of data type registries that facilitate subset identification (and retrieval)

Sources: RDA Data Type Registries Working Group, EUDAT

Data sharing requires that data can be parsed, understood and reused by both people and applications other than those that created the data. Ideally, the metadata will contain exhaustive information about all relevant aspects, e.g., measurement units, geographical reference systems, variable names, etc. However, even if present, such information may not be readily interpretable – it may be expressed in different languages, or contain non-standard terminology. There is a need for a support system that allows for a precise characterisation of the parameter

descriptions in a way that can be accessed and understood by both human users and machine-actionable workflows.

Registries containing persistently and uniquely identified Data Type definitions offer one solution that is highly configurable and can be adapted to needs of specific scientific disciplines and research infrastructures. In addition to the basic properties listed above, the type registry entries can also contain relationships with other types (e.g., parent and child, or more complex ones), pointers to services useful for processing or interpretation, or links to data convertors. Data providers can choose to register their own data types (possibly using their own namespace), apply definitions provided by others, or apply a mix of these approaches. The PIDs of the applicable data types are then inserted into the data objects' metadata, and can also be exposed via cataloguing services and search interfaces.

The RDA Data Type Registry working group has designed a prototype registry server, which is currently being tested by a number of RIs and organisations. In a second phase, the RDA group will continue the development of the registry concept by formulating a data model and expression for types, designing a functional specification for type registries, and investigating different options for federating type registries at both technical and organisational levels. The adoption of unambiguous and clear annotation of data, as offered by Data Types, should go a long way towards allaying researchers' concerns that their data will be "misused", either in an erroneous fashion, or for inappropriate purposes.

3.2.5 A longer term horizon

As discussed in a recent report from the RDA Data Fabric Interest Group (Balmas 2015), both the increasing amounts of available data and the rapidly evolving ecosystem of computing services, there will have to be an intensifying focus on interconnectedness and interoperability in order to make best use of the funding and resources available to scientists (and society). Tools and technologies including cloud-based processing and storage, and increasing application of machine-actionable workflows including autonomous information searches and data analyses, will all rely on sustainable and reliable systems for identification and citation of data.

Based on this, we have identified a couple of likely trends for the period up to the year 2020:

- A move towards automation of those aspects of the research data lifecycle that will involve basic tasks like assigning identifiers and citing or referring to all kinds of resources – including data and metadata objects, software, workflows, etc.
- Evolution towards more complex "collections" of research resources, like Research Objects, that will necessitate a more flexible approaches towards both strategies for identification and detailed, unambiguous citation or referencing parts of such objects.
- Much more tightly integrated systems for metadata, provenance, identification and citation will evolve (pushed by data producers, publishers and data centres), offering rapid and trusted feedback on data usage and impact.

3.2.6 Relationships with requirements and use cases

Requirements

There are strong connections between the RI requirements gathered for *identification and citation* with those related to other topics, including *cataloguing*, *curation*, *processing* and *provenance*. A majority of RIs are very concerned with how to best encourage and promote the use of their data products in their designated scientific communities and beyond, but at the same time, it is considered a high priority to implement mechanisms and safeguards that can ensure that the data producers (especially principal investigators and institutes in charge of data collecting and processing) receive proper credit and acknowledgments for their efforts. Here, it seems obvious that consistent allocation of persistent identifiers, and the promotion of standards for using these when citing data use in reports and publications, will go a long way to

fulfil these needs. In addition, efforts to standardise the practices and recipes for identifying subsets of complex data collections, and subsequent extraction of micro-attribution information related to these subsets, would ensure a fair distribution of professional credit asked for by researchers and funding agencies alike.

Work Packages

The overarching objective of the ENVRIplus Work Package 6 is to improve the efficiency of data identification and citation by providing recommendations and good practices for convenient, effective and interoperable identifier management and citation services. WP6 will therefore focus on implementing data tracing and citation functionalities in environmental RIs and develop tools for the RIs, if such are not otherwise available.

Use cases

Of the proposed ENVRIplus case studies[66], those of interest from an I&C perspective are mainly IC_01 "Dynamic data citation, identification & citation", IC_06 "Identification/citation in conjunction with provenance" and IC_09 "Use of DOIs for tracing of data re-use". (At the time of writing, these are under review or preparation, with some likelihood of a merger of the three.) The primary aim of IC_01 is to provide demonstrators of the RDA Data Citation Working Group's recommendation [Rauber 2016] for a query-centric approach to how retrieval, and subsequent citation, of dynamic data sets should be supported by the use of database systems that track versions. This may be combined with support also for collections of data sets, which can be seen as a sub-category of dynamic datasets, thus addressing also the goals of IC_09. IC_06 is aimed at identifying good practices for using PIDs for recording provenance throughout the data object lifecycle, including workflows and processing.

3.2.7 Summary of analysis highlighting implications and issues

Tools and services now under development that will allow seamless linking of data, articles, people, etc. are likely to have a large impact on individual researchers, institutions, publishers and stakeholders by allowing streamlining of the entire data management cycle, virtually instantaneous extraction of usage statistics, and facilitation of data mining and other machine-actionable workflows.

While DOIs for articles, and ORCID identifiers for researchers, are now an accepted part of the scientific information flow, publishing of data may not even consider identifiers for other resources (except for publications, for which DOIs are well established). To speed up the adaptation, both current and future technologies for (data) identification and citation must not only be flexible enough to serve a wide range of existing research environments, but they also have to be shown to provide clear benefits to both producers, curators and end users.

Indeed, while some research communities and infrastructures have fully embraced the consistent use of PIDs for data, metadata and other resources throughout the entire data lifecycle, many others are only beginning to think about using them. Important reasons for this hesitancy or tardiness include a substantial knowledge gap, perceived high investment costs (both for personnel, hardware and software), and a lack of support from the respective scientific communities to change engrained work practices.

ENVRIplus is expected to play an important role in defining best practices for first applying identifiers to data and other research resources – including the researchers themselves – and secondly, how use them for citations and provenance tracking. This will be achieved by 1) designing and building demonstrators and implementations based on concrete needs and requirements of ENVRIplus member RIs; and 2) providing documentation and instructional materials that can be used for training activities.

Further discussion of the data identification and citation technologies can be found in Section 4.2.5. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

3.3 Curation technologies

Keith Jeffery, British Geological Survey (BGS)

3.3.1 Introduction, context and scope

"Digital curation is the selection, preservation, maintenance, collection and archiving of digital assets. Digital curation establishes, maintains and adds value to repositories of digital data for present and future use. This is often accomplished by archivists, librarians, scientists, historians, and scholars" (Wikipedia).

It should be noted that Cataloguing, Curation and Provenance are commonly grouped together since the metadata, workflow, processes and legal issues associated with each have more than 70% intersection and therefore rather than generating independent systems a common approach is preferable. Moreover, there are strong interdependencies with identification and citation, with AAAI, with processing, with optimisation, with modelling and with architecture.

3.3.2 Sources of state of the art technology information used

Relevant sources are the Data Curation Centre (DCC), Open Archival Information System (OAIS) (both discussed below) and Research Data Alliance (RDA), which has several relevant groups notably preservation[69] but also active data management plans[70] and reproducibility[71].

3.3.3 Short term analysis of state of the art and trends

The ideal curation state is aimed to ensure the availability of digital assets through media migration to ensure physical readability, redundant copies to ensure availability, appropriate security and privacy measures to ensure reliability and appropriate metadata to allow discovery, contextualisation and use, including information on provenance and rights. The current practice commonly falls far short of this with preservation commonly linked with backup or recovery (usually limited to the physical preservation of the digital asset) and lacking the steps of curation (selection, ingestion, preservation, archiving (including metadata) and maintenance. Furthermore, in the current state while datasets may be curated it is rare for software or operational environments to be curated. Including these necessary to achieve reusability [Belhajjame 2015]. Collecting them automatically has been demonstrated by [Santana-Perez 2016], where processes in a virtual environment are monitored and their interactions with external resources recorded. The collected information is used to automatically create a virtual image in which the job can be deployed and re-run on the cloud.

Curation Lifecycle

The desirable lifecycle is represented by a DCC (Digital Curation Centre) diagram[72] (Figure 5).

Data Management Plan

Increasingly research funders are demanding a DMP (Data Management Plan). Different organisations have proposed different templates and tools for plans but that of DCC is used widely[73] as is the US equivalent[74]. A DMP is defined (Wikipedia) "A data management plan or

[69] https://rd-alliance.org/groups/preservation-e-infrastructure-ig.html
[70] https://rd-alliance.org/groups/active-data-management-plans.html
[71] https://rd-alliance.org/groups/reproducibility-ig.html
[72] http://www.dcc.ac.uk/resources/curation-lifecycle-model
[73] http://dmponline.dcc.ac.uk
[74] http://dmp.cdlib.org

DMP is a formal document that outlines how you will handle your data both during your research, and after the project is completed".

OAIS Reference Model

OAIS (Open Archival Information Systems Reference Model — ISO 14721:2003) provides a generic conceptual framework for building a complete archival repository, and identifies the responsibilities and interactions of Producers, Consumers and Managers of both paper and digital records. The standard defines the processes required for effective long-term preservation and access to information objects, while establishing a common language to describe these. It does not specify an implementation, but provides the framework to make a successful implementation possible, through describing the basic functionality required for a preservation archive. It identifies mandatory responsibilities, and provides standardised methods to describe a repository's functionality by providing detailed models of archival information and archival functions [Higgins 2006]. A set of metadata elements in a structure has been proposed[75].

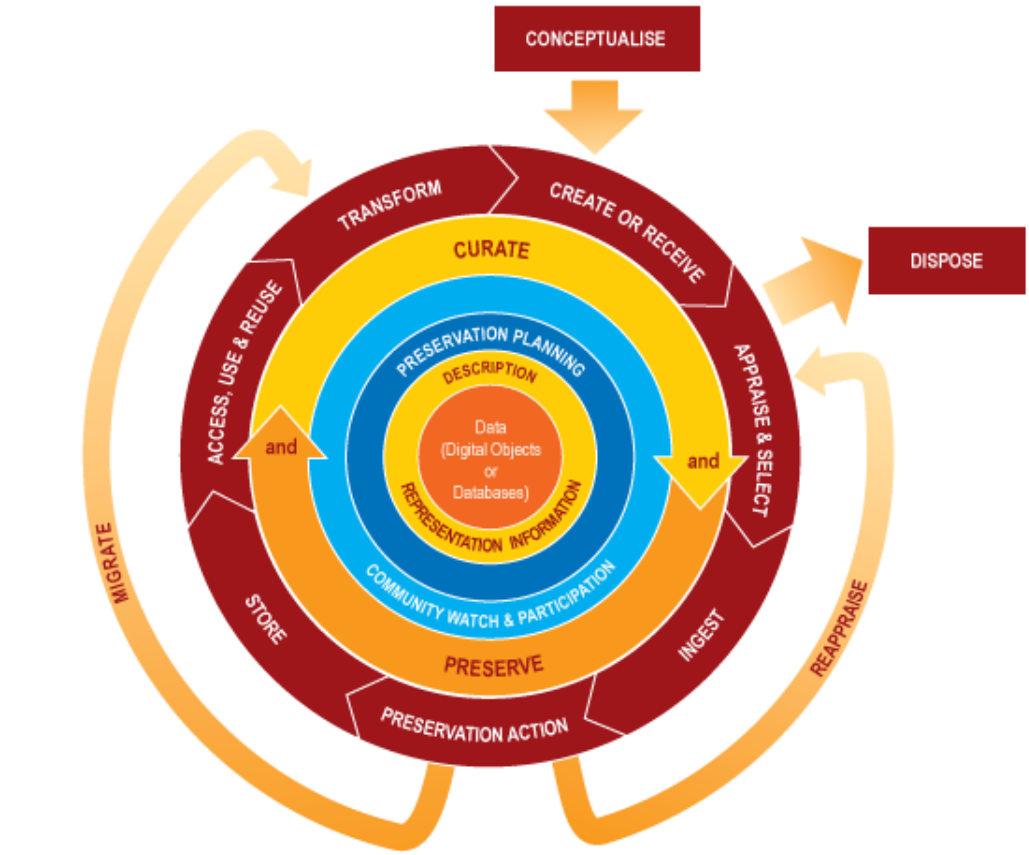

FIGURE 5: THE CURATION LIFECYCLE MODEL

Problems to be Overcome

The following are some important problems that need to be addressed for curation:

1. **Motivation**: There is little motivation for researchers to curate their digital assets. At present curation activity obtains no 'reward' such as career preferment based on data citations. In some organisations curation of digital assets is regarded as a librarian function but without the detailed knowledge of the researcher the associated metadata

[75] http://www.oclc.org/content/dam/research/activities/pmwg/pm_framework.pdf

is likely to be substandard. Increasingly funding agencies are demanding curation of digital assets produced by publicly funded research.

2. **Business model**: Curation involves deciding what assets to curate and of those, for how long they should be kept. Determining an appropriate duration of retention for a digital asset is a problem; economics and business models do not manage well the concept of infinite time. First a business justification is needed in that (a) the asset cannot be collected again (i.e., it is a unique observation, experiment); (b) the cost of collecting again (by the same or another researcher) is greater than the cost of curation.
3. **Metadata**: Metadata collection is expensive unless it is automated or at least partially automated during the data lifecycle by re-using information already collected. Commonly, metadata is generated separately for discovery, contextualisation, curation and provenance when much of the metadata content is shared across these functions. A comprehensive but incrementally completed metadata element set is required that covers the required functions of the lifecycle. It needs sufficient application domain data that other specialists in that domain will be able to find and correctly interpret the associated data.
4. **Process**: The lifecycle of digital research entities is well understood and it needs process support. The incremental metadata collection aspect is critically important for success. Workflow models – if adapted to such an incremental metadata collection with appropriate validation –are likely to be valuable here [Jeffery 2006].
5. **Curation of data**: It may be considered that curation of data is straightforward –but it is not. First the dataset may not be static (by analogy with a type-specimen in a museum); both streamed data and updateable databases are dynamic thus leaving management decisions to be made on frequency of curation and management of versions with obvious links to provenance. Issues related to security and privacy change with time and the various licences for data use each have different complexities. The data may change ownership or stewardship. Derivatives may be generated and require management including relationships with the original dataset and all its attendant metadata.
6. **Curation of software**: Software written 50 years ago is unlikely to compile (let alone compose with software libraries and execute) today. Indeed, many items of software, such as the workflows behind a scientific method, will either not run or give different results, six months later. Since many research propositions are based on the combination of the software (algorithm) and dataset(s) then the preservation and curation of the software becomes very important. It is likely that in future it will be necessary to curate not only the software but also a specification of the software in a canonical representation so that the same software process or algorithm can be reconstructed (and ideally generated) from the specification. This leaves the question of whether associated software libraries are considered part of the software to be curated or part of the operating environment (see below). Very often software contains many years-worth of intellectual investment by collaborating experts. This makes it very valuable and hard to replace. Taking good care of such assets will be a requirement for most research communities.
7. **Curation of operational environments**: It is necessary to record the operational environment of the software and dataset(s). The hardware used – whether instrumentation for collection or computation devices – has characteristics relating to accuracy, precision, operational speed, capacity and many more. The operating system has defined characteristics and includes device drivers – i.e., a software library used by the application. It is a moot point whether software libraries belong to the application software or to the operational environment for the purposes of curation. Finally the management ethos of the operational environment normally represented as policies requires curation.

3.3.4 A longer term horizon

There is some cause for optimism:

1. Media costs are decreasing – so more can be preserved for less;
2. Awareness of the need for curation is increasing; partly through policies of funding organisations and partly through increased responsibility of some researchers;
3. Research projects in ICT are starting to produce autonomic systems that could be used to assist with curation.

However, the major problem is the cost of collecting metadata for curation. Firstly, incremental collection along the workflow with re-use of existing information should assist. Workflow systems should be evolved to accomplish this. Secondly, improving techniques of automated metadata extraction from digital objects may reach production status in this timeframe[76].

3.3.5 Relationships with requirements and use cases

All the requirements obtained from the interviews and the use cases indicated some awareness of the need for digital curation. However, few RIs had advanced towards providing systems to achieve curation and even those that had advanced had not a full data management plan (including business case) in place.

3.3.6 Issues and implications

1. Commonality of metadata elements across curation, provenance, cataloguing (and more) so a common metadata scheme should be used;
2. Metadata collection is expensive so incremental collection along the workflow is required: workflow systems should be evolved to accomplish this and scientific methods and data management working practices should be formalised using such workflows to reduce chores and risks of error as well as to gather the metadata required for curation;
3. Automated metadata extraction from digital objects shows promise but production system readiness is some years away
4. ENVRIplus should adopt the DCC recommendations
5. ENVRIplus should track the relevant RDA groups and – ideally – participate

Further discussion of the curation technologies can be found in Section 4.2.6. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

3.4 Cataloguing technologies

Thomas Loubrieu, L'Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER)

3.4.1 Introduction, context and scope

The technological review for cataloguing covers a subset of the different concepts to be managed in catalogues, as seen in requirement section 2.3.3.

- **Reference catalogues**: persons and organisations, publications, research objects.
- **Federated catalogues**: datasets, resources, physical samples, procedures and software

[76] http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/automated-metadata-extraction

Activity and event logs have not been considered in the technology review because the subject is not mature enough in RI and ICT to manage this information in catalogue yet. As a matter of priority, we focus on references and federated catalogues described above.

The review gives an overview of the software applications or systems and interface standards used for cataloguing related information.

3.4.2 Sources of state of the art technology information used

The standards considered are provided by the following bodies:

- W3C: www.w3.org
- ISO: www.iso.org/iso/home.html
- OGC: www.opengeospatial.org
- RDA working groups (e.g., on metadata)
- Domain-specific standards (e.g., CERIF and geoscienceml).

We identify whether tools are open-source. They may be desktop or server side (with web interfaces) software.

3.4.3 Short term analysis of state of the art and trends

References catalogues

- Persons and organisations: The most popular system for person's identification and cataloguing is currently ORCID. They are involved in THOR project, which helps to connect together datasets, papers and researcher information. They are also working on organisation cataloguing[77].
- Publications management systems cited by RI are Web of Knowledge[78] and Scopus[79].
- For research objects no technology has been cited. Further investigation would be required before developing catalogues for research objects.

Federated catalogues

- **Dataset** catalogues are managed at the RI level with CKAN in the Open Data world and RDA or geonetwork in the ISO and OGC contexts. CKAN is open-source application software developed by the Open Knowledge Foundation[80]. This application is now very popular in the open data projects and is used by EUDAT for the B2FIND function. Geonetwork is open-source catalogue application software for spatially referenced resources and especially datasets. It is very popular in the GIS community and will allow RIs to fulfil INSPIRE requirements for data discovery[81]. Both applications are web servers and can be used and managed on line. It appears to be pragmatic and feasible to harvest existing CKAN and geonetwork in one CKAN central server (e.g., at EUDAT)— Figure 6.

[77] http://orcid.org/blog/2016/03/09/organisations-missing-link
[78] http://apps.webofknowledge.com/
[79] http://www.scopus.com/
[80] For details see http://ckan.org/
[81] For details see http://geonetwork-opensource.org/

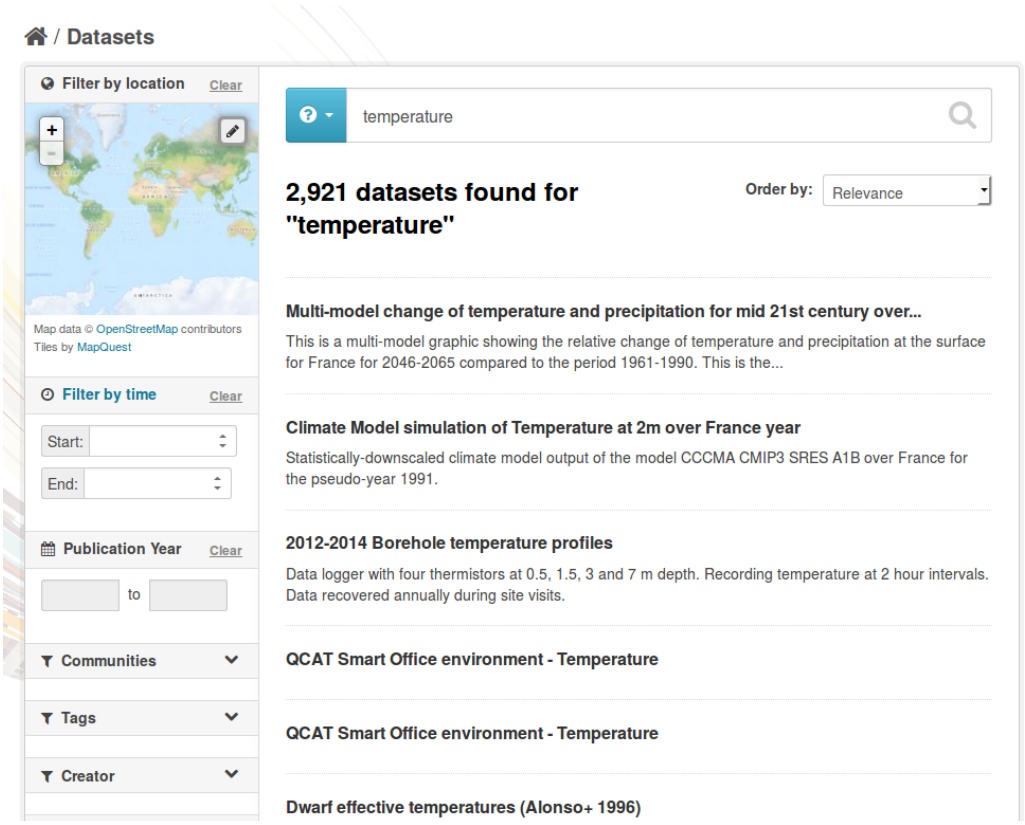

FIGURE 6: CKAN SERVER PROVIDED FOR EUDAT/B2FIND SERVICE

- **Resources**, especially observation equipment are managed in dedicated systems at the RI level, however two standards are popular to describe these items: SSN ontology from W3C and SensorML from OGC. The Semantic Sensor Network ontology together with connected ontologies (e.g., PROV-O) is a W3C standard dedicated to the description of sensors, observations, and related concepts[82]. Although no end-user tools are available yet to implement this ontology the BODC for project SenseOcean[83] is proposing developments in this perspective [Kokkinaki 2016]. OGC is also standardising observation system description with the SensorML standard[84] - see Figure 7. The sensorML standard is part of the Sensor Web Enablement OGC initiative[85]. INSPIRE is recommending this for sensor data sharing and these technologies have been tested and assessed for air or water quality decision support systems with, for example, 52°North software solution [Bröring 2011]. It is currently being developed as well in the marine community to ease the preservation and accessibility of observation context information. The idea developed in Oceans of tomorrows projects (e.g., FixO3 or Nexos) is to enable SWE and especially sensorML from on-board the sensor or instrument to streamline data flow to the data centres (i.e., plug and play sensors). The standardisation of instrument manufacturers specifications is also a goal which is looked at with sensor registries such as the yellow pages developed for EMSO RI[86] which aims at being standardised in sensorML V2[87].

[82] http://www.w3.org/2005/Incubator/ssn/ssnx/ssn
[83] http://www.senseocean.eu/
[84] http://www.opengeospatial.org/standards/sensorml/
[85] http://www.opengeospatial.org/ogc/markets-technologies/swe/
[86] http://www.esonetyellowpages.com/
[87] See export at http://www.ifremer.fr/isi/sensorNanny/emso-yp-sml/

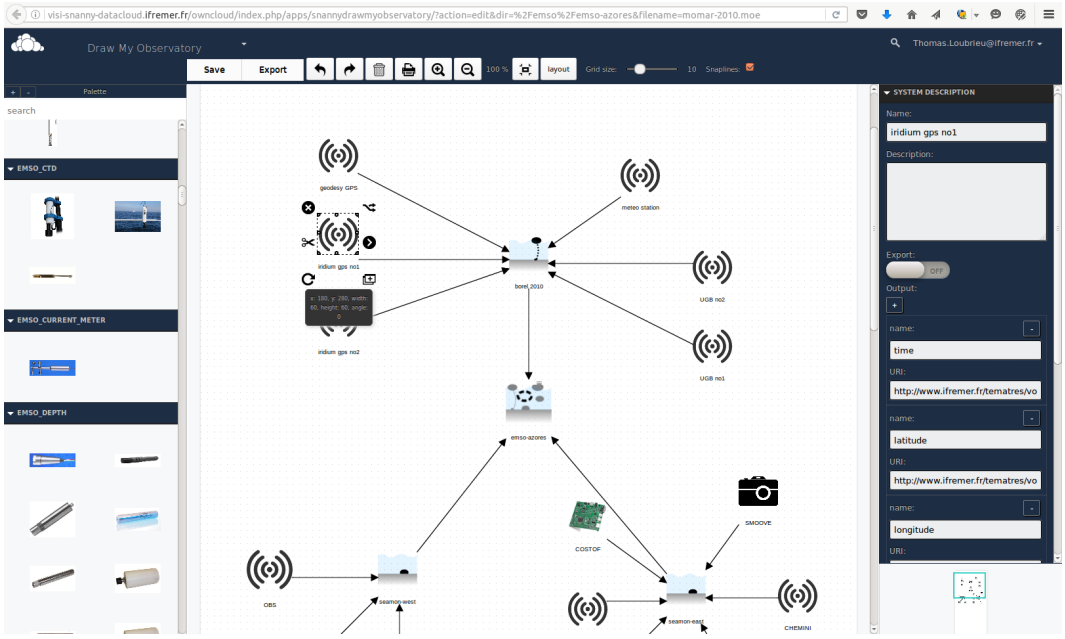

FIGURE 7: SENSORML COMPLIANT EDITOR FOR MARINE OBSERVATION SYSTEM (EMSO RI)

- **Physical samples** are managed in dedicated systems. No common standard has been identified yet. For solid earth specifically, GeoSciML (Figure 8) provides a standard for boreholes and laboratory analysis specimens[88]. In biology specimens records are not standardised and further analysis would be required to review off-the-shelf available software such as Collection management systems[89], or dedicated systems specific to an RI.

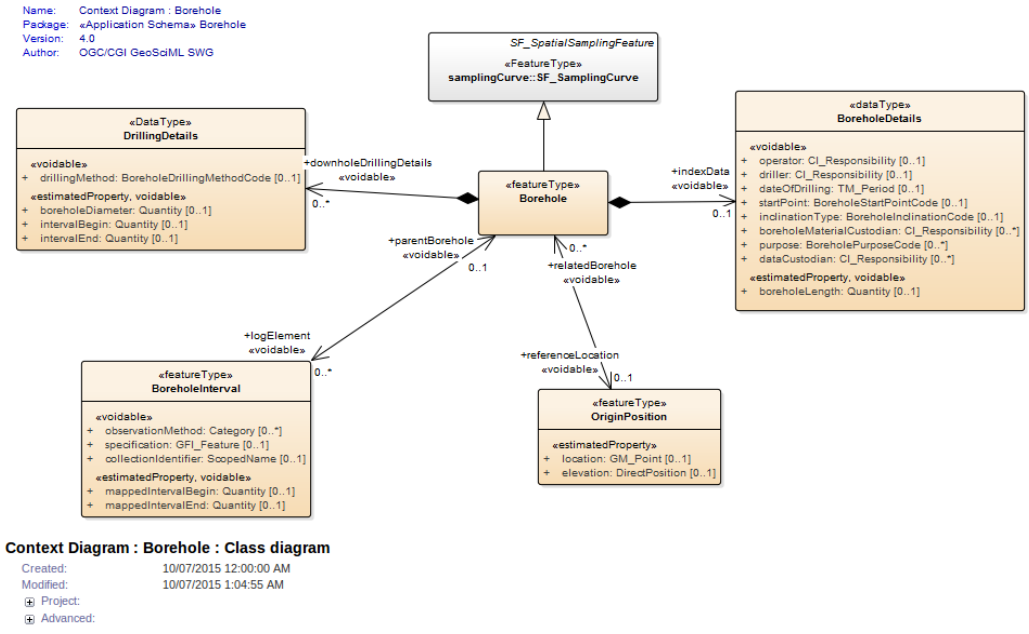

FIGURE 8: BOREHOLE DESCRIPTION IN GEOSCIML

- **Procedures and software**: although we are not aware of standards covering the description of software applications and libraries, there are *de facto* central

[88] http://www.geosciml.org/geosciml/4.0/documentation/html
[89] http://en.wikipedia.org/wiki/Collections_management_%28museum%29

infrastructures very popular on which to host software code, documentation and even project management tools (e.g., a bug tracking system). In the past sourceforge was the most popular. Nowadays, gitHub[90] is more popular. GitHub can be accessed via an API[91], which could be useful to harvest in a catalogue of information related to software and algorithms. No specific tools or standards are identified to document procedures. Generic documents or scientific papers are used to describe the procedures.

Overall solutions: CERIF proposed by EPOS provides an overall conceptual model for managing the above information (see Figure 9).

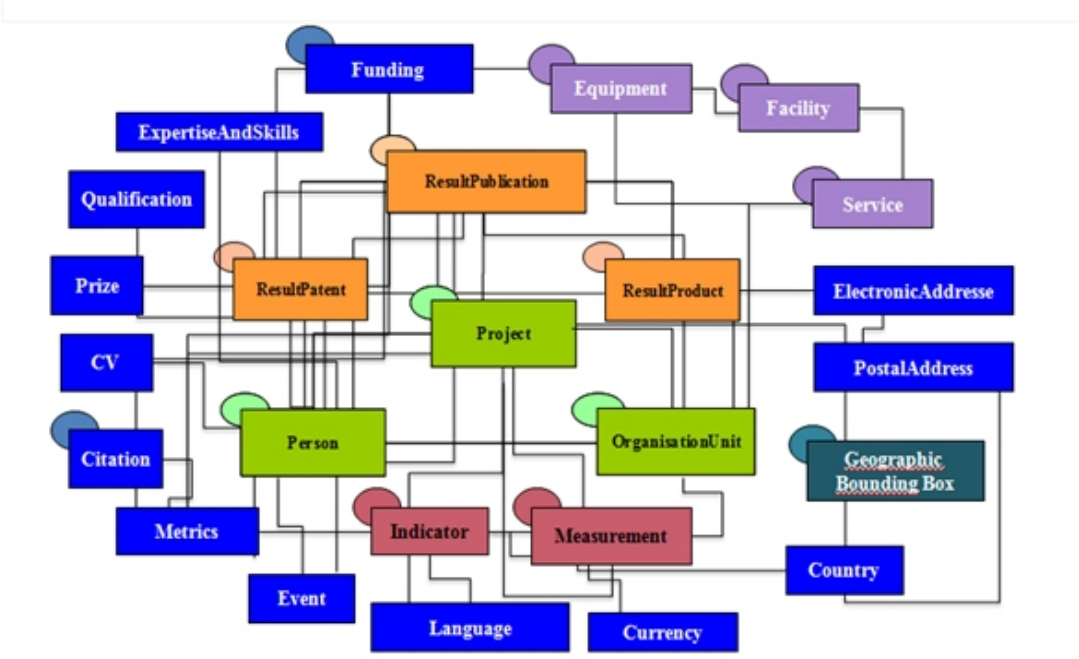

FIGURE 9: CERIF GENERAL DATA MODEL

Catalogue federation will be done by aggregating RI catalogue distributed instances content through common dedicated standards (e.g., CSW/ISO19139 for datasets, sensorML for sensor or instruments, etc.) into state-of-the-art catalogue applications for each type of catalogue (e.g., CKAN for datasets). To prepare the future (see below) and enable cross-catalogue federation, the availability of semantic capabilities (e.g., RDF and SPARQL interfaces) will be considered.

3.4.4 A longer term horizon

In ENVRIplus, catalogue federation will rely on the most popular solutions in each field, datasets, observation systems, samples, software and procedures, and we can expect cross-fertilisation between Research Infrastructures on this subject and rely on catalogue specific official processes (e.g., CSW/ISO19139 for datasets) or *de facto* (CKAN restful API) standards to federate them. *This cross-fertilisation of cataloguing strategies may be a good focus for a think tank.*

Beyond the interoperability or federation of catalogues across RIs per type of object (e.g., datasets) we can expect interoperability between catalogues of different items (e.g., datasets and observation systems). Actually some of the tools identified to implement catalogues already provide generic SPARQL interfaces[92], which can be foreseen as a semantic interfaces between

[90] https://github.com/
[91] https://developer.github.com/v3/
[92] For example, http://www.geocat.net/query-geonetwork-with-sparql

catalogues of different types of object. In this perspective, the availability of such semantic interfaces on top of catalogue implementation will be a selection criterion.

One expectation from the *provenance* activity coupled with *cataloguing* is to provide guided user searches over catalogues by exploiting statistical results mined from previous searches.

It is clear that many *de jure* and *de facto* metadata standards for cataloguing exist and are used. Even stating that an RI uses CKAN doe s not indicate homogeneity since the semantics can be very different for different implementations, and many RIs extend existing metadata standards. Therefore, ENVRIPlus – as it wishes to promote interoperability among RIs – will need to manage metadata heterogeneity – for datasets, software components, workflows, persons, resources, publications etc. One approach (being used in EPOS) is to choose a rich metadata model (in this case CERIF) and provide matching and mapping software and conversion software for the commonly used metadata schemes in the RIs within EPOS. These include DC, DCAT, CKAN, eGMS, ISO19115/INSPIRE and others, but in each metadata case with different dialects. Remembering that a key performance measure is researcher productivity, we need to be careful not to drown researchers in a sea of incomprehensible metadata. One strategy is to include in the preferences available to users, the ability to select a view, e.g., by ticking the facets of metadata that are of interest in a menu generated from that community's total metadata set. The system then generates a query that selects that subset and interposes it for this user, so the user sees a database view that matches their interests.

3.4.5 Relationships with requirements and use cases

As seen earlier the requirement analysis guides us toward:

1. identification and selection of core central catalogues for persons and documentation (e.g., scientific papers). Software might be as well managed in central repository.
2. Federation of distributed catalogues and central harvesting for datasets, observation systems and samples.

The catalogue developments will be applicable in the following use cases:

- **IC_1, Dynamic data citation**: dataset catalogue metadata can be re-used for registering datasets (e.g., at dataCite) while avoiding multiple edits for the author.
- **IC_2, Provenance:** information gathered in provenance management processes should be hosted by catalogues. For example, the author of datasets tracked through the data lifecycle by the provenance process will be stored in a catalogue.
- **IC_8, Cataloguing, curation and provenance:** is the implementation case for catalogues fulfilling curation and provenance requirements.
- **IC_9, Provenance – use of DOI for tracing data re-use:** datasets and scientific paper catalogues will be used as background for this use case.
- **IC_11, semantic linking framework:** interoperability and semantic linking across catalogues (e.g., datasets with observation systems and persons) will be provided by this use case.
- **TC_4: sensor registry:** will actually be the implementation of one catalogue of sensor or instrument model and instances together with maintenance monitoring tools.

3.4.6 Issues and implications

The harmonisation of item descriptions in catalogues across RIs is the primary challenge of the catalogue topic. The implication is that the catalogue development will not cover every RI for every type of catalogue, but demonstrate the interoperability of some RI systems and the actual value added by ENVRIplus to provide cross-disciplinary catalogues.

In addition, the adherence and actual involvement of key partners in RIs so that information required to populate the catalogue are actually available in the central or federated catalogues will also be an issue. The implication is that the development, as foreseen in use case agile-task-force teams, and subsequently RI involvement in catalogue will be driven by the good will and availability of the key partners.

Further discussion of the cataloguing technologies can be found in Section 4.2.7. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

3.5 Processing technologies

Leonardo Candela, Consiglio Nazionale delle Ricerche (CNR) and Rosa Filgueira, University of Edinburgh.

3.5.1 Introduction, context and scope

There are a great many requirements for processing at every stage of the data lifecycle from validating, error correcting and monitoring during data acquisition to transformations for comprehensible presentations of final results. Every step in between has major processing requirements. All forms of data preparation, filtering and transformation to achieve consistent input to subsequent stages in the data lifecycle or the next step in a scientific method. Analysis, pattern matching and statistical reduction to extract relevant signals from complex and noisy data. Large-scale simulations to generate the implications of current models, correlation of those results with well-prepared derivatives from observations and then refinement of the models.

A lot of technologies and approaches have been developed to support these tasks including:

- ***High Performance Computing solutions***, i.e., aggregated computing resources thus to realise an "high performance computer" (including processors, memory, disk and operating system);
- ***Distributed Computing Infrastructures***, i.e., distributed systems characterised by heterogeneous networked computers called to offer data processing facilities. This includes high-throughput computing and cloud computing;
- ***Scientific workflow management systems (SWMS)***, i.e., systems enacting the definition and execution of *scientific workflows* consisting of [Liew 2016]: a list of tasks and operations, the dependencies between the interconnected tasks, control-flow structures and the data resources to be processed;
- ***Data analytics frameworks and platforms***, i.e., platforms and workbenches enabling scientists to execute analytic tasks. Such platforms tend to provide their users with implementations of algorithms and (statistical) methods for the analytics tasks.

These classes of solutions and approaches are not isolated, rather they are expected to rely on each other to provide end users with easy to use, efficient and effective data processing facilities, e.g., SWMS rely on distributed computing infrastructures to actually execute their constituent tasks.

In Europe, PRACE definitely represents the major initiative for High Performance Computing. Similarly, EGI is a point of reference for distributed computing. Both these initiatives are discussed in detail in other parts of this deliverable (see Section 3.11.3.2) and will not be further analysed in this section. In this section we will thus focus on Scientific Workflow Management Systems and Data Analytics frameworks and platforms.

Over the last two decades, many large-scale scientific experiments take advantage of scientific workflows to model data operations such as loading input data, data processing, data analysis, and aggregating output data.

The term workflow refers to the automation of a process, during which data is processed by different logical data processing activities according to a set of rules, along with the attendant tasks of, for example, moving data between workflow processing stages. Workflow management systems (WMS) [Bux 2013] aid in the automation of these processes, freeing the scientist from the details of the process, since WMS manage the execution of the application on a computational infrastructure.

Scientific workflows allow scientists to easily model and express the entire data processing steps and their dependencies, typically as directed Acyclic Graph (DAG), whose nodes represent workflow steps that are linked via dataflow edges, thus prescribing serial or parallel execution of nodes.

Scientific workflows have different levels of abstraction: abstract and concrete. An abstract workflow models data flow as a concatenation of conceptual processing steps. Assigning actual methods to abstract tasks results in a concrete workflow.

There are four key properties of scientific workflows, which are handled differently in each scientific workflow management:

- **Reusability**: Workflow management systems have to make it easier for workflow designer to reuse their previously developed workflows in their under development workflows. Many workflows provide mechanisms for tracing provenance and methodologies that foster reproducible science [Santana-Perez 2015].
- **Performance**: Workflow optimisation is not a trivial task, there are different methods can be applied on a workflow to reduce the execution time [Spinuso 2016].
- **Design**: Almost all the modern workflow management systems provide a rich graphical user interface for creating workflows. The aim of providing graphical composition mechanism is to ease the step of describing workflows for the workflow developers.
- **Collaboration**: Due to the increase in the number of workflows and collaborative nature of scientific research projects developing share and collaboration mechanisms through the network and Internet for workflows is a must. Some projects such myExperiment [De Roure 2009], Wf4Ever [Belhajjame 2015], and Neuroimaging workflow reuse [Garijo 2014], are specially focused on this approach.

Scientific workflows perform two basic functions. They manage (a) the execution of constituent codes and (b) information exchanged between them. Therefore, an instantiation of a workflow must represent both the operations and the data products associated with a particular scientific domain. It should be assumed that individual operations and data products were developed independently in an uncoordinated fashion. Workflows must be usable by the target audience (computational scientists) on target platforms (computing environments) while being represented by abstractions that can be reused across sciences and computing environments and whose performance and correctness can be modelled and verified.

In parallel with scientific workflows, a series of platforms and frameworks have been developed to simplify the execution of (scientific) distributed computations. This need is not new, it is actually rooted in high-throughput computing which is a well-consolidated approach to provide large amounts of computational resources over long periods of time. The advent of Big Data and Google MapReduce in the first half of 2000 brings new interests and solutions. Besides taking care of the smart execution of user-defined and steered processes, platforms and environments start offering ready to use implementations of algorithms and processes that benefits from a distributed computing infrastructure.

3.5.2 Sources of state of the art technology information used

Two major sources of information have been used, literature available discovered by the web and technologies web sites. In particular, the following websites have been source of information:

- Apache Airavata website airavata.apache.org
- Apache Spark website spark.apache.org
- dispel4py website dispel4y.org
- Galaxy website galaxyproject.org
- gCube website www.gcube-system.org
- Kepler website kepler-project.org
- KNIME website www.knime.org
- Pegasus website pegasus.isi.edu
- Taverna website www.taverna.org.uk
- Triana website www.trianacode.org
- Wf4Ever website www.wf4ever-project.org
- WINGS website www.wings-workflows.org

3.5.3 Short term analysis of state of the art and trends

Several technologies and trends characterise the data processing domain.

For **Scientific Workflow Management Systems (SWMS)** [Liu 2015] several have developed a user-friendly way for designing and implementing computational scientific procedures under the workflow paradigm, providing GUIs and tools for easing the task of handling large and complex computational processes in science. Examples of them are:

- **Pegasus** [Deelman 2015] supports execution of workflows in distributed environments such as campus clusters, grids and clouds. Pegasus Workflow Management Service maps an application onto available resources pertaining to the cluster while keeping the internal and external dependencies of the workflow in order. Pegasus workflow has been used to powers LIGO gravitational wave detection analysis.
- **Triana** [Churches 2006] is an open source graphical problem-solving environment that allows you to assemble and run a workflow through a graphical user interface while minimizing the burden of programming.
- **Taverna** [Wolstencroft 2013] provides an easy to use environment to build, execute and share workflows of web services. It has been developed for the enactment of bioinformatics workflows. It emphasizes usability, providing a graphical user interface for workflow modelling and monitoring as well as a comprehensive collection of predefined services
- **Galaxy** [Blankenberg 2011] is a web-based system that aims to bring computational data analysis capabilities to non-expert users in the biological sciences domain. The main goals of the Galaxy framework are accessibility to biological computational capabilities and reproducibility of the analysis result by tracking the information related to every step on the process. The Galaxy workflow model does not follow the DAG paradigm, as it allows to define loops, being a directed cyclic graphs (DCGs) approach.
- **KNIME** [Beiskenr 2013] shares many characteristics with Taverna, with a stronger focus on user interaction and visualisation of results, yet with a smaller emphasis on web service invocation. Furthermore, KNIME focuses on workflows from the fields of data mining, machine learning, and chemistry, while Taverna is more concerned with integration of distributed and possibly heterogeneous data. A graphical user interface facilitates design and execution monitoring of workflows.
- **Kepler** is a frequently used graphical SWMS. Similar to Taverna and KNIME, it provides an assortment of built-in components with a major focus on statistical analysis. Kepler

workflows are written in MoML (an XML format) or KAR files, which are an aggregation of files into a single JAR file. Kepler is built on top of the Ptolemy II Java library, from which it inherits the concepts of Directors and Actors. The former ones control the execution of the workflow, while the actors execute actions when specified by directors.

- **Apache Airavata** [Marru 2011] is an open source, open community SWMS to compose, manage, execute, and monitor distributed applications and workflows on computational resources ranging from local resources to computational grids and clouds Airavata builds on general concepts of service-oriented computing, distributed messaging, and workflow composition and orchestration.

These examples are task-oriented, that is their predominant model has stages that correspond to tasks, and they organise their enactment on a wide range of distributed computing infrastructures (DCI), normally arranging data transfer between stages using files [Vahi 2013]. These systems have achieved substantial progress in handling data-intensive scientific computations; e.g., in astrophysics, in climate physics and meteorology, in biochemistry, in geosciences and geo-engineering and in environmental sciences. In this category we could also include other works like Swift [Wilde 2011], Trident [Simmhan 2009], WS-PGRADE/gUSE [Kozlovszky 2014], SHIWA/ER-flow.

Alternative approaches to task-oriented workflows are the ***stream-based workflows***. This mirrors the shared-nothing composition of operators in database queries and in distributed query processing that has been developed and refined in the database context. Data streaming was latent in the auto-iteration of Taverna, it has been developed as an option for Kepler, and it is the model used by Meandre [Acs 2010], and by Swift (which supports the data-object-based operation using its own data structure). Data streaming pervaded the design of Dispel [Atkinson 2013]. Dispel was proposed as a means of enabling the specification of scientific methods assuming a stream-based conceptual model that allows users to define abstract, machine-agnostic, fine-grained data-intensive workflows. dispel4py [Filgueira 2016] implements many of the original Dispel concepts, but presents them as Python constructs. It describes abstract workflows for data-intensive applications, which are later translated and enacted in distributed platforms (e.g., Apache Storm, MPI clusters, etc.).

Bobolang [Falt 2014], a relative new workflow system based on data streaming, has linguistic forms based on C++ and focuses on automatic parallelisation. It also supports multiple inputs and outputs, meaning that a single node can have as many inputs or outputs, as a user requires. Currently, it does not support automatic mapping to different Distributed Computing Infrastructures (DCIs).

For ***data analytics frameworks and platforms***, a lot of variety exists including:

- ***Apache Mahout*** is a platform offering a set of machine-learning algorithms (including collaborative filtering, classification, clustering) designed to be scalable and robust. Some of these algorithms rely on Apache Hadoop, others are relying on Apache Spark.
- ***Apache Hadoop*** is a basic platform for distributed processing of large datasets across clusters of computers by using a MapReduce strategy. In the reality this is probably the most famous open-source implementation of ***MapReduce***, a simplified data processing approach to execute data computing on a computer cluster [Li 2014]. Worth to highlight that one of the major issues with MapReduce – resulting from the *"flexibility"* key feature, i.e., "users" are called to implement the code of map and reduce functions – is the amount of programming effort. In fact, other frameworks and platforms are building on it to provide users with data analytics facilities (e.g., Apache Mahout).
- ***Apache Spark*** is an open-source, general-purpose cluster-computing engine which is very fast and reliable. It provides high-level APIs in Java, Scala, Python and R, and an optimised engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming.

- ***gCube Data Analytics*** [Candela 2013], [Coro 2014] is an open-source solution conceived to offer an open set of algorithms with the as-a-Service paradigm. The platform relies on a set of DCIs for executing the computing tasks including D4Science and EGI. This platform is equipped with more than 100 ready-to-use algorithm implementations which include real valued features clustering, functions and climate scenarios simulations, niche modelling, model performance evaluation, time series analysis, and analysis of marine species and geo-referenced data. New algorithms can be easily integrated. In fact, the platform comes with a development framework dedicated to this (Java algorithms as well as R scripts are well supported). Once integrated, each algorithm is automatically exposed via a REST-based protocol (OGC WPS) as well as via a web-based GUI that is a complete dashboard for executing computations by guaranteeing Open Science practices (e.g., every computation leads to a "research object" recording and making available every "piece" of the task).
- ***iPython/Jupyter*** [Pérez 2007] is a notebook-oriented interactive computing platform which enacts to create and share "notebooks", i.e., documents combining code, rich text, equations and graphs. Notebooks support a large array of programming languages (including R) and communicate with computational kernels by using a JSON-based computing protocol. Similar solutions include: knitr which works with the R coding language and Dexy is a notebook-like program that focuses on helping users to generate papers and presentations that incorporate prose, code, figures and other media.

The heterogeneity characterising these systems make evident that when discussing data processing "technologies" there are different angles, perspectives and goals to be taken into account. When analysing technologies from the scientist-perspective, the following envisaged trends should be taken into account:

- Technology should be "ease of (re-)use", i.e., it should not distract effort from the pure processing task. Scientists should be exposed to technologies that are flexible enough to enable them to quickly specify their processing algorithm/pipeline. It should not require them to invest effort in learning new programming languages or in deploying, configuring or running complex systems for their analytics tasks. Methods and algorithms are expected to be reused as much as possible, thus data processing should enable them to be "published" and shared.
- "as-a-Service" rather than "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 make it possible to rely on economies of scale and keep the costs low.
- 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. Certain tasks must be executed on specific infrastructures; certain tasks are conceived to crunch data that cannot be moved on other machines from where they are stored.

These trends actually suggest that scientists are looking for "workbenches" / "virtual research environments" / "virtual laboratories" [Candela 2013b] providing them with easy to use tools for accessing and combining datasets processing workflows that behind the scene / transparently exploit a wealth of resources residing on multiple infrastructures and data providers (according to their policies). Such environments should not be pre-cooked / rigid, rather they should be flexible thus to enable scientists to enact their specific workflows. They should provide their users with appropriate and detailed information enacting to monitor the execution of such a workflow and be informed of any detail occurring during the execution. Finally, they should promote "open science" practices, e.g., they should record the entire execution chain leading to a given result, they should enact others to repeat/repurpose an existing process.

3.5.4 A longer term horizon

Data processing is strongly characterised by the "one size does not fit all" philosophy, it does not exist and will never exist a single solution that is powerful and flexible enough to satisfy the needs arising in diverse contexts and scenarios.

The tremendous velocity characterising technology evolution calls for implementing data sustainable processing solutions that are not going to require radical revision by specialists whenever the supporting technologies evolve. Whenever a new platform capable of achieving better performance than existing ones becomes available, users are enticed to move to the new platform. However, such a move does not come without pain and costs.

Data analytics tasks tend to be complex pipelines that might require combining multiple processing platforms and solutions. Exposing users to the interoperability challenges resulting from the need to integrate and combine such heterogeneous systems strongly reduce their productivity.

There is a need to develop data processing technologies that tend to solve the problem by abstracting from (and virtualising) the platform(s) that take care of executing the processing pipeline. Such technologies should go in tandem with optimisation technologies (see Section 3.7) and should provide the data processing designer with fine-grained processing directives and facilities enabling to specify in detail the processing algorithm.

3.5.5 Relationships with requirements and use cases

Most of the RIs that participate in ENVRIplus have computer-based scientific experiments, which need to handle massive amounts of data being some of them generated every day by different sensors/instruments or observatories. In most cases, they have to handle primary data streams as well as data from institutional and global archives. Their live data flows from global and local networks of digital sensors, and streams from many other digital instruments. Often, they employ the two-stage handling of data – established initial collection with quality monitoring, then an open ended exploration of data and simulation models where researchers are responsible for the design of methods and the interpretation of results. These researchers may want to 're-cook' relevant primary data according to their own needs. Their research context has the added complexity of delivering services, such as hazard assessments and event, e.g., earthquake, detection and categorisation, which may trigger support actions for emergency responders. They therefore have the aspiration to move innovative methods into service contexts easily.

Data streaming is essential to enable users such scientists from Atmosphere, Biosphere, Marine and Solid Earth domains, to move developed methods between live and archived data applications, and to address long-term performance goals. The growing volumes of scientific data, the increased focus on data-driven science and the areal storage density doubling annually (Kryder's Law), several stress the available disk I/O – or more generally the bandwidth between RAM and external devices. This is driving increased adoption of data-streaming interconnections between workflow stages, as these avoid a write out to disk followed by reading in, or double that I/O load if files have to be moved. Therefore, data-streaming workflows are gaining more and more attention in the scientific communities.

Another aspect to be considered is that, scientific communities tend to use wide range e-Infrastructures for running their data-intensive applications, e.g., HPC clusters, supercomputers, and cloud resources. Therefore, workflow systems that are able to run them at scale on different DCIs without users making changes to their codes are currently in trend.

It is also necessary to provide facilities to run data-intensive applications across platforms on heterogeneous systems, because data can be streamed to and from several DCIs for performing various analyses. For these DCIs, it is not feasible to store all data since new data constantly

arrive and consumes local store space. Therefore, after data are processed and become obsolete, they need to be removed for newly arrival data. So, data-stream workflow systems should be combined with traditional SWMS systems, which effectively coordinate multiple DCIs and provide functions like data transfers, data clean-up, data location and transfer scheduling.

All in all, the requirements for data processing are very heterogeneous, evolving and varied simply because diverse are the needs when moving across communities and practitioners. Moreover, even within the same community there are diverse actors having different perceptions, ranging from data managers that are requested to perform basic data processing tasks to (data) scientists willing to explore and analyse available data in innovative ways. When analysed from the perspective of (data) scientists the problem tends to become even more challenging because data are heterogeneous and spread across a number of diverse data sources, thus before being analysed for the sake of the scientific investigation, the data need to be acquired and "prepared" for the specific need. Steps will be needed to refine the understanding of these requirements to identify consistent and significant groups where the supplied toolkit for e-Infrastructures may offer common, sharable solutions. Developing that clarity may be another focus for a think tank.

3.5.6 Issues and implications

Scientific workflows have emerged as a flexible representation to declaratively express complex applications with data and control dependences. A wide range of scientific communities are already developing and using scientific workflows to conduct their science campaigns. However, managing science workflows for synergistic distributed and extreme scale use cases is extremely challenging on several fronts workflow management system design, interaction of workflow management with OS/R and provisioning/scheduling systems, data movement and management for workflows, programming and usability, advanced models, provenance capture and validation to name a few.

A major challenge for ENVRIplus RIs applications is the integration of instruments into the scientist's workflow. Many scientists retrieve the data from a (web and/or archive) facility provided by their RIs and then realise some post analyses. Not many RIs offer the possibility to work with life data streamed directly from their instruments/sensors. Therefore, how the ICT workflows community can enable a seamless integration of live experimentation with analysis in a way that increases the overall turnaround time and improves scientific productivity can be identified as one of the mayor challenges, which involve:

- **Provisioning:** Models, algorithms, and mechanisms for resource provisioning: compute, data storage, and network. This includes open questions like How to efficiently determine the resources necessary for workflow execution over time? What information needs to be exchanged between the WMS and resource provisioning systems? How does the WMS adapt to the changes in resource availability?
- **Execution**: Examining the interplay between the WMS and system-side services (data movers, schedulers, etc.), WMS and the operating system or hardware present on the HPC platform. Issues of not only performance but also energy efficiency need to be taken into account. Support streaming data models and manage trade-offs of performance, persistence and resilience of data movements.
- **Adaptation**: Novel approaches to workflow resilience and adaptation. This includes how does the WMS discover hard and soft failures? There are several open questions that need to be addressed: Can provenance information help in detecting some of these anomalies, and the corresponding root causes? How does the WMS adapt to changes in the environment, to failures, to performance degradations? How is the resource provisioning, workflow scheduling, etc. impacted? How do we steer and reschedule workflows when there are failures?

- **Provenance**: What information needs to be collected during execution to support provisioning, execution, and adaptation. What metrics and metadata need to be in the provenance store and what should be the level of detail? Frequency of provenance information collection. Identification and interaction with all the system layers to collect the provenance data. Best strategy to store the provenance data. Development of provenance analysis models to analyse large and complex provenance information.
- **Analytical Modelling**: Exploration of more complex hardware and workflow designs, including novel memory architectures with in situ analysis and co-processing.
- **Collaboration**: another important aspect of the problem is the ability to support workflows within a scientific collaborator and related to that how to support the execution of a set of workflows (a workflow ensemble) on behalf of the user of collaboration, and how to describe and map collaboration workflows.

Besides complex scientific workflows, a lot of scientists are willing to specify their data processing algorithms by realising what falls under the "research software" umbrella. This represents a valuable research asset that is gaining momentum thanks to the open science movement. A lot of such a software is actually implemented by people having limited programming skills and computing resources. In these scenarios, environments conceived to use the software as-is and – with minor directives/annotations – enact its execution by relying on a distributed computing infrastructure are of great help [Coro 2014], e.g., this might enable the scientist to easily execute the code on a number of machines greater than the one he/she usually use, this might enable to expose the algorithm "as-a-Service" and thus to include it in scientific workflows.

Further discussion of the processing technologies can be found in Section 4.2.8. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

3.6 Provenance technologies

Barbara Magagna, Umweltbundesamt GMBH (EAA)

3.6.1 Introduction, context and scope

Provenance, deriving from the French term 'provenir' with the meaning 'to come from', was originally used to keep track of the chain of ownership of cultural artefacts, such as paintings and sculptures as it determines the value of the artwork. But this concept becomes more and more important also in the data-driven scientific research community. Here it is used synonymously with the word lineage meaning origin or source. The knowledge about provenance of data produced by computer systems could help users to interpret and judge the quality of data a lot better. In the W3C PROV[93] documents provenance is defined as information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

3.6.2 Sources of state of the art technology information used

As this topic is intensively studied both from the research viewpoint and from the viewpoint of those deploying and using provenance in production contexts, there are a large number of relevant papers and reports, cited from the text and further identified in the Reference Section from page 193 onwards. The following urls identify other useful sources:

- https://www.w3.org/TR/prov-overview

[93] https://www.w3.org/TR/prov-overview/

- https://rd-alliance.org/groups/research-data-provenance.html
- https://rd-alliance.org/sites/default/files/Krohnposter.pdf
- http://de.slideshare.net/drshorthair/om-alignment-with-ssn-prov-oboe-bfo
- http://twiki.ipaw.info/bin/view/Challenge/WebHome
- https://github.com/NCEAS/open-science-codefest/wiki/ProvenanceR
- https://www.dataone.org/webinars/provenance-and-dataone-facilitating-reproducible-science
- http://eprints.soton.ac.uk/271449/1/opm.pdf
- http://d2i.indiana.edu/provenance
- https://kepler-project.org/users/add_on_modules/provenance
- http://www.taverna.org.uk/documentation/taverna-2-x/provenance/
- http://www.mygrid.org.uk/projects/semantic-provenance-project/
- https://www.eudat.eu/semantics
- http://sead-data.net/

3.6.3 Short term analysis of state of the art and trends

Already by early 2000, provenance of the scientific results was regarded as important as the result itself. [Moreau 2007] considers that, in order to support reproducibility, workflow management systems are required to track and integrate provenance information as an integral product of the workflow. Consequently [Tan 2007] distinguishes between workflow provenance (or coarse-grained), which refers to the record of the entire history of the derivation of the final output of the workflow and data (or fine-grained) provenance, which gives a detailed account of the derivation of a piece of data that is in the result of a transformation step specified in a database query. [Krohn 2014] calls the latter the database provenance with its sub concepts *why*, *where* and *how provenance*. These describe relationships between data in the source and in the output, for example, by explaining *where* output data came from in the input [Bunemann 2001], showing inputs that explain *why* an output record was produced [Bunemann 2001] or describing in detail how an output recording was produced [Cheney 2009]. [Krohn 2014] adds to this characterisation a third type – provenance of web resources with its sub concept access provenance including both actions of publication and consumption of data. [Hartig 2009] provides a base for research on the provenance of linked data from the Web. [Park 2008] describes republishing as the process of transforming sensor data across the Internet. [Lebo 2014] introduces PROV Pingback which enables parties to discover what happened to objects they created after they have left their domain of influence following the Linked Data principles.

Researchers still face the challenging issue that the provenance of the data products they create is often irretrievable. In many cases the tools for composing lineage metadata are not provided with the software used for scientific data processing. [Bose 2005] sees also the problem that no definitive method, standard or mandate exists for preserving lineage of computational results. While this was true in the early 2000 the provenance community reached a significant milestone in 2013 when the World Wide Web Consortium (W3C) published its PROVenance documents. Although combining PROV with Linked Data offers great potential for discovery, access and use of provenance data, the research community needs practical answers about how to do it. Solutions are necessary to bridge the gap between existing systems built on technologies not well suited to adopting Linked Data design and an interconnected Web of provenance with other systems [Lebo 2014]. [Stehouwer 2014] comes to the same conclusion: there seems to be consensus that it would be very good to move away from manually executed or *ad-hoc*-script-driven computations to automated workflows, but there is still a reluctance to take this step. Traditional approaches of provenance management have focused on only partial sections of data lifecycle and they do not incorporate domain semantics, which is essential to support domain-specific querying and analysis by scientists [Sahoo 2011]. Often analysis has to be performed on scientific information obtained from several sources and generated by computations on distributed resources. This unleashes the need for automated data-driven applications that also

can keep track of the provenance of the data and processes with little user interaction and overhead [Altintas 2006]. Comprehensive provenance frameworks as proposed by [Sahoo 2011], [Garijo 2014a], [Myers 2015] or [Filgueira 2015] seem to be the adequate answer to overcome these challenges. These approaches differ from each other and are described below in more detail.

The following section specifies some basic issues related to provenance (see Simmhan 2005): uses, subject, representation, storage, dissemination, tools, collection supported by scientific workflows and by semantic based provenance systems.

Different **uses of provenance** can be envisaged, while currently specific provenance systems typically only support a couple of them [Simmhan 2005]:

Data quality: Lineage can help to estimate data quality and data reliability based on the source data and transformations. It is also used for proof statements on data derivations.

Audit trail: provenance can trace the audit trail of data, determine resource usage and detect errors in data generation. The process that creates an audit trail runs typically in a privileged mode, so it can access and supervise all actions from all users. This makes not only the data lineage transparent but also the use of data after its publication, which could expose sensitive and personal information. It is questionable if usage tracking should be a by-product of provenance which normally should just focus on the origins and transformations of the data product rather than on its users [Bier 2013].

Replication recipes: detailed provenance information can allow repetition of data derivation.

Attribution: pedigree of data can give credit and legal attribution to the data producers, enable its citation and determine liability in case of erroneous data. Summaries of such records are useful when funders review the value of continuing support for data services.

Informational: a generic use of provenance is to query based on lineage metadata for data discovery. By browsing it, a context to interpret data is provided.

The **subject of provenance** information can be of different types as already mentioned above depending on its transparency:

Data-oriented provenance is gathered about the data product and is explicitly available.

Process-oriented (deduced indirectly) provenance focuses on the deriving processes inspecting the input and output data products.

The ***granularity*** at which provenance is detected determines the cost of collecting and storing the related information. The range spans from provenance on attributes and tuples in a database to provenance of collections of files.

Representation of Provenance: different techniques can be used depending on the underlying data processing system.

Annotation: metadata including derivation history of a data product is collected as annotations and descriptions. This information is pre-computed and thus readily usable as metadata.

Inversion: derivations can be inverted automatically to find the source data supplied to them to derive the output data e.g., queries, user-defined functions in databases. This method is more compact.

Provenance related metadata is either directly attached to a data item or its host document or it is available as additional data on the Web [Hartig 2009]. Both types may be represented in RDF using vocabularies or it may be data of another form. The most common *representation languages* used are

- XML
- RDF/OWL using domain ontologies

- CERIF
- dispel4py

Various vocabularies and ontologies exist that allow users to describe provenance information with RDF data.

Provenance models:

During a session on provenance standardization at the International Provenance and Annotation Workshop (IPAW'06) the first Provenance Challenge on a simple example workflow was set up in order to provide a forum for the community to understand the capabilities of different provenance systems and the expressiveness of their representations (Moreau 2007). After the Third Provenance Challenge, the Open Provenance Model (OPM) consolidated itself as the *de facto* standard for representing provenance and was adopted by many workflow systems. The interest of having a standard led to the W3C Provenance Incubator Group, which was followed by the Provenance Working Group. This effort produced the family of PROV specifications[94], which are a set of W3C recommendations on how to model and interchange provenance in the Web.

OPM[95]: In OPM (Open Provenance Model) provenance is represented by graphs. It is used to describe workflow executions. The nodes in this graph represent three different types of provenance information: resources created as *artefacts* (immutable pieces of state), steps used as *processes* (actions or series of actions performed on artefacts) and the entities that control those processes as *agents*. The edges are directed and have predefined semantics depending on the type of their adjacent nodes: *used* (a process used some artefact), *wasControlledBy* (an agent controlled some process), *wasGeneratedBy* (a process generated an artefact), *wasDerivedFrom* (an artefact was derived from another artefact) and *wasTriggeredBy* (a process was triggered by another process). *Roles* are used to assign the type of activity that artefacts, processes and agents played in their interaction and *accounts* are particular views on the provenance of an artefact. OPM is available as two different ontologies which are built on top of each other: the lightweight OPM Vocabulary (OPMV) and the OPM Ontology (OPMO) with the full functionality of the OPM model.

The *PROV* model is very much influenced by OPM. Here resources are modelled as *entities* (which can be mutable or immutable), the steps used as *activities*, and the individuals responsible for those activities as *agents*. Seven types of relationships are modelled: *used* (an activity used some artefact), *wasAssociatedWith* (an agent participated in some activity), *wasGeneratedBy* (an activity generated an entity), *wasDerivedFrom* (an entity was derived from another entity), *wasAttributedTo* (an entity was attributed to an agent), *actedOnBehalfOf* (an agent acted on behalf of another agent) and *wasInformedBy* (an activity used an entity produced by another activity). Roles are kept to describe the type of relationship and the means to qualify each of the relationships using an n-ary pattern are provided. OPM introduces the concepts *plan* associated with a certain activity and PROV statements grouped in *bundles* defined as entities.

[94] https://www.w3.org/TR/prov-overview/
[95] http://eprints.soton.ac.uk/271449/1/opm.pdf

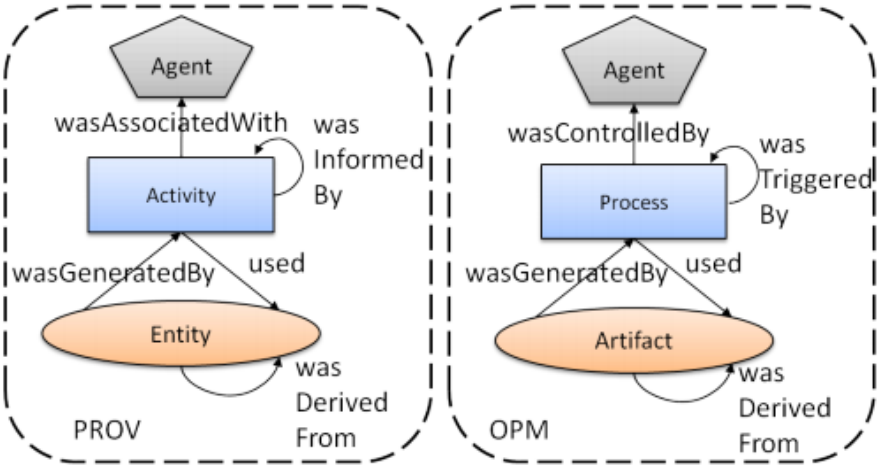

FIGURE 10: THE COMMUNALITIES BETWEEN PROV (LEFT) AND OPM (RIGHT) [GARIJO 2014A].

The PROV family of documents provides among others an ontology (PROV-O), the data model (PROV-DM) and an XML schema (PROV-XML).

Provenir [Sahoo 2011]: is a domain-upper ontology provenance ontology used in translational research. It is consistent with other upper ontologies like SUMO (Suggested Upper Merged Ontology), BFO (Basic Formal Ontology) and DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering). Provenir extends primitive philosophical ontology concepts of continuant and occurent along with ten fundamental relationships. The three top-level classes are data, process and agent, where data is specialised in the classes data_collection and parameter (spatial, temporal and thematic). Provenir is used in the semantic provenance framework (SPF) as explained below.

P-PLAN [Garijo 2014a]: in order to be able to represent workflow templates and workflow instances [Garijo 2014a] extended PROV. The *plan* concept is derived from prov:Plan, the *step* concept represents the planned execution activities and the inputs of a step are modelled as a *variable* with the properties: type, restrictions and metadata.

OPMW [Garijo 2014a]: is designed to represent scientific workflows at a fine granularity. OPMW extends P-plan, PROV and OPM. It is able to model the links between a workflow template, a workflow instance created from it and a workflow execution that resulted from an instance. Additionally, it supports representation of attribution metadata about a workflow. OPMW is used as provenance representation model in the WEST workflow ecosystem.

O&M alignments with PROV[96]: To be compliant with the OGC standard ISO 19156 (Geographic Information – Observation and Measurement) Simon Cox (2015) made efforts to align O&M with PROV. In O&M an observation is an action whose result is an estimate of the value of some property of the feature-of-interest, obtained using a specified procedure.

Provenance storage: in case the data is fine grained, provenance information can become larger than the data it describes. This determines its scalability. This is particularly true when annotations are added manually instead of automatically collecting them.

Provenance dissemination: In order to use provenance, a system should allow rich and diverse means to access it. These can include provenance mining, visualisation and browsing. If provenance is stored in RDF/OWL it is possible to query using SPARQL. Many tools have been developed for PROV for this purpose. A visualisation tool like PROV-O-viz produces derivation graphs that users can browse and inspect [Garijo 2014a].

[96] http://de.slideshare.net/drshorthair/om-alignment-with-ssn-prov-oboe-bfo

Provenance collection: might be performed by stand-alone tools such as ProvenanceR[97], which enables provenance capture in R but these are of more useful when **embedded in a workflow system.**

Provenance collection supported by scientific workflow systems: data analysis can be facilitated by scientific workflow systems that have the ability to make provenance collection a part of the workflow. Here the provenance should include information about the context in which the workflow was used, execution that processed the data and the evolution of the workflow design. Among the most popular of these are Taverna, Kepler and Pegasus. Here only a few a described in some detail – see also descriptions in Section 3.5.

Kepler: is a cross-project collaboration to develop a scientific workflow system for multiple disciplines that provides a workflow environment in which scientists can design and execute workflows. Kepler uses Ptolemy II software, a Java-based system and a set of APIs. The focus is to build models based on the composition of existing components, called 'actors', and observe the behaviour of these simulation models when executed using different computational semantics called 'directors'. Formerly a Provenance Recorder had been implemented to be configured as a 'director' with a standard configuration menu and becoming part of the workflow definition [Altintas 2006]. Today the Kepler Provenance enriches the capabilities of the workflow as add-on module suite. Provenance is toggled on and off in the Kepler toolbar. When on and when running a workflow with a supported director (SDF, DDF, or PN), execution details are recorded into a database in the KeplerData/modules/provenance directory. This powerful feature is leveraged by modules such as Reporting and the Workflow Run Manager, which provides a GUI to manage and share your past workflow runs and results[98].

The *dispel4py* data-streaming system [Filgueira 2015], [Spinuso 2016]: is a versatile data-intensive kit presented as a standard Python library. It describes abstract workflows for stream-based applications, which are later translated and enacted in distributed platforms. It allows users to define abstract, machine-agnostic, fine-grained data-intensive workflows. Scientists can easily express their requirements in abstractions closer to their needs without demanding knowledge of the hardware or middleware context in which they will be executed. A processing element (PE) is a computational activity. It encapsulates an algorithm or a service, and is instantiated as node in a workflow graph. Users only have to use available PEs from the dispel4py libraries and registry, and connect them as they need in graphs which leads to extensive re-usability. The provenance management system of dispel4py consists of a comprehensive system which includes extensible mechanisms for provenance production, a web API and a visualisation tool. The API is capable of exporting the trace of a run in the W3C-PROV JSON representation to facilitate interoperability with third party tools.

Provenance collection supported by semantic-based provenance systems:

Taverna: is an open source and domain-independent Workflow Management System comprising a suite of tools to design and execute scientific workflows. It has been created by the myGrid team and is funded by FP7 projects BioVeL, SCAPE and Wf4Ever. It is written in Java and includes the Taverna Engine (used for enacting workflows) that powers both Taverna Workbench (the client application) and Taverna Server (executing remote workflows). Taverna automates experimental methods through the use of a number of different services from a diverse set of domains. It enables a scientist who has a limited background in computing, limited technical resources and support, to construct highly complex analyses over data and computational resources. Workflow sharing is arranged via myExperiment. Taverna can capture provenance of workflow runs, including individual processor iterations and their inputs and outputs. This provenance is kept in an internal database which is then used to populate the history results in the results perspective in the Taverna Workbench. The provenance trace can be used by the

[97] https://github.com/NCEAS/open-science-codefest/wiki/ProvenanceR
[98] https://kepler-project.org/users/add_on_modules/provenance

Taverna-PROV plugin to export the workflow run, including the output and intermediate values, and the provenance trace as a PROV-O RDF graph which can be queried using SPARQL and processed with other PROV tools, such as the PROV Toolbox. Within Taverna, a workflow can be annotated to give attribution to the Authors of a workflow (or nested workflow)[99]. Although Taverna is not semantic based it supports the semantic description of workflows.

The semantic provenance framework (SPF) (Sahoo 2011): provides a unified framework to effectively manage provenance of translational research data during pre and post-publication phases. It is underpinned by an upper-level provenance ontology (Provenir) that is extended to create domain-specific provenance ontologies to facilitate provenance interoperability, seamless dissemination of provenance, automated querying with SPARQL and analysis. To collect provenance information at a first stage existing data stored in RDB was converted to RDF with help of D2RQ using the domains-specific Parasite Experiment ontology (PEO). On a second stage an ontology-driven web form generation tool called Ontology-based Annotation Tool (OntoANT) was developed to dynamically generate web forms for use in research projects to capture provenance information consistent with PEO in RDF. The SPF stores both the dataset and provenance information together in a single RDF graph. This allows for application-driven distinction between provenance metadata and data, and additionally facilitates that updates of data are seamlessly applied to the associated provenance.

The WEST workflow ecosystem [Garijo 2014a]: integrates different workflow tools with diverse functions (workflow design, validation, execution, visualisation, browsing and mining) created by a variety of research groups. Workflow representation standards and semantic technologies are used to enable each tool to import workflow templates and executions in the format they need. WEST uses and extends the Open Provenance Model and the W3C PROV standard by P-Plan which is able to represent plans. The extension is considered necessary because the OPM and PROV models are not able to represent workflow templates and workflow instances. The OPMW vocabulary is designed to represent scientific workflows at a fine granularity built upon P-Plan, OPM and PROV, and allowing the linking between a workflow template, a workflow instance created from it, and a workflow execution that resulted from an instance. [Garijo 2014a] demonstrate the efficiency of such an approach by the usage of different tools such as WINGS for generating workflows, workflow execution engines such as Pegasus, the FragFlow system for workflow mining, Prov-o-viz for visualising provenance structures, WExp for exploring different workflow templates, the Organic Data Science Wiki, an extension of semantic wikis for workflow documentation and Virtuoso as workflow storage and sharing repository.

Life Science Grid (LSG) (Cao 2009): is a cyber-infrastructure framework supporting interactive data exploration and automated data analysis tools. It uses the Karma provenance framework[100] developed at Indiana University to capture raw provenance events and to format them according to the Open Provenance Model specification. Additionally, it integrates automated semantic enrichment of the collected provenance metadata using the Semantic-Open Grid Service Architecture (S-OGSA) semantic annotation framework developed at University of Manchester.

The Sustainable Environmental Actionable Data (SEAD)[101]: provides data curation and preservation services to deploy those services for beneficial use to active research groups. It intends to support the 'long-tail' of smaller projects in sustainability science. Assuming that metadata could be used to help organise and filter data during research, the SEAD approach allows data and metadata to be added incrementally, and the generation of citable persistent identifiers for data. It comprises three primary interacting components: Project Spaces, Virtual Archive and Researcher Network. The Project Space is a secure, self-managed storage with tools that allow research groups to assemble, semantically annotate and work with data resources. The web application leverages the Tupelo semantic content middleware developed at NCSA,

[99] http://www.taverna.org.uk/documentation/taverna-2-x/provenance/
[100] http://d2i.indiana.edu/provenance_karma
[101] http://sead-data.net/

which provides a blob plus RDF metadata abstraction over an underlying file system and RDF store. The web application itself is an extension to the Java-based Medici semantic content management web application. SEAD has also added a set of restful web services that can be used within the R analysis application to read and write data with desired provenance and metadata. A SPARQL-query service is also implemented. The Virtual Archive is a service that manages publication of data collections from Project Spaces to a range of long-term repositories. It is a federated layer over multiple repositories that manages an overall packaging and publication workflow and provides a global search capability across data published via SEAD. It leverages the Komadu provenance service[102] which is a stand-alone provenance collection tool that can be added to an existing cyberinfrastructure for the purpose of collecting and visualising provenance data. It supports the W3C PROV specification. Komadu is the successor of the Karma provenance tool which is based on OPM.

Another semantic tool which can be adopted for provenance information collection is B2NOTE[103]: The EUDAT project developed a first prototype version using python and common semantic python libraries like RDFlib and SPARQLWrapper. This webservice allows annotation of imported text/documents with terms coming from Bioportal, EnvThes and GEMET from EIONET. This prototype is currently being tested and extended using the Django RESTful framework to be further integrated with the LTER/LifeWatch portal.

3.6.4 A longer term horizon

- In order for data-driven research to be reproducible it is an essential requirement to define unambiguously all data inputs, analysis steps and data products, as well as software and algorithms used with *persistent identifiers*. This will allow for connections to cataloguing and maintenance of provenance records, supporting automated metadata extraction and production for machine-actionable workflows.
- Future provenance management developments will have to implement *interoperability* functions of workflows. The need for global inter-disciplinary collaborations will continue to grow with demands for scientific data to be shared, processed and managed on different *distributed computational infrastructures*.
- Provenance management should embrace the *whole life cycle* of data and incorporate *domain semantics* by encouraging and building on controlled vocabularies formalised as ontologies – see Section 3.9, which is essential to support domain-specific querying and analysis by scientists. The approach used for provenance representation has a significant impact on the storage, dissemination, and querying phases of the provenance life cycle [Sahoo 2011].
- Provenance analytics and visualisation techniques will receive more attention in future applied research [Spinuso 2016]; so far it has been largely unexplored. By analysing and creating insightful visualisations of provenance data, scientists can debug their tasks and obtain a better understanding of their results. [Davidson 2008], [Cao 2009].

3.6.5 Relationships with requirements and use cases

Requirements:

There is a big interest among the RIs to get clear recommendations from ENVRIplus about the information range provenance should provide. This includes drawing an explicit line between metadata describing the 'dataset' and provenance information. Also it should be defined clearly whether usage tracking should be part of provenance.

[102] http://d2i.indiana.edu/provenance_komadu
[103] https://www.eudat.eu/semantics

It is very important to provide support for automated tracking solutions and provenance management APIs to be applied in the specific e-science environments. Although there are some thesauri already in use there is a demand for getting a good overview of the existing vocabularies and ontologies that are ready to use or that need to be slightly adapted for specific purposes.

Work Packages:

There is a strong relationship between WP 6 and the WP 8 task 3 *Provenance* as there must be a direct link between the data and its lineage that can be followed by the interested user. The recommendations provided for data identification and citation should be used in provenance service solutions. Provenance tracking is also an important feature for the tasks 7.1 processing and 7.2 optimisation. The connections with the tasks 8.1 curation and 8.2 cataloguing are evident as well as all of these recommendations must be built upon the same data model, semantically and technically speaking, as defined in the task 5.3 semantic linking framework and integrated in the task 5.4 interoperation based architecture design.

Relationships with use cases as foreseen in WP9:

- **IC_1, Dynamic data citation**: Connections to cataloguing and maintenance of provenance records, supporting automated metadata extraction and production for machine-actionable workflows
- **IC_2, Provenance:** aims amongst others at defining a minimum information set that has to be tracked, finding a conceptual model for provenance which conforms to the needed information, maps existing models to the common model, and finds a repository to store the provenance information.
- **IC_06, Identification/citation in conjunction with provenance**: is aimed at identifying good practices for using PIDs for recording provenance throughout the data object lifecycle, including workflows and processing.
- **IC_8, Cataloguing, curation and provenance:** is the implementation case for catalogues fulfilling curation and provenance requirements.
- **IC_9, Provenance – use of DOI for tracing data re-use:** provenance capture techniques will be used as background for this use case.
- **IC_11, semantic linking framework:** interoperability and semantic linking across catalogues (e.g., datasets with observation systems and persons) upon a common data and metadata model will be provided by this use case.

3.6.6 Issues and implications

- Commonality of metadata elements across curation, provenance, cataloguing (and more) thus a common metadata and provenance scheme based on widely adopted international standards should be used.
- Link to existing vocabularies and ontologies to enable domain semantic provenance representation thus a strong collaboration with the semantic working group.
- Having better visualisation tools at hand for provenance dependencies will increasingly help to reduce the RIs reluctance to adopt workflow solutions with provenance functionalities – thus it is important to follow related developments and to try to implement the most relevant one(s) in the provenance service.
- ENVRIplus should consider collaborating with EUDAT on the development of provenance tools as foreseen in WP 8 and influence the General Execution Framework (GEF) so that it supports the provenance-collection functionality.
- ENVRIplus should follow the RDA provenance working groups and participate.
- Provenance in ENVRIplus is a task which is due in a later stage of the project. Thus it is a must to follow in the meantime tools and services now under development that will

allow seamless linking of data, articles, people supporting streamlining of the entire data management cycle, virtually instantaneous extraction of metadata and provenance information, and facilitating data mining and other machine-actionable workflows.

Further discussion of the provenance technologies can be found in Section 4.2.9. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

3.7 Optimisation technologies

Paul Martin, Universiteit van Amsterdam (UvA)

The optimisation work is scheduled for later in ENVRIplus. Hence, this section is preliminary, as the challenges on which optimisation must focus are not yet decided. However, virtually every RI agrees that the priorities should be chosen to improve productivity of researchers.

3.7.1 Introduction, context and scope

System-level environmental science involves large quantities of data, often diverse and dispersed insofar as there are many different kinds of environmental data commonly held in small datasets. In addition, the velocity of data gathered from detectors and other instruments can be very high. Data-driven experiments require not only access to distributed data sources, but also parallelisation of computing tasks for the processing of data. The performance of these applications determines the productivity of scientific research and some degree of optimisation of system-level performance is urgently needed by the RI projects in ENVRIplus as they enter production.

This topic focuses on how to improve many of the common services needed to perform data analysis and experiments on research infrastructure, with an emphasis on how data is delivered and processed by the underlying e-infrastructure. There needs to be consideration of the service levels offered by e-infrastructures, and of the available mechanisms for controlling the system-level quality of service (QoS) offered to researchers. This topic should therefore focus on the mechanisms available for making decisions on resources, services, data sources and potential execution platforms, and on scheduling the execution of tasks. The semantic linking framework developed in Task 5.3 on linking data, infrastructure, and the underlying network can be used to embed the necessary intelligence to guide these decision procedures (semi-)autonomously.

Ultimately, based on the relevant task (7.2) of the ENVRIplus project, we will need to:

1. Provide an effective mapping between research-level quality attributes (ease-of-use, responsiveness, workflow support) to infrastructure-level quality attributes on computing, storage and network services provided by underlying e-infrastructures.
2. Define test-bed requirements for software and services, and identify conditions for operating final software and services inside each domain, and between multiple domains.
3. Extend and customise existing optimisation mechanisms for computing and storage resources, and provide an effective control model between processes of data analysis and the underlying e-infrastructure resources, making the application performance as easy as possible to control at runtime.

Thus the focus of the technology review in ENVRIplus from the optimisation perspective is to determine two things:

1. What the RI projects already have at their disposal for effective data access, delivery and processing.

2. What mechanisms can be used to meet RI projects' processing and optimisation requirements[104].

The optimisation section of the ENVRIplus technology review focuses on the second point above; the first point should be addressed in other parts of section 3, particularly section 3.5.

3.7.2 Short term analysis of state of the art and trends

In principle, optimisation can be conducted at every level of interaction—at the social level between investigators, at the human-computer interface level between researchers and their tools, at the service level, at the functional level, at the infrastructure level, and so forth. Any number of optimisations can be applied at each of these levels based on an understanding of the technologies and engineering currently being used at that particular level—a thousand different bespoke manipulations in order to ensure perfect operation.

In reality, while there will always be scope for hand-crafted solutions to every problem where the payoff is sufficient to offset the effort required to understand, produce and maintain those solutions, what is increasingly necessary is the ability to produce *generically* optimisable systems. As described in the optimisation requirements analysis (Section 2.3.6), there exist different ways for human experts to embed their insight into the operation of a system:

- The investigator engaging in an interaction can directly configure the system based on their own experience and knowledge of the infrastructure—this is the bespoke optimisation already alluded to.
- The creator of a service or process can embed their own understanding in how the infrastructure operates—this is key to producing high quality software, middleware or e-infrastructure, but it is not always applicable in broader contexts.
- Experts encode their expertise as knowledge stored within the system, which can then be accessed and applied by autonomous systems embedded within the infrastructure—this is the approach that is being adopted in ENVRIplus, in its formal modelling, semantic linking, interoperable architecture design, and provenance support.

To embed knowledge into the system, it is necessary to do so at multiple levels, and it is necessary to link those different levels—from the abstract requirements of researchers to the fundamental characteristics of the infrastructure. This has been the focus of the technology review for optimisation in this instance.

Optimisation is conducted according to certain metrics measured at various levels from different perspectives. From the high-level user perspective, these metrics concern *quality of service (QoS)*.

Most experimental or analytical tasks, especially when distributed, are subject to degraded performance when limited by the underlying infrastructure, especially when that infrastructure is shared with other applications. Thus most QoS research is focused on telephony and the Internet. The International Telecommunication Union defined a standard for telephony QoS in 1994, that was revised in 2008 [ISO 2008]; the ITU later defined a standard for information technology QoS in 1997 [ISO 1997]. Regardless of context, QoS requirements are generally the same; the application requires certain levels of performance in terms of speed, stability, smoothness, response, etc. Advances in distributed computing drive research into service-based infrastructures that provide assets on-demand, reacting to changes in the system in real-time [Menychtas 2009]. Thus the notion of QoS, wherein an application requires a certain level of performance (speed, stability, smoothness, etc.) from components, has been subjected to greater scrutiny of late as the demand to move more and more quality-critical applications onto

[104] https://wiki.envri.eu/display/EC/Processing+requirements, https://wiki.envri.eu/display/EC/Optimisation+requirements

the Internet raises reliability issues that may not be resolvable by blanket over-provisioning of computational and network resources. Li et al. [Li 2012] proposes a taxonomy for cloud performance which can be generalised to Grid and other virtual infrastructure contexts, constructed across dimensions of performance features and experiments. Aceto et al. [Aceto 2013] stress the importance of monitoring of virtualised environments.

If a system provides the ability to prioritise different applications, processes, users, or data-flows as opposed to simply making a best-effort attempt to do everything, then technical factors that influence the ability to fulfil QoS requirements include the reliability, scalability, effectiveness, sustainability, etc. of the underlying infrastructure and technology stack. Other factors however include the information models used to describe applications and infrastructure that then can be used to infer how to manage QoS requirements; for example, [Kyriazis 2008] demonstrates how QoS might be specified and verified when mapping workflows onto Grid environments.

On the platform level, the QoS of the application and QoE of users are ensured by dynamically allocating resources with the fluctuations of workload. There are only limited resources and the computing and networking infrastructures also have a maximum capacity. Therefore, all the resources have to be shared in a virtualised manner. So the challenge is to determine the resource requirements of each application and allocate resources most efficiently. The state of the art of this problem can be classified into resource *provisioning*, resource *allocation*, resource *adaptation* and resource *mapping* [Manvi 2014].

Workflows provide a means for researchers and engineers to configure multi-stage computational tasks, whether as part of the generic operation of a research infrastructure or as part of a specific experiment. Workflows are typically expressed as directed (a)cyclic graphs. A key property is that workflows provide a means to manage dataflow. There are a number of different workflow management systems that could be enlisted by research infrastructure for framing workflows [Deelman 2009]—e.g., Taverna, Pegasus and Kepler. The specification of workflows for complex experiments provides structural information to the operating environment about how different processes interrelate, and thus provides guidance as to how data and processes need to be staged in order to better support research activities. Given information about all the different workflows concurrent in a system, it is also then possible to regulate the scheduling of resources to best optimise *overall* system performance.

Conscripting elastic virtualised infrastructure services permits more ambitious data analysis and processing workflows, especially with regard to 'campaigns' where resources are enlisted only for a specific time period. Resources can be acquired, components installed, and processes executed with relatively little configuration time provided that the necessary tools and specifications are in place. These resources can then be released upon the completion of the immediate task. However, in the research context, it is necessary to minimise the oversight and 'hands-on' requirement for researchers, and to automate as much as possible. This requires specialised software and intelligent support systems; such software either does not currently exist, or operates still at too low a level to significantly reduce the technical burden imposed on researchers, who would rather concentrate on research than programming.

Finally, the adoption and collection of precise provenance information permits deep analysis of historical data and resource use, which can be used to refine decision procedures and so enhance the overall performance of the system.

3.7.3 A longer term horizon

In the longer term, the increasing complexity and use of virtualised infrastructure will widen the gulf between researchers and the hands-on engineering necessary to manually configure the acquisition, curation, processing and publication of datasets, models and methods. Thus context-aware services will be required at all levels of computational infrastructure to manage and control the staging of data and the provisioning of resources for researchers autonomously, and these services will have to be aware of the state of the entire systems, catering not to the whims

of individual researchers, but taking into account the wider use of the system by entire communities. The establishment of such topics will be wholly dependent on integrative thinking—taking heed not just of developments in individual areas of (for example) workflow management, provenance and cataloguing, but also the development of techniques to promote interoperation between all parts of research infrastructure.

3.7.4 Relationships with requirements and use cases

The optimisation topic is strongly related to the compute, storage and networking topic, the processing topic and the provenance topic in particular:

- The focus of optimisation is on more efficient use of underlying e-infrastructure, especially of the kind provided by initiatives such as EGI.
- The target of optimisation is on better data retrieval and processing.
- Autonomous optimisation relies on knowledge embedded in the datasets, services and resources involved in data retrieval and processing tasks—a significant portion of which is generated as part of provenance services.

There are a number of ENVRIplus use-cases for which the optimisation task is a potential contributor[55]:

- The **data subscription service**, for the transport and staging of data onto cloud resources.
- Implementing a prototype **cross-RI provenance model** using workflow management systems and EUDAT services requires intelligent data movement and resource management.
- **Re-processing of data by users using their own algorithms** requires smart resource control.

3.7.5 Issues and implications

It is possible to automate large portions of research activity—however this is contingent on the existence of good formal descriptions of data and processes, and on there being good tool support for initiating and informing the automated procedures with regard specific experiments and applications.

The optimisation of resources is dependent on the requirements of researchers. The quality of service offered is based on certain taxonomies used to frame constraints that are then translated into requirements for the configuration of networks and infrastructure. Three branches can be distinguished in a classical performance taxonomy [Barbacci et al. 1995]:

- *Concerns* list quality of service attributes that may be of concern to researchers.
- *Factors* lists properties of the environment that may impact concerns.
- *Methods* lists the mechanisms at the disposal of the system that can be used to monitor concerns.

It is necessary to identify the concerns of researchers in specific use-cases investigated within ENVRIplus, and to analyse the factors dictating performance in current research infrastructures. The role of Task 7.2 in ENVRIplus is to provide methods for monitoring and responding to selected concerns.

The broader implications of generic optimisation of infrastructure and resources extends to the increasing prevalence of and reliance upon virtualised infrastructure and networks. Being able to generate a deeper understanding of how different kinds of task impose different requirements

on different underlying infrastructure by being able to reason from the level of user-level quality constraints down to physical resource specifications is invaluable if we wish to be able to handle ever more extensive computational research. This is particularly true if we want to keep the accessibility of research assets as open to the broader research community as possible, rather than within the hands of a few well-resourced experts—in this light, we need to consider infrastructure as a utility, one that is intelligent and self-organising.

Further discussion of the optimisation technologies can be found in Section 4.2.10. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

3.8 Architectural technologies

Keith Jeffery, British Geological Survey (BGS), Malcolm Atkinson, University of Edinburgh and Alex Hardisty, Cardiff University.

3.8.1 Introduction, context and scope

As defined in Wikipedia[105], **Information technology architecture** is:

> *"… the process of development of methodical information technology[106] specifications, models and guidelines, using a variety of Information Technology notations, for example UML[107], within a coherent Information Technology architecture framework[108], following formal and informal Information Technology solution, enterprise, and infrastructure architecture processes. These processes have been developed in the past few decades in response to the requirement for a coherent, consistent approach to delivery of information technology capabilities. They have been developed by information technology product vendors and independent consultancies, based on real experiences in the information technology marketplace and collaboration amongst industry stakeholders, for example the Open Group[109]. Best practice Information Technology architecture encourages the use of open technology standards and global technology interoperability. Information Technology Architecture can also be called a high-level map or plan of the information assets in an organisation, including the physical design of the building that holds the hardware."*

It is fair to say that architecture, framework, reference model, scheme, design and fabric are all used with various meanings in the literature. It is generally agreed that the architecture describes – as a design or a model – the data structures and semantics, the software components, the compositions and workflows and the interactions between components and users as functional aspects. It also describes non-functional aspects – usually treated as constraints – for security, privacy, rights, costs, performance.

In the case of ENVRIplus the different RIs are in very different stages of maturity. Some plan to offer (and some indeed already offer) a user portal to access datasets and – in a few cases – processing capabilities. Some provide APIs to processing and urls or other addressing mechanisms to datasets.

It is assumed that ENVRIplus will offer a reference architecture and standard component software, i.e., a toolkit, for constructing an access mechanism – probably through each one's

[105] https://en.wikipedia.org/wiki/Information_technology_architecture
[106] https://en.wikipedia.org/wiki/Information_technology
[107] https://en.wikipedia.org/wiki/Unified_Modeling_Language
[108] https://en.wikipedia.org/wiki/Architecture_framework
[109] https://en.wikipedia.org/wiki/The_Open_Group

portal – to each other's e-RI. Given the use of standard software components and a standard architecture this should allow peer-to-peer interoperable access among ENVRIplus e-RIs and a superset system (outside of ENVRIplus) to provide a 'virtual research environment' capability providing user-driven interoperation across the various e-RIs as envisaged for the EU VRE4EIC project[110], which has ENVRIplus and EPOS as project partners through UvA and INGV (Figure 11).

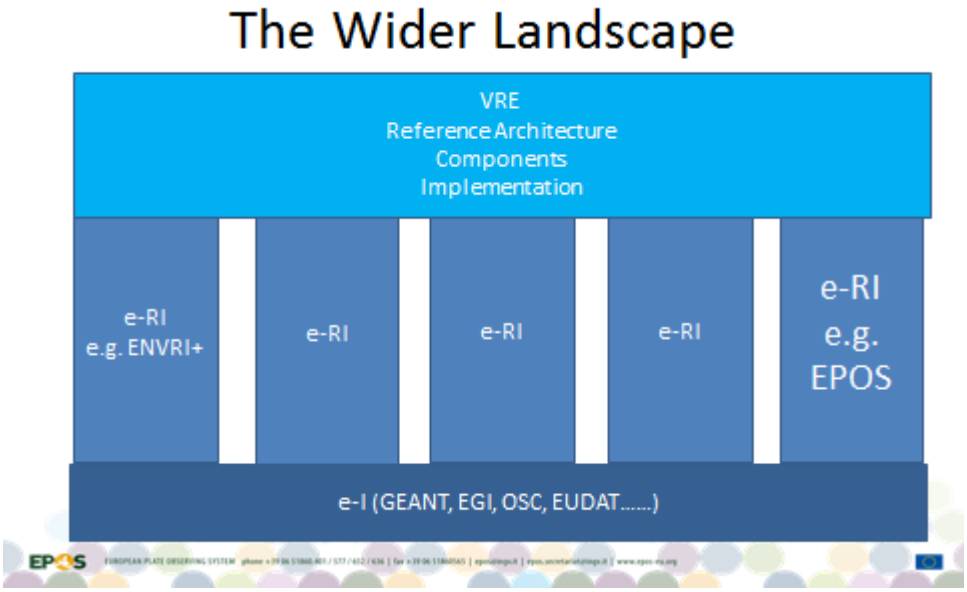

FIGURE 11: THE WIDER LANDSCAPE

This has implications in terms of what each e-RI (at ENVRIplus level or within-ENVRIplus e-RIs) need to provide to allow (a) portal and API access to the e-RI within ENVRIplus; (b) portal and API access to ENVRIplus acting as a portal across its e-RIs; (c) access from a VRE to e-RIs such as ENVRIplus (Figure 12).

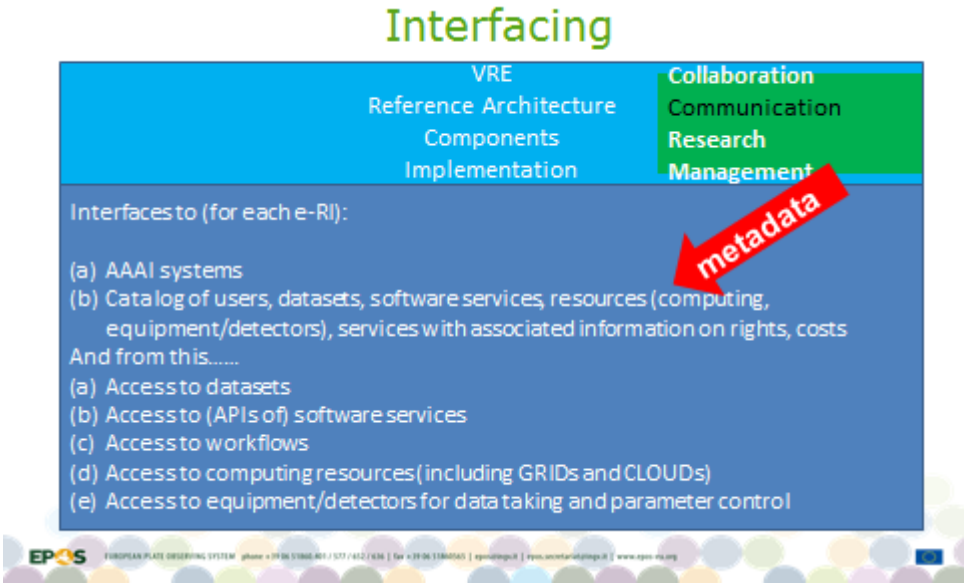

FIGURE 12: INTERFACE REQUIREMENTS

The e-RIs may themselves have user and API access to the RIs within them through a portal such as the ICS (Integrated Core Services) being constructed within the EPOS-IP project (Figure 13).

[110] http://www.vre4eic.eu/

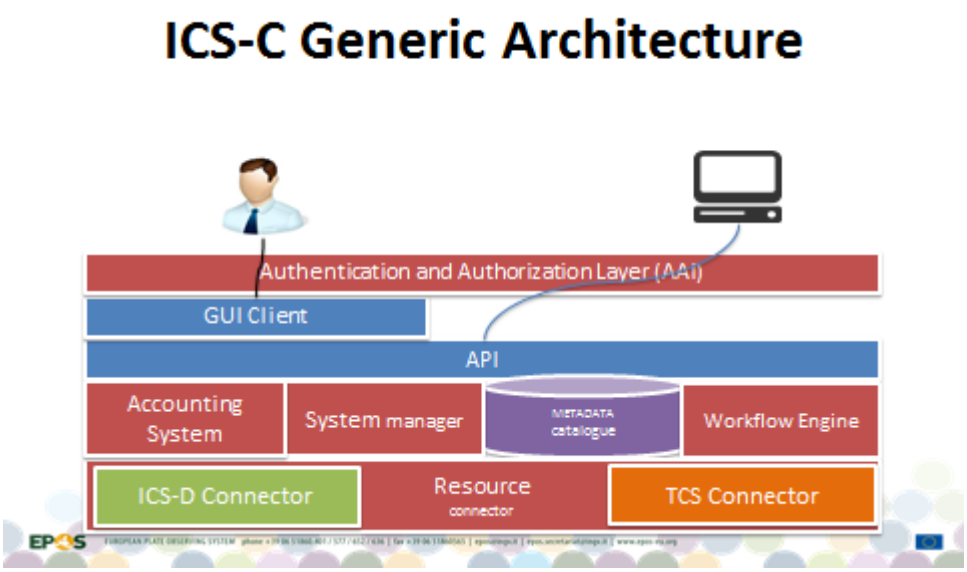

FIGURE 13: EPOS-IP ICS

A first step would be to agree a (set of) Reference Configuration(s) expanding on these concepts and relationships between them, to permit relevant interfaces to be properly specified.

This nicely illustrates conflicting pressures that an e-Infrastructure architecture needs to resolve. The primary goal is to improve researcher productivity as far as possible. For most researchers this requires consistency, automation, tuning in their virtual research environment and the portals and tools they use to do their work. They may further tune this environment using group and personal preference setting. To achieve inter-RI or cross-discipline access and work, they would each like to remain in this productive context, and access data and tools from outside their RI through the same interfaces and with the same tools. However, those architecting the e-Infrastructure for their RI and for the other RIs need to consider feasibility and sustainability. If a direction taken leads to engineering that is too difficult to build, that demands excessive resources, or is so complex that it cannot be maintained for the longer term – see Section 4.2.4 – then the architects must steer the e-Infrastructure away from this—*it is their responsibility to take the long-term view and balance concerns.* A succession of pair-wise arrangements between RIs for specific data can lead to such problems. Initially there are a small number of maintained data integration relationships, but potentially there are n(n-1) such pairings for a wide range of categories of data, and *a non-linear growth in workload is not sustainable.* The architect therefore has to:

1. Recognise such potential problems, explain them and persuade RI construction teams not to start down such paths.
2. Devise a compromise strategy that balances the conflicting issues in a way that is acceptable to the clients, in ENVRIplus's case the RIs.
3. Once the clients' decision makers have agreed, explain that solution to the construction teams and monitor the construction sites to ensure compliance.

A possible compromise solution in this case may be achieved by integrating the ideas of the data-intensive federation framework (DIFF) – see Section 4.2.3 – with the approach adopted by EU VRE4EIC project described above. The DIFF provides a set of APIs that are consistent and sustained, that the portals and tools that each RI uses can interface with. It also hosts and enforces the rules agreed between data providers on the use of their data and a growing repertoire of recipes for data translation that it applies on behalf of its clients. Building, maintaining and supporting those recipes by hand is also likely to prove infeasible, but VRE4EIC uses description-driven logic to generate the recipes. Hence, a compromise is feasible, following such a path. However, architects also have to deliver against the clients scheduling expectations.

The feasible path still involves R&D and so may not be available in time. In which case a temporary solution has to be fabricated that leaves an opportunity to fit the better compromise in later.

This exposes the critical importance of decision making. The above decisions have long term consequences, affect many individuals and organisations, and may affect costs substantially and delay or improve time to production. Such decisions need to be taken by properly constituted bodies. The decisions need investment in investigations and evidence collecting, and then judgement drawing on all aspects of the relevant expertise. See Section 5.2 item 8 for further discussion of decision making issues and the end of Section 4.2.14 for an example of some critical decisions and how they may be partitioned.

3.8.2 Sources of state of the art technology information used

UML, although a (graphical) language standardised by OMG (Object Management Group) in fact causes the architect/designer to consider carefully the architecture of an ICT system[111]. UML has static structure diagrams, dynamic behaviour diagrams and interaction diagrams. It is based on object-oriented approaches and as such suffers from the close integration of data and processing. Extended entity-relationship modelling provides a formalism for structure diagrams which is process-independent.

ODP (Open Distributed Processing) has a reference model (RM-ODP)[112] based on the ANSA project of Andrew Herbert 1984-1998. It has the concept of viewpoints: Enterprise, Information, Computational, Engineering, and Technology. In support of flexibility and independence, the languages used to express the concepts and relations in each viewpoint are abstract. They are not directly translatable to formal system specifications, or to software or data structures in a physical system, which is the aim of CASE (Computer-Aided Software Engineering) Systems. Something additional is needed. Unified Modelling Language (for example) can be used to practically represent ODP-oriented systems designs. UMP Profile Plugins based on UML4ODP [ISO/IEC 19793:2015] make this possible in commercially available IDEs / tool-chains such as those from IBM, Sparx Systems[113], No Magic[114], etc. Model-driven approaches (below) facilitate transformation of computation-independent models (ODP models being one example) to platform-independent models and platform-specific models; again, supported by a wide-range of tools including those from Sparx Systems and No Magic.

MDD: In recent years the concept of model-driven systems engineering (or Model Driven Development) has emerged[115]; the idea is that the system architecture can be described by a model from which the actual system (software, data storage structures and semantics, constraints, …) can be generated (semi-)automatically. With its roots in CASE (Computer-Aided Software Engineering) tools from the 1980s the aim is to translate from conceptual specifications to physical systems thus improving the efficiency of the systems-development process (reduced cost and time) and the effectiveness (improved quality) while retaining the clear linkage to user requirements at conceptual level (enterprise validation).

3.8.3 Short term analysis of state of the art and trends

It is fair to say that the current state of systems development is somewhat chaotic. Fashions come and go, each with a group of enthusiastic consultants claiming to have the 'magic wand' to

[111] https://en.wikipedia.org/wiki/Unified_Modeling_Language
[112] https://en.wikipedia.org/wiki/RM-ODP
[113] http://sparxsystems.com/
[114] http://www.nomagic.com/
[115] http://www.theenterprisearchitect.eu/blog/2009/11/25/15-reasons-why-you-should-start-using-model-driven-development/

make systems development rapid, inexpensive, high quality and matching user requirements. However, some immediate trends are clear:

1. Virtualisation: the user neither knows nor cares where and how the information processing is done as long as their requirements (functional and non-functional) are respected in Service level agreements, quality of service agreements etc.
2. Interoperation: to satisfy the desire for end-users to be able to access not only resources in their domain of interest but across domains –such as the 20 e-RIs of ENVRIplus;
3. Re-use of standard components of software as building blocks joined together like LEGO; this has implications for standardisation of the structure of APIs and messaging interfaces;
4. The definition of data structures and semantics separately from software – a move away from object-orientation – in order to be able to use generic software components;
5. The need for systems to be distributed, partitioned, parallel and (mobile) client-device-independent;
6. The need for systems to handle data streams from instruments/detectors and for users to be able to control the parameters of data-taking;
7. Composition of software components linked to datasets as workflows with parallel/sequential, distributed/centralised control and exception management.

3.8.4 A longer term horizon

The current trends that adopt virtualisation more widely are likely to continue, as will the demand for increased access, processing and ease of use – increasingly through visualisation and mobile devices. This implies the need for an integrated catalogue to provide a 'view' over the ENVRIplus RIs and within which the ENVRIplus RIs can update to ensure they are appropriately represented. The catalogue would then be used for RI-to-RI interoperation (e.g., a query to the portal of one RI could be extended to the portal of another RI) or an ENVRIplus super-portal could be created utilising the catalogue to form workflows dispatched to appropriate RIs for data access and processing. A toolkit that supports each RI installing a way of federating its response to queries and requests with other RIs may be the best way forward – see Section 4.2.3. Whatever solution path is adopted, it will require maintenance – see Section 4.2.4.

3.8.5 Relationships with requirements and use cases

The existing use cases and derived requirements all point to the need for: a) integration mechanisms to overcome data heterogeneity – both syntactic and semantic; b) improved re-use of common software components at any one RI developed by another RI; c) re-use of workflows perhaps provided as services at each RI; and d) improved best practice in curation and provenance recording.

3.8.6 Issues and implications

The ENVRIplus architecture for interoperability has to accommodate (i.e., provide a superset view) over the heterogeneity of the components RIs in the aspects of data, software components, users, resources (computers, equipment). The heterogeneity may be encapsulated at each RI within services, ideally common in functionality and non-functional aspects across all RIs but implemented specifically at each RI. However, much research use is likely to be working with other services so the canonical common catalogue will be required with appropriate software to provide access to the assets recorded there in order to construct workflows to meet end-user requirements.

A simple reference configuration embodying the concepts and relations expressed in Figure 11 and Figure 12, and explained in the text can assist to reach common standing of the points at which standard interfaces / APIs need to be specified.

Further discussion of the architecture technologies can be found in Section 4.2.11. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

3.9 Technologies for semantic linking

Paul Martin, Universiteit van Amsterdam (UvA).

3.9.1 Introduction, context and scope

The role of the semantic linking model is to provide a framework for translating between the different standards used for data and process specification in the environmental sciences in the context of the ENVRI reference model. This model should provide a formal basis on which to improve the interoperability of RI services and products, by focusing on the vocabularies used by the ENVRI RIs, feeding into the design of the abstract architecture for interoperable RIs in general. The model also serves to provide the machine-readable formalisation of the ENVRI reference model (or at least its concept model).

Ultimately, based on the relevant task (5.3) of the ENVRIplus project, we will need to:

1. Capture the conceptual vocabulary of the ENVRI reference model and the correspondences between different concepts described by different viewpoints.
2. Define a framework by which existing standards, taxonomies and ontologies can be mapped to the reference model and via the reference model to each other.
3. Provide tool support for defining new mappings between standards, and for searching the semantic space defined by the resulting interlinking.

Thus the purpose of the technology review in ENVRIplus from the linking model perspective is to determine what technologies are available for ontology specification and formal verification, and what technologies exist that could help us to develop new (or adapt existing) tools.

3.9.2 Short term analysis of state of the art and trends

Combining all environmental domains into one single RI is neither feasible in development nor manageable in operation. During the past several years, interoperability between infrastructures has been extensively studied, with different interoperability solutions proposed for different levels of interoperation: between computing infrastructures [Charalabidis 2012], [Ngan 2011], between middleware [Blair 2012], and between computational workflows [Zhao 2006]. These solutions iteratively build adapters or connectors between two infrastructures and then derive new service standards via focusing community efforts. Such iteration promotes the evolution of services in infrastructures, but cannot fully realise infrastructure interoperability while these solutions only focus on specific layers of the global problem without considering the overall e-science context [Riedel 2009]. Meanwhile, White et al. [White 2012] argued the importance of an ontological reference model in the development of interoperable services in infrastructure.

The linking framework for ENVRIplus [Martin 2015] is being founded on semantic web technologies [Berners-Lee 2001], though the core principles are *technology-agnostic*. Key among these technologies is the Resource Description Framework (RDF)[116] that has come to be used as a generic means to describe information implanted in web resources; building upon RDF, the Web Ontology Language (OWL)[117] is a knowledge representation language used to describe ontologies, and is a significant factor in many semantic infrastructure modelling projects [Zhao

[116] http://www.w3.org/TR/rdf-syntax-grammar
[117] http://www.w3.org/TR/owl-overview

2011], [Baldine 2010]. Within ENVRIplus, the core of the linking framework would be the OIL-E ontologies, which are described in OWL. OWL is well-used in the semantic description domain, but limitations of OWL include the inability to describe integrity constraints or perform closed-world querying [Motik 2006], which might otherwise be useful in (for example) certain well-prescribed areas of the ENVRI reference model. There are also various problems with dealing with diverse schemas, incomplete metadata and the limitations of query interfaces [Gölitz 2007].

The notion of mapping out the topology of standards in environmental science, research practice and e-Infrastructure reflects very much the linked open data approach. The linked data approach offers certain advantages, such as ensuring openness, shareability and reusability [Ferris 2014]. There is however a lack of good tool support for linked data solutions [Enoksson 2009], which is one of the areas that Task 5.3 is intended to address.

Semantic linking is often investigated in the context of ontology matching, mapping or alignment. The key task is to compare similarity between entities from different semantic models and measure the similarity distances at different layers: the *data* layer, comparing data values and objects; the *ontology* layer, comparing the labels and concepts of entities; and the *context* layer, comparing semantic entities with inclusion of application contexts. We posit that the five viewpoints of the ENVRI reference model are applicable for grouping the different modelling contexts of concern to environmental science research infrastructures.

Different metadata standards have been observed from those RIs that are in operation, including NASA DIF [Miled 2001] and SensorML[118] in EMSO, ISO 19115 [ISO 2014] geospatial metadata in SeaDataNet and ISO 19139 [ISO 2007] geospatial XML in EUROGOOS, and a combination of ISO 19115, INSPIRE[119] and NetCDF-CF[120] based standards in IAGOS [Boulanger 2014]. In addition, we have observed the use of Dublin Core [ISO 2009], ISO 19156 [ISO 2011], SeaDataNet Cruise Summary Reports[121] metadata, CERIF [Jeffery 2014], and CSMD[122]. These standards can be linked via the information viewpoint of the ENVRI reference model and mapped to functional subsystems of RIs. There is prior work mapping information viewpoint concepts in the reference model to concepts found in those standards [Zhao 2014].

The typical process for semantic linking involves several iterations of the following steps: 1) pre-processing of features by a small set of excerpts of the overall ontology definition to describe a specific entity; 2) definition of the search space in the ontology for candidate alignment; 3) computation of the similarity between two entities from different ontologies; 4) aggregation of the different similarity results of each entity pair, depending on the algorithms used; and 5) derivation of the final linking between entities using different interpretation mechanisms, including the analysis of human experts.

The linking component of OIL-E glues concepts both *inside* ENVRI-RM and *between* ENVRI-RM and external concepts belonging to outside vocabularies. The ENVRI-RM ontology only contains a limited set of vocabularies derived from common functionality and patterns, so linking ENVRI-RM with external RI-specific concepts will enable RI-specific extensions to the ENVRI-RM vocabulary. Similarly, linking ENVRI-RM with external vocabularies provides bridge between those vocabularies and ENVRI-RM, and indirectly between the vocabularies themselves. Notably, the internal correspondences between different ENVRI-RM viewpoints (enterprise, information, etc.) can potentially be used to indirectly link external vocabularies of quite different foci (data, services, infrastructure, etc.).

Distributed applications and systems can be described using published ontologies, permitting services both internal and external to a system to potentially interact with application

[118] http://www.opengeospatial.org/standards/sensorml
[119] http://inspire.ec.europa.eu/
[120] http://cfconventions.org/latest.html
[121] http://www.seadatanet.org/Standards-Software/Metadata-formats/CSR
[122] https://icatproject.org/user-documentation/csmd/

components without having had to be explicitly designed to do so, provided that they can process the ontology used to describe the component.

There already exists work on doing this kind of semantic modelling of computing and network infrastructure, however the modelling of applications running on cloud platforms is less well-developed—in [Ortiz 2011], the author articulates some of the challenges facing standardisation of cloud technologies, and the lack of concrete formal models is a major factor. Even excluding the cloud however, information models for modern computing infrastructure are often lacking in some dimension. For example, modern infrastructure modelling languages must be able to model virtualisation and management of virtualised resources as well as physical resources.

In [Ghijsen 2013], the authors describe the Infrastructure and Network Description Language (INDL), a product of the Open Grid Forum (OGF) Network Markup Language Working Group (NML-WG). INDL is designed to be extensible, linkable to existing information models, and technology independent. NDL-OWL [Baldine 2010] provides a Semantic Web model for networked cloud orchestration modelling network topologies, layers, utilities and technologies. It extends the Network Description Language upon which INDL is based and uses OWL. Meanwhile [Zhao 2010] presented a workflow planning system called NEtwork aware Workflow QoS planner (NEWQoSPlanner) based on INDL; NEWQoSPlanner is able to select network resources in the context of workflow composition and scheduling.

3.9.3 A longer term horizon

The generation of formal descriptions for complex entities is essential for the mechanisation of processes involving those entities—this is not in question. What is in question is the extent to which different systems can be integrated within common models with shared vocabularies, and to what extent we must accept the existence of proliferation of alternative models, and thus have to expend effort in bridging between the resulting heterogeneous concept spaces.

3.9.4 Relationships with requirements and use cases

The linking model is strongly tied to the reference model, which provides its core vocabulary. The linking model should also itself contribute vocabulary and relations that are useful for the interoperable architecture design task.

Regarding use-cases, any of the use-cases might benefit from a linking of formal descriptions, depending on the extent to which the use-cases cross between domains, or make use of formal descriptions that need linking to the reference model concepts. Particular ENVRIplus cases[55] where linking between different existing standards and vocabularies might be useful include:

- Identifying **trends in the emergence of mosquito born diseases** requires interaction between a number of different data centres and compute providers.
- The **description of a national biodiversity data archive centre** requires a formal model for how data from a national facility is to be delivered to and integrated with Europe-wide data providers.
- **Domain extension of existing thesauri**.

3.9.5 Issues and implications

The question that underlies the semantic linking task is: how do we make it easier to map between different vocabularies? Autonomous mapping processes are highly error prone, and extremely sensitive to the quality of the underlying taxonomies or ontologies. Manual mapping requires expert oversight, but can be supported by tools. Current work on OIL-E to and from CERIF mapping within the VRE4EIC project[110] should yield useful results here.

The base contribution of a linking model in the environmental science research infrastructure domain is the ability to map out the space of existing standards, models and vocabularies being used in different datasets, architecture designs, instrument specifications, service profiles, etc. used by different research communities, and the ability to associate them via the viewpoints of the ENVRI reference model or its successors. This in and of itself would constitute a useful contribution, since as it stands it requires substantial research to truly understand the full current research landscape, and even experts' views are often narrow, focused on a particular domain or a particular geographic region (i.e., the standards produced within their home continent).

Further discussion of the semantic linking technologies can be found in Section 4.2.12. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

3.10 Technologies for the reference model

Alex Hardisty and Abraham Nieva de la Hidalga, Cardiff University (CU).

3.10.1 Introduction, context and scope

So, what is a Reference Model (RM)? A good place to start is with a Wikipedia article on reference models[123]. Its opening paragraph explains an RM as "***an abstract framework consisting of an interlinked set of clearly defined concepts produced by an expert or body of experts in order to encourage clear communication. A reference model can represent the component parts of any consistent idea, from business functions to system components, … …***". It goes on to say that an RM can "***… then be used to communicate ideas clearly among members of the same community***". This then, is the essence of an RM. It's a descriptive conceptual framework, establishing a common language of communication and understanding, about elements of a system and their significant relationships, within a community of interest. That's particularly important when, as in the environmental research infrastructures (RI) sector that community of interest brings together significant numbers of experts from vastly different scientific and technical backgrounds to talk about building distributed ICT infrastructures.

The present topic is concerned principally with the ENVRI Reference Model[124] and is closely related to the topic of the Linking model (see Section 0), which depends upon it. However, reference models are cutting across all aspects of infrastructure design and technology review. Thus, this topic relates to all the topics of the technology review (see Section 3).

3.10.2 Sources of state of the art technology information used

Wikipedia provides general introductory level information on reference models and reference architectures. ISO/IEC publishes relevant international standards. Various Web resources have been used and are mentioned / linked in the text. Other sources are directly referenced from the text and listed in the bibliography.

3.10.3 Short term analysis of state of the art and trends

State of the art

The ENVRI Reference Model (ENVRI RM)[124] is presently work in progress. Based on RM-ODP [ISO/IEC 10746], version 1.1 has been published in summer 2013 as a deliverable of the ENVRI project. It is derived from commonalities of requirements collected from 6 research

[123] https://en.wikipedia.org/wiki/Reference_model
[124] http://envri.eu/rm

infrastructures. In the ENVRIplus project there is a task 5.2 to review and improve the RM, based on new requirements analysis of 20 research infrastructures[125]. At present the ENVRI RM is introduced through a sub-systems view of research infrastructure but this needs to shift to a data lifecycle oriented approach. The sub-systems perspective has to be more properly assigned only within the Engineering Viewpoint where it can support the complete lifecycle of research data (from design of experiments that produce new data through acquisition, curation and publishing of that data, to its use in processing and analysis to reach scientific conclusions) according to specific scope and needs of individual RIs.

Moving forward with the RM in ENVRIplus

Use of reference models, and particularly viewpoint models such as ENVRI RM keeps the design discussion centred at the right level (see remarks on raising the level of discourse) while accommodating the perspectives of difference stakeholders. They allow moving from a high level description of RIs for researchers and sponsors that is founded on the science to be carried out, to a lower, more detailed design level for IT developers and technicians, concerning engineering and technology aspects. By using the ENVRI RM, RIs can create a set of models that separate concerns neatly but at the same time keep the consistency of the RI systems as a complete entity, as well as accommodating relevant policy constraints.

Validating the present ENVRI RM based on review of requirements from a wider set of RIs and completing and evolving the RM for easier use are main activities in the ENVRIplus project now. Another important activity is to explore ways in which RIs communities can be helped, and assisted to become self-sufficient. Working in conjunction with several use cases teams (see below) and producing specialised e-learning materials are two strands of planned activity. As well as delivering content specifically about the 'internals' of the ENVRI RM, training will also give guidance for different situations on how to use various parts of the RM. This will be very much driven by case examples and, over time we expect to see emergence of common re-usable patterns that can be applied elsewhere.

It would be interesting to find an early adopter RI prepared to invest in exploring the potential of the available tools (see above), casting a model in UML4ODP perhaps.

Problems to be overcome: Adoption

In the research infrastructures sector we have to move to an RM oriented approach for three reasons. Firstly, so that we can achieve interoperability within and between different infrastructures. Secondly, because there are multiple players and stakeholders in the sector that have to work together and talk to one another. And thirdly, so that the sector can achieve the economies of scale within and across infrastructures that we need for attracting the attention of industry. There is a role for bespoke design and development due to the unique attributes of individual infrastructures but wherever possible, off-the-shelf capabilities should be adopted first. We can do this more easily when we have a commonly accepted conceptual foundation upon which to base procurement. Achieving a shift in culture and mind set of the community is a significant issue to be overcome. It needs to balance the costs of replacing existing technology and the consequent impact on working practices with the long-term costs of support and maintenance – see Section 4.2.4

Problems to be overcome: Complexity

RMs are a systems modelling way of thinking that draws together all the conceptual elements and relationships in a large class of very complex distributed systems. Systems thinking gives us a means to cope with that complexity. It helps us to better deal with change in the (scientific) business, leading to more agile styles of thinking and response. Understanding relationships between the various parts of a research infrastructure helps us to understand the possible collective (emergent) behaviours of the infrastructure and to practically engineer and manage

[125] The present document, so the task is just commencing at the time of writing this.

real systems. Thus (and according to APG) a reference model is really a framework from which a portfolio of services can be derived.

Complexity can be off-putting. [Hardisty 2015] has suggested ways to engage with RMs for the first time and how, particularly to get the best out of the ENVRI RM124. A Forbes article on Enterprise Architecture [Bloomberg 2014] also offers several suggestions that are transferable to the present context. You don't have to take reference models too literally. You don't have to "do" all of the RM to benefit from it. Just pick and choose what works for you. It's basically a toolkit. You can use it in several different ways - to baseline what you already have and to clean up; to target desired outcomes and plan out how to achieve them; or in combination to deal with a troublesome area (pain point) – first by baselining it, then by targeting it and then iterating until the pain has gone away.

Problems to be overcome: Tooling and skills development

Effective software systems engineering depends on having robust and capable Integrated Development Environment (IDE) within which all the processes of software design, implementation and test can take place. As noted above, industry-standard design tools are beginning to support the necessary concepts but their penetration and use in research infrastructures sector is still quite low. The level of architecting skills to be found among practitioners in research infrastructures is also quite low. This has to be addressed by targeted recruiting and specialised training.

Use of RMs in other sectors

RMs have been used widely in the telecoms, healthcare and defence sectors, as well as among architects of enterprise and public sector systems. All these sectors are characterised by their need for "infrastructure at scale". They involve multiple vendors who have to work, if not together then to a common framework of principles and concepts to bring about widespread interoperability. It's easy to make a phone call to more or less anywhere on the planet, or to receive streaming video there. That is the result of using reference models and standardising interfaces between sub-systems and components from different vendors.

One view of reference models, particularly expressed by practitioners at Armstrong Process Group (APG)[126] is that they are a 'supporting capability' in the Enterprise Architecture value chain. Putting that into the ENVRI context is to say that RMs have relevance to and use for understanding and analysing the environmental science enterprise prior to and as part of planning and implementing (engineering) research infrastructures.

During 2013 the ESFRI cluster projects covering the biomedical sciences (BioMedBridges), physics (CRISP), social science and humanities (DASISH), and environmental sciences (ENVRI) came together to identify common challenges in data management, sharing and integration across scientific disciplines [Field 2013]. Reference models were identified as a common interest of all the clusters. Subsequently, RMs were ranked as one of the top three issues needing to be addressed jointly across all RIs at the European level.

UML4ODP and tooling for software / systems engineering

Recently revised, UML4ODP [ISO/IEC 19793:2015] allows systems architects to express their systems architecture designs in a graphical and standard manner using UML notation. This is exciting because it means, for example that the ENVRI RM and all its concepts can be built into software engineering IDEs[127] with all that implies for inheritance, compliance with agreements and standards, etc. This makes it possible for industry-standard model-based systems engineering tools, such as Sparx Systems' Enterprise Architect[128], IBM Rational Software Architect[129] or MagicDraw[130] to deal with ODP based designs and thus to inherit concepts from

[126] http://www.aprocessgroup.com/
[127] https://en.wikipedia.org/wiki/Integrated_development_environment
[128] http://www.sparxsystems.com/products/ea/index.html
[129] http://www-03.ibm.com/software/products/en/rational-software-architect-family

an RM once that RM is encoded as a UML4ODP representation[131]. This has been explored, for example in the healthcare context by [Lopez 2009]. However, as far as we know there are no open-source IDE tools specifically supporting UML4ODP at this time. Eclipse[132] has general support for UML but not specifically for UML4ODP.

On the other hand, the ODP and ENVRI reference models can also be represented as an ontology (see Section 0) expressed, for example in OWL and RDF, which means it can then be used in a knowledge base over which reasoning can take place. This has multiple applications.

Supporting the European Open Science Cloud (EOSC)

Early in April 2016 a High-Level Expert Group reported its strategic advice on the future European Open-Science Cloud (EOSC)[133] to the European Commission. "***By mapping the route to a European Open-Science Cloud***", says expert group member Paul Ayris, "***the group's ultimate goal is to create a trusted environment for hosting and processing research data to support world-leading EU science. Cloud computing can change the way that research in Europe is done. The creation of an open-science commons would allow European researchers to collaborate, share and innovate using shared infrastructures, tools and content.***"

EOSC[134] is envisioned as a federated environment, made up of contributions from many stakeholders at both national and institutional levels. The desire for minimal international guidance and governance, combined with maximum freedom of implementation means that moving towards some kind of framework of reference as the basis of the open science commons[135] is inevitable. Robust standards for exchanging information between different heterogeneous parts of the federated cloud environment will be paramount. Developing these in an open and transparent manner will be difficult and costlier without a framework of reference (such as the ENVRI Reference Model) within which to situate them.

The ENVRI RM can be used for describing the EOSC.

On one level, there is an implied assumption that cloud computing (as understood in common parlance) is the basis of the EOSC. This is a technology assumption (and therefore also partially an Engineering assumption). However, the true scope of EOSC has to be thought of in terms much wider than just technology and engineering; especially as the former is subject to rapid evolution. Consideration has to be given to the business of the EOSC, to the data and information it is expected to handle, and to the nature of the computation (in its widest sense) to be applied in order to create the '***trusted environment for hosting and processing***'.

EOSC implies more than is just meant by the term "cloud", as often used in common parlance to mean cloud computing. EOSC bundles: a) financial and business models, that are Science viewpoint; b) data and information to be handled, that are Information Viewpoint; c) shared provisioning, operations management, and systems support that is organisational, and involves multiple viewpoints; d) a hardware-level protection regime, involving Engineering and Technology viewpoints; e) a whole open-ended set of ways of building and deploying executable machine images; which has Computational, Engineering and Technology; f) a range of ways of allocating resources and scheduling work, again Computational, Engineering and Technology viewpoints; g) a variety of AAAI strategies; and h) a variety of collaboration and isolation regimes. EOSC will not be a single platform or a single technology but a heterogeneous collection of virtual and dynamic configurations responding to the circumstances of the moment. Initiatives

[130] http://www.nomagic.com/
[131] The Information Viewpoint of the ENVRI RM has been cast in UML4ODP during its development in the ENVRI project. Work remains to be done to cast the other viewpoints in UML4ODP.
[132] https://eclipse.org/
[133] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[134] Announced by the EC on 19th April 2016, http://europa.eu/rapid/press-release_IP-16-1408_en.htm?locale=en
[135] https://www.opensciencecommons.org/

such as Kubernetes[136], for example and our own ENVRI Linking Model (0) are exploring ways of developing smart mappings to cope with this.

Cloud is not easy, certainly if you're doing most of the things the RIs are expected to be doing. For them, using methods with the ENVRI RM to unpick the elements that make up cloud might be useful.

Alignment to Research Data Alliance (RDA)

By engaging the scientific communities to address the issues such as data identification and citation, discovery, access, sharing, etc., the Research Data Alliance (RDA)[137] has a role to further promote the maturation and adoption of practices for open research data and open science.

One product of RDA thus far is the results from its Data Foundation and Terminology Working Group [DFT: Results RFC]. This is a set of core terms for classifying data objects and repositories, and a model of relationships between the terms. These DFT core terms correspond more or less with some main concepts in the Information Viewpoint of the ENVRI RM but the scope is limited to that.

Part of the envisaged evolution of the ENVRI RM during the ENVRIplus project will involve RDA alignment.

In general terms, the "digital transformation agenda" (encompassing cloud infrastructure, continuous delivery of IT services, DevOps, agile software development, etc.) acts as a significant driver. Bots, services, APIs and apps - this is a catch-all for the general trend in consumer computing towards a world of smart applications, interacting with services (both bot and human) via a range of APIs. Knowing all the APIs, where they are and how they relate to one another in terms of compatibility and composition potential will be a crucial development to watch as it spills over from mainstream consumer computing into enterprise and academic/research sectors. To what extent do current RMs overtly accommodate this trend? To what extent do RIs realise the impact it will have for them? One possible argument is that it's just engineering and that all the logical stuff is already provided for.

Wider uptake and dependence on RMs for design, planning and change management becomes apparent. Design patterns, based on a widely accepted conceptual understanding of the archetypical architecture(s) of research infrastructures become more prominent.

Architectures become agile and dynamic, requiring continuous re-appraisal and evolution of RMs to suit new circumstances.

3.10.4 Relationships with requirements and use cases

TC_16 Description of a National Marine Biodiversity Data Archive Centre[138] seeks to integrate the DASSH Data Archive Centre[139] with other European marine biological data (e.g., data curated by EMSO, SeaDataNet, JERICO and EMBRC) as a joint contribution to EMODNET Biology, the COPERNICUS provider. This is a typical test case for the ENVRI Reference Model.

Using the ENVRI Reference Model (RM), IC_12 Implementation of ENVRI(plus) RM for EUFAR and LTER[138] seeks to describe two RIs with (in part) very different framework requirements. EUFAR (European Facility for Airborne Research) is an emerging RI to coordinate the operation of instrumented aircraft and remote sensing instruments for airborne research in environmental and geo- sciences. LTER (Long-Term Ecosystem Research) is a global effort aiming at providing information on ecosystem functioning and processes as well as related drivers and pressures on ecosystem scale (e.g., a watershed).

[136] http://kubernetes.io/
[137] https://rd-alliance.org/
[138] https://wiki.envri.eu/display/EC/Use+Cases
[139] A UK national facility for archival of marine species and habitat data, http://www.dassh.ac.uk/

A number of other use cases[138] (for example: SC_3, TC_2, TC_4, IC_3) would probably also benefit from applying RM thinking and concepts in their analysis and design. Each of these use cases contains one or more detailed scenario descriptions and explanations that could benefit from being thought about from the different viewpoints of science ("the business"), information and computation. Ultimately, engineering and technology aspects also become important.

3.10.5 Issues and implications

Reference Models (RM) and the ENVRI RM in particular have a significant role to play in fostering the use of common language and understanding in the architectural design of environmental research infrastructures. Adoption and use contributes significantly towards the goal of interoperability among research infrastructures. However, there are social barriers to be overcome. These have to be addressed by marketing, education and training.

Lack of training is a key issue, and with it the lack of skilled architects.

RMs have been ranked by the first round of ESFRI research infrastructure cluster projects as one of the top three issues needing to be addressed jointly across all RIs at the European level.

Further discussion of the reference model technologies can be found in Section 4.2.13. This takes a longer term perspective and considers relations with strategic issues and other technology topics

3.11 Technologies for providing compute, storage and network resources

Yin Chen, EGI. Alex Hardisty, Cardiff University (CU).

3.11.1 Introduction, context and scope

What are *e-Infrastructures*? The e-Infrastructure Reflection Group (e-IRG) [e-IRG White Paper 2013], defines them to include: access to high-performance computing and high-throughput computing; access to high end storage for ever increasing data sets; advanced networking services to connect computing and storage resources to users and instruments; middleware components to enable the seamless use of the above services, including authentication and authorisation; and generic services for research, providing support for research workflows using combinations of the above (sometimes called virtual laboratories or virtual research environments). In particular, it envisions e-Infrastructures where the principles of global collaboration and shared resources are intended to encompass the sharing needs of all research activities.

The European Strategy Forum on Research Infrastructures (ESFRI) presented the European roadmap[140] for new, large-scale Research Infrastructures. These are modelled as layered hardware and software systems that support sharing of a wide spectrum of resources, spanning from instruments and observations, through networks, storage, computing resources, and system-level middleware software, to structured information within collections, archives, and databases. The roadmap recognises that the special “e-needs” of research infrastructures should be met by e-Infrastructures.

Environmental and Earth sciences have been supported by national and institutional investments for a great many years. These have led to a diversity of significant computing resources and support services that are the precursors of today’s pan-European e-Infrastructure. They now coexist with and participate in today’s pan-European e-Infrastructures.

[140] ESFRI Roadmap: http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

The contemporary supported strategies lead to the development of e-Infrastructures in Europe, connecting them into continent-wide e-Infrastructures. This is to allow researchers from different countries to work together using shared resources, including computers, data and storage. Important pan-European large-scale e-Infrastructures include: EGI, EUDAT, PRACE, GÉANT, OpenAIRE, and Helix Nebula. Each has own focused areas, e.g., EGI provides pan-European federated computing and storage resources; PRACE federates pan-European High Performance Computing (HPC) resources; EUDAT focuses on providing services and technology to support the life-cycle of data. GÉANT is the pan-European data network for the research and education community, interconnecting National Research and Education Networks (NRENs) across Europe. OpenAIRE is a network of Open Access repositories, archives and journals that support Open Access policies. The Helix Nebula initiative is providing a public-private partnership by which innovative cloud service companies can work with major IT companies and public research organisations. These e-Infrastructures provide generic IT resources and services solutions to support multiple European scientific research activities. The benefits to adopt and make good use of these resources for a scientific community and a research infrastructure include:

- Having ready-to-use compute and storage resources and services solutions for scientific collaborations;
- Avoiding duplicated development and effort;
- Enlarged community network and user bases – since these pan-European e-Infrastructures have already been attracting many international collaborations and users;
- Sharing state-of-art experience by research communities already using the e-Infrastructure.

This section gives an overview of current e-infrastructure for European research, along with some of the forthcoming developments and innovations. The focus is on pan-European scale infrastructure broadly classified into high-throughput computing (HTC or "cloud"; e.g., EGI), high-performance computing (HPC; e.g., PRACE), open-access publications repositories and catalogues (Pubs; e.g., OpenAIRE) and data storage and services (Data; the EUDAT CDI). The figure also includes a social dimension, characterising interactions by expert groups. The focus reflects the pan-European scale of the Research Infrastructures (RI) represented in ENVRIplus.

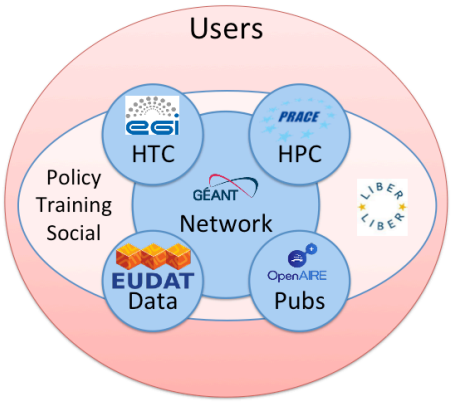

FIGURE 14: CLASSIFYING EUROPEAN E-INFRASTRUCTURES

In general, all of the current European scale e-infrastructures seek to include partners in all European Member States, thereby providing a one-stop-shop for continental-scale interactions while at the same time providing access to local and regional activities in the individual Member States. At a European level, the e-infrastructure is often presented in terms of:

1. Computer networking
2. High capacity computing and high-throughput computing
3. Data storage and management

4. User tools (Virtual Research Communities and Virtual Research Environments).

In the sections 3.11.2 - 3.11.6 that follow we focus on the first three of these.

3.11.2 Sources of state of the art technology information used

The technology information is provided by e-Infrastructure providers, including: EGI.eu, and CSC (representative of EUDAT). Information also refers to ESFRI Strategy Report on Research Infrastructure Roadmap 2016 [ESFRI 2016].

3.11.3 Short term analysis of state of the art and trends

3.11.3.1 Networking

GÉANT

The model for research and education networking in Europe is of a single national entity per country (the National Research and Education Network – NREN) connecting to a common pan-European backbone infrastructure, GÉANT. In combination these networks provide a powerful tool for international collaborative research projects – particularly those with demanding data transport requirements. NRENs[141] are able to connect individual sites (universities, research centres, other related not-for-profit institutions) to their high-bandwidth infrastructures or arrange point-to-point services for bilateral collaborations. GÉANT provides a single point of contact to coordinate the design, implementation and management of network solutions across the NREN and GÉANT domains. In addition to its pan-European reach, the GÉANT network has extensive links to networks in other world regions including North America, Latin America, the Caribbean, North Africa and the Middle East, Southern and Eastern Africa, the South Caucasus, Central Asia and the Asia-Pacific Region. In addition, there is on-going work to connect to Western and Central Africa[142].

The GÉANT network (like the majority of NRENs) has a hybrid structure – operating a dark-fibre network and transmission equipment wherever possible and leasing wavelengths from local suppliers in more challenging regions. This structure allows the operation of both IP and point-to-point services on a common footprint. Since 2013, GÉANT has migrated to a new generation of both transmission and routing equipment platforms. The resulting network is seen as a significant increase in the bandwidth available along with an improved range of network services. GÉANT's pre-provisioned capacity on each of the core network trunks (covering western and central Europe) is around 500Gbps and an advanced routing/switching platform delivers IP, VPN and point-to-point services with greater flexibility to all European NRENs.

The GÉANT project provides more than just a physical network infrastructure. Its service development and research activities address directly the needs of the research and education community both by providing advanced international services on the NREN and GÉANT backbones, and also by developing software and middleware to target network-related issues from campus to global environments. The GÉANT backbone currently offers:

- GÉANT IP – a high quality IP service providing robustness and high levels of availability, high-bandwidth and global reach.
- GÉANT Plus – point-to-point services offering guaranteed routing, latency and stability on the full GÉANT footprint.
- GÉANT Lambda – offering guaranteed capacity of 10Gbps or 100Gbps on dedicated wavelengths over the GÉANT-operated optical fibre.

[141] GÉANT NRENs list: http://www.geant.org/About/Membership/Pages/MAandGAreps.aspx

[142] Countries interconnect with GÉANT http://www.geant.org/Networks/Global_networking/Pages/Home.aspx.

- VPN (Virtual Private Network) services, which can provide bespoke network architectures for multi-site collaborations.

Services under development in GÉANT include[143]:

- Software-defined networking to facilitate faster and easier network configuration.
- Authentication and Authorisation (AAI) services – designed to address international multi-domain environments.
- A centrally procured cloud service to leverage economies of scale across the European NREN constituency.

3.11.3.2 Computing

PRACE

PRACE[144] provides high-end computing resources to European top science. The largest 3-5 PRACE systems are generally referred to as "tier-0". These systems are in general significantly larger than other European computer systems accessible to researchers. The resources are accessible to applicants based on a twice-yearly Calls for Proposals. Preparatory access proposals, allowing users to develop software or test out novel ideas, are also accepted.

Over a series of 'implementation projects', including pre-commercial procurement PRACE include a range of activities that are interesting for the biological and medical science communities: training courses, software development, HPC technology tracking and access to prototype resources. The fourth implementation project (PRACE-4IP, 2015-2017) is working now towards transition to PRACE 2; strengthening the internationally recognised PRACE brand; preparing strategies and best practices towards exascale computing, coordinating and enhancing the operation of the multi-tier HPC systems and services, and supporting and educating users to exploit massively parallel systems and novel architectures.

It is important to note that the explosion in the data generation capacity of scientific equipment and sensors is creating a new class of researchers who have different demands in terms of their use of high performance computing (HPC) power, and of how and where their data is stored. Traditionally, researchers need PRACE and other similar supercomputing capability/capacity to execute large-scale compute-costly software codes for modelling and simulations. It is often the case that input data needed by these codes is moved ('staged ') to the HPC facility. It may even be (semi-permanently) kept there. The output results are either also kept there or are staged back to the researcher. Results are often used multiple times to compare with other results and models so that they don't have to be re-generated. In contrast, the new type of users wants to process and analyse their data that is too massive (voluminous) to be staged. This introduces new problems around locating HPC close to well-founded repositories where data should be kept. Finding the balance between optimal HPC location (moving execution towards the data), costs of data staging, and changes in community working practices around data deposition is where the challenges lie. See also additional explanation about positioning EUDAT below (in section 3.11.3.3).

EGI

The EGI infrastructure is a publicly funded e-infrastructure giving scientists access to more than 650,000 logical CPUs, 550 PB of storage capacity to drive research and innovation in Europe. Resources are provided by about 350 resource centres distributed across 53 countries in Europe, the Asia-Pacific region, Canada and Latin America. EGI also federates publicly funded cloud providers across Europe for the implementation of an European data cloud to support open science. EGI supports computing (including closely coupled parallel computing normally associated with HPC), compute workload management services, data access and transfer, data

[143] For full details of GÉANT services see http://www.geant.org/Services.
[144] PRACE: http://www.prace-ri.eu

catalogues, storage resource management, and other core services such as user authentication, authorisation and information discovery that enable other activities to flourish. User communities gain access to EGI services by partnering with EGI, either directly through federating their own resource centres, or indirectly by accessing national or regional resource centres that already support their communities.

Existing high-level services:

- Federated IaaS Cloud: Run compute- or data-intensive tasks and host online services in virtual machines or docker containers on IT resources accessible via a uniform interface. Store/retrieve research data at multiple distributed storage service providers. Share applications, tools and software for data processing and analysis.
- High-Throughput Data Analysis: Run compute-intensive tasks for producing and analysing large datasets and store/retrieve research data efficiently across multiple service providers.
- Federated access to computing and data: Manage service access and operations from heterogeneous distributed infrastructures and integrate resources from multiple independent providers with technologies, processes and expertise offered by EGI.
- Consultancy for user-driven innovation: Expertise to assess research computing needs and provide tailored solutions for advanced computing.

High-level services under development:

- Open Data Platform: Store and discover research data, publish with open or controlled access, access and reuse data with the EGI computing services.
- Accelerated computing: Run computational tasks on specialised processors (accelerators) with traditional CPUs from multiple providers allowing for faster real-world execution times.
- Community-specific tools: To provide access to specialised tools for data analysis contributed by the community.

Project positioning with respect to related initiatives

- EUDAT2020: EGI enables reuse of research data available from their services.
- PRACE: EGI complements PRACE HPC services with cloud and HTC capabilities, altogether addressing the different computing needs of the research community.
- GÉANT: EGI relies on connectivity for distributed access to data and computing.
- OpenAIRE: use of dissemination/discovery services for research outputs supported by EGI.
- VRE projects: EGI provides hosting environments for services developed by VRE projects and co-creates community specific tools.
- On-going project such as, INDIGO-DataCloud and AARC: EGI adopt their software and technical solutions.

EGI matured its portfolio of solutions that help accelerate data-intensive research. The most relevant developments in EGI for ENVRIplus are:

1. Launch of EGI Federated Cloud

EGI opened the 'EGI Federated Cloud' as a production infrastructure in May 2014[145]. Based on open standards, it is an interconnected grid of institutional clouds offering unprecedented versatility and cloud services tailored for European researchers. With the EGI Federated Cloud, researchers and research communities can:

- Deploy scientific applications and tools onto remote servers (in the form of Virtual Machine images).
- Store files, complete file systems or databases on remote servers.

[145] http://go.egi.eu/cloud

- Use compute and storage resources elastically based on dynamic needs (scale up and down on-demand).
- Immediately address workloads interactively (no more waiting time as with grid batch jobs)
- Access resource capacity in 19 institutional clouds[146].
- Connect their own clouds into a European network to integrate and share capacity, or build their own federated cloud with the open standards and technologies used by the EGI Federated Cloud.

Since its launch, the EGI Federated Cloud has attracted more than 35 use cases from various scientific projects, research teams and communities. Among these there are several applications from environmental sciences.

2. Simplifying access to EGI for the 'long tail of science'

While processes to gain access to EGI are well established across the NGIs (National Grid Initiatives) for entire user communities, individual researchers and small research teams sometimes struggle to access compute and storage resources for the implementation of their applications. Recognising the need for simpler and harmonised access for individual researchers and small research groups (i.e., 'long tail of science') the EGI community has launched (December 2015) a prototype platform[147] providing integrated services from the NGIs to those researchers and small research teams who work with data but have limited or no expertise in using distributed systems. The platform lowers the barrier to access grid and cloud infrastructure via a centrally operated access management portal and an open set of virtual research environments designed for the most frequent use cases. The project defines security policies and implements new security services that enable personalised, secure and yet simple access to e-infrastructure resources via the virtual research environments for individual users. The platform authenticates users via the EduGAIN federation and other username–password based mechanisms, complementing the long established certificate-based access mechanisms.

3. EGI-Engage

One of the main objectives of the Horizon 2020 funded EGI-Engage project (2015 – 2017, €8.7m) is to expand the capabilities of EGI (e.g., cloud and data services) and the spectrum of its user base by engaging with large Research Infrastructures (RIs), the long tail of science, and with industry/SMEs (Small and medium-sized enterprises). The key engagement instrument for this is a network of eight competence centres, in which National Grid Initiatives (NGIs), user communities, technology and service providers work together to collect requirements, integrate community-specific applications into state-of-the-art services, foster interoperability across e-infrastructures, and evolve services through a user-centric development model. The competence centres provide state-of-the-art services, training, technical user support and application co-development to specific scientific domains. The following science communities (including 3 from environmental sciences) have dedicated competence centres in EGI-Engage:

1. Earth-science research (EPOS)
2. EISCAT 3D
3. Life-science research (ELIXIR)
4. Biodiversity and ecosystem research (LifeWatch)
5. Biobanking and medical research (Biobanking and Bimolecular Research Infrastructure, BBMRI-ERIC)
6. Structural biology and brain imaging research (MoBrain supporting WeNMR and Integrating Structural Biology – INSTRUCT)
7. Arts and Humanity (DARIAH)
8. DisasterMitigation

[146] The number is growing; see up to date values at https://wiki.egi.eu/wiki/Fedcloudtf:ResourceProviders#Fully_integrated_Resource_Providers
[147] https://access.egi.eu/start

The Helix Nebula Marketplace

The Helix Nebula initiative is a public-private partnership by which innovative cloud service companies can work with major IT companies and public research organisations. The Helix Nebula Marketplace (HNX) is the first multi-vendor product of the initiative, delivering easy and large-scale access to a range of commercial Cloud Services through an innovative open source broker technology. A series of cloud service procurement actions, including joint pre-commercial procurement co-funded by the European Commission are using the hybrid public-private cloud model to federate e-infrastructures with commercial cloud services into a common platform delivering services on a pay per use basis.

3.11.3.3 Data

Research Data & EUDAT Services

EUDAT is a pan-European data infrastructure initiative in a consortium of 33 partners, including research communities, national data and high performance computing (HPC) centres, technology providers, and funding agencies from 14 countries. EUDAT aims to build a sustainable cross-disciplinary and cross-national data infrastructure that provides a set of shared services for accessing and preserving research data.

The EUDAT *Collaborative Data Infrastructure* (CDI) is a defined data model and a set of technical standards and policies adopted by European research data centres and community data repositories to create a single European e-infrastructure of interoperable data services. The EUDAT CDI is realised through ongoing collaboration between *service providers* and *research communities* working as part of a common framework for developing and operating an interoperable layer of common data services. The scope of the CDI covers data management functions and policies for upload and retrieval, identification and description, movement, replication and data integrity. EUDAT's vision is to enable European researchers and practitioners from any research discipline to preserve, find, access, and process data in a trusted environment. The CDI is conceived as a network of collaborating, cooperating centres, combining the richness of numerous community-specific data repositories with the permanence and persistence of some of Europe's largest scientific data centres. At the heart of the CDI is a network of distributed storage systems hosted at the major scientific data centres. Between them, these centres manage more than 100 PB of high performance, online disk in support of European research, plus a greater amount of near-line tape storage. EUDAT's strength lies in the connections between these centres, the resilience resulting from the geographically distributed network, and its ability to store research data right alongside some of the most powerful supercomputers in Europe.

Currently, EUDAT is working with more than 30 research communities covering a wide range of scientific disciplines and has built a suite of integrated services (Table 18 below) to assist them in resolving their technical and scientific challenges.

Covering both access and deposit, from informal data sharing to long-term archiving, and addressing identification, discoverability and computability of both long-tail and big data, EUDAT services aim to address the full lifecycle of research data.

TABLE 18: THE EUDAT SERVICE CATALOGUE

Service	Function	Status	Individual Researcher	RI/ Community Manager	Service Provider
Data Discovery					
B2FIND	Multi-disciplinary joint MD catalogue	Active	X	X	
Metadata Catalogue	MD extraction, MD store, index	Under develop.		X	X
Data Hosting, Registration & Management & Sharing					
B2DROP	Cloud storage, sync & exchange	Active	X	X	X

Service	Function	Status	Individual Researcher	RI/ Community Manager	Service Provider
B2SAFE	Policy-driven data management	Active		X	X
B2SHARE	Repository for sharable digital objects	Active	X	X	X
B2HANDLE	Policy-based prefix & PID management	Active		X	X
Data Type Registry		Under develop.		X	
Data Access, Interface & Movement					
B2ACCESS	Federated multi-protocol IAM	Active	X	X	X
Generic API	Common data interface service	Under develop.	X	X	X
B2STAGE	Data staging service CDI → ext.	Active	X	X	X
Subscription	Data transfer subscription	Under develop.		X	X
Consultancy					
Training	on services & data management	Active	X	X	X
Consultancy	on licensing, certification, data privacy, data system design	Active	X	X	X
Helpdesk	Support and enabling	Active	X	X	X
Operations					
Service Hosting	PaaS, IaaS, SaaS	Under develop.		X	X
Monitoring	Availability & reliability monitoring	Active		X	X
Accounting	Storage & Data Usage Reporting	Under develop.		X	X
SLC Management	Service Portfolio & Catalogue	Active	X	X	X
Coordination	Project Implementation, Service & Resource Provisioning	Active	(X)	X	X
Site Registry	Site, Service & Service Groups	Active	(X)	X	X

These services have been developed together with research communities coming mostly from the environmental sciences (EPOS, ICOS, EISCAT, ENES, LTER, DRIHM), life sciences (ELIXIR, VPH, BBMRI, ECRIN, DIXA), and social sciences and humanities (CLARIN, CESSDA, DARIAH), In October 2015, EUDAT issued a public call for data pilot projects and received 24 applications including 9 from Earth and Environmental Sciences, Energy and Environment disciplines, 6 from the Biomedical and Life Sciences, 5 from the Social Sciences and Humanities, and 4 from Physical Sciences and Engineering. Altogether these pilots represent a potential user base of 40,000 researchers.[148]

EUDAT distinguishes three main types if users/customers of its services and infrastructure:

- Individual researchers: Those wishing to share data with colleagues or collaborators, or those wishing to discover and re-use data as part of their on-going research.
  - These users are anybody – researchers (from academia and industry), citizen scientists, policy makers, and members of the public – anyone wanting to share or re-use European research data in simple, powerful ways.
  - As a user your main responsibility is to adhere to the terms and conditions of the B2 services provided by the EUDAT consortium
- Organized research communities: Those concerned with the management of their research e-infrastructure and/or community-specific data repositories who wish to join their repositories formally with the CDI network or deploy EUDAT services on top
  - These research communities are organized research groups (e.g., EC projects), research Infrastructures (e.g., ESFRI) or universities and libraries – anyone interested

[148] https://www.eudat.eu/eudat-communities-pilots

in archiving, replicating, processing and cataloguing data on behalf of the research community they support.
    - They either use EUDAT services a) as they are, according to the service terms and conditions, b) through an agreement with a specific service provider, or c) by joining the CDI as a node/service provider.
- CDI Service providers
    - Service Providers wish to use and/or deploy CDI services to support or augment their existing role and service portfolio – to provide long term preservation of important digital assets, offer wider accessibility, intelligent caching of data near compute, data integrity checking and so on.

Positioning EUDAT

In defining the EUDAT CDI's position with respect to other e-infrastructure initiatives and organisations, EUDAT regards any and all e-infrastructures (including, though not limited to, PRACE, EGI, HelixNebula, OpenAIRE) as *organisational end-users* of EUDAT's services. The CDI Gateway API defines a clear contract with external end-users and consequently a set of stable targets for computational jobs (scripts, programs or workflows) running on external infrastructure.

The key value that EUDAT's implementation of the CDI brings to any external user is a *well-defined API to EUDAT services* and *coherent service offerings* across all EUDAT partner sites. These common, coherent service interfaces create the line of demarcation between the EUDAT CDI and the other e-Infrastructures – the boundary of the domain of registered data. Other infrastructures then have clear ways to interact with the EUDAT CDI. Across the *network* they can:

- retrieve metadata records by PID (e.g., HTC workflows, HPC programs, publication repositories & catalogues);
- retrieve open access data by PID (e.g., HTC workflows, HPC programs, publication repositories & catalogues);
- subscribe to metadata feeds using OAI-PMH (e.g., publication catalogues);
- where authorised: create (upload) data & metadata and receive a registered PID (e.g., HTC workflows, HPC programs & scripts);
- where authorised: update or delete data and/or metadata by PID (e.g., HTC workflows, HPC scripts).

This model positions the EUDAT CDI as the home for persistent, shared, re-used research data.

EUDAT is about preserving research data for reuse, and an aspect of making digital data reusable lies in providing the capabilities for efficient computation on them. EUDAT2020 enables data analytics by staging data to dedicated analysis systems – leveraging the computing capacity made available via EGI and PRACE. EUDAT has issued two joint public calls in 2015 with PRACE allowing PRACE users which have been granted PRACE computing resources to store the data resulting from simulations into EUDAT. It is also working with EGI to strengthen interoperability between the two infrastructures with a view to connect data stored in the EUDAT Collaborative Data Infrastructure to high throughput and cloud computing resources provided by EGI. EUDAT develops solutions for data coupled computing, including big data frameworks and workflow systems for initiating computing tasks on datasets located in the EUDAT infrastructure. EUDAT B2STAGE library allows to stage data to HPC computing environments and is being developed further to add support for Hadoop and Spark big data systems. EUDAT also offers a hosting environment for the deployment and provision of data analytics services directly at the data centres – building on the Service Hosting Framework successfully trialled in the first EUDAT project to provide a flexible virtual computing environment at participating data centres, a highly-configurable cluster computing platform sited right alongside the data archives.

Publications, data and OpenAIRE

OpenAIRE[149] enables researchers to deposit research publications and data into Open Access repositories and provides support to researchers at the national, institutional and local level to guide them on how to publish in Open Access (OA) and how to manage the long tail of science data within the institution environment. This complements national initiatives in several European countries. If researchers have no access to an institutional, national or a subject repository, Zenodo[150], hosted by CERN, enables them to deposit their articles, research data and software. Zenodo exposes its contents to OpenAIRE and offers a range of access policies helping researchers to comply with the Open Access demands from the EC and the ERC (European Research Council). It now uses CERIF for its metadata. Zenodo has also been extended with important features that improve data sharing, such as the creation of persistent identifiers for articles, research data and software. OpenAIRE has recently moved from a DC-like metadata catalogue to CERIF in OpenAIREplus.

Open Science Commons of EGI

EGI developed its 'Open Science Commons' vision[151] inspired by the emerging open access policy in the European Research Area. The goal of open access is to ensure that research results are made freely available to end users and that they are reusable. Research results and resources thus become a shared community resource (i.e., a commons). In order for this to happen, researchers need to change their own behaviours and they need to be supported with services that simplify the sharing of research results, their discovery and reuse. In the EGI-Engage project EGI is developing the concept of a federated open research data platform, an innovative solution enabling to publish data, link to open access repositories, and offering easy integration into processing capabilities (e.g., EGI Federated Cloud). Furthermore, the federated cloud infrastructure, including existing publicly funded institutional cloud and expanding to commercial clouds, will evolve to offer IaaS, PaaS and SaaS for specific communities, the long-tail of research and the industrial/SME sector. In collaboration with other e-infrastructures, services will be tailored to meet the needs of the long tail of research and their evolution will be driven by the requirements of the RIs on the ESFRI roadmap that participate in the EGI Engage project through Competence Centres.

Research Data Alliance

Together, and with many other organisations the pan-European e-Infrastructure initiatives are contributing to international cooperation in addressing issues around large-scale data infrastructures through the recently formed international Research Data Alliance (RDA)[152]. Launched as a community-driven organization in 2013 by the European Commission, the United States National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation, the Research Data Alliance (RDA) has the goal of building the social and technical infrastructure to enable open sharing of data.

With close to 4,000 members from 110 countries (April 2016), RDA provides a neutral space where its members can come together through focused global Working and Interest Groups to develop and adopt infrastructure that promotes data-sharing and data-driven research, and accelerate the growth of a cohesive data community that integrates contributors across domain, research, national, geographical and generational boundaries.

In Europe, the work of the RDA has been supported by several projects funded under FP7 and H2020.

[149] https://www.openaire.eu/
[150] https://zenodo.org/
[151] https://www.opensciencecommons.org/
[152] https://rd-alliance.org/

3.11.4 A longer term horizon

The recent revised ESFRI Roadmap 2016 [ESFRI 2016], highlights the notion of a *European e-infrastructure Commons,* referring to the framework for an easy and cost-effective shared use of distributed electronic resources for research and innovation across Europe and beyond. The concept is outlined by the e-Infrastructure Reflection Group (e-IRG) based on the identification of the need for a more coherent e-infrastructure landscape in Europe. According to the e-IRG report[153],

"*An essential feature of the Commons is the provisioning of a clearly defined, comprehensive, interoperable and sustained set of services, provisioned by several e-infrastructure providers, both public and commercial, to fulfil specific needs of the users. This set should be constantly evolving to adapt to changing user needs, complete in the sense that the needs of all relevant user communities are served and minimal in the sense that all services are explicitly motivated by user needs and that any overlap of services are thoroughly motivated. The Commons has three distinct elements:*

- A platform for coordination of the services building the Commons, with a central role for European research, innovation and research infrastructure communities.
- Provisioning of sustainable and interoperable e-infra structure services within the Commons, promoting a flexible and open approach where user communities are empowered to select the services that fulfill their requirements.
- Implementation of innovation projects providing the constant evolution of e-infrastructures needed to meet the rapidly evolving needs of user communities."

In summary, the ultimate vision of the Commons is to reach integration and interoperability in the area of e-infrastructure services, within and between member states, and on the European level and globally. This e-infrastructure Commons is also a solid basis for building the *European Open Science Cloud* as introduced in the description of the Digital Single Market [COM(2015) 192 final], [SWD(2015) 100 final], already containing most of the ingredients needed for an integrated European platform for Open Science [ESFRI 2016].

To support this vision, it would request a long-term agenda for supporting a coherent, innovative and strategic European e-infrastructure policy making and the development of convergent and sustainable e-infrastructure services. Today (April 2106) the EC announces the European Cloud Initiative[154] - €6.7billion of public and private investment in European Open Science Cloud (2016), opening up by default all scientific data (2017), flagship initiative on quantum technology (2018), development and deployment of European high performance computing, data storage and network infrastructure (2020), including by acquiring two prototype next-generation supercomputers of which one would rank among the top three in the world, establishing a European big data centre, and upgrading the backbone network for research and innovation (GEANT).

3.11.5 Relationships with requirements and use cases

ENVRIplus has already been collaborating with these pan-European e-Infrastructures, such as EGI and EUDAT. EUDAT services are chosen (by some of Research Infrastructures) for data management. Other RIs will benefit from feedback on their initial experiences.

In ENVRIplus WP9, EGI will provide computing and storage resources for deploying services developed by ENVRIplus development WPs. The task begins with identifying a number of community use cases, and the feasibility of deployments of the use cases are evaluated by e-

[153] http://e-irg.eu/documents/10920/11274/e-irg-white-paper-2013-final.pdf
[154] http://europa.eu/rapid/press-release_IP-16-1408_en.htm?locale=en

Infrastructure experts. 5-6 use cases are selected which will have resources and technical supports from EGI for deployments.

3.11.6 Issues and implications

Interoperable access to these e-Infrastructures remains as a challenging issue. In this sense, ENVRIplus is in good position to provide real use cases/requirements to influence the future implementations of these e-Infrastructures.

Further discussion of the provision of computational, storage, network and software technologies can be found in Section 4.2.14. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

4 Assessment of achievements, gaps and impact

This section assesses the achievements in the two parts, requirements gathering and technology review, and their relationships. It also assesses the work's implications for the planned work and for additional actions. Finally, it categorises the outcomes in terms of their short-term and longer-term implications.

4.1 Assessment of requirements gathering

The requirements gathering campaign, built on the understanding developed during the preceding ENVRI project, and on the intensive discussions that shaped the ENVRIplus bid. Its primary purpose was to sufficiently understand the *combined* requirements of the RIs, many of which are new since ENVRI, and all of which have developed substantially, to be sure that the work undertaken in Theme 2 is the best possible match to the current and anticipated requirements.

There were the following subsidiary purposes:

1. To stimulate dialogue and effective communication within ENVRIplus, particularly between experts in RIs with ICT experts.
2. To initiate a resource for recording and analysing requirements that will be sustained and useful throughout the project and beyond.
3. To help with awareness raising and training by identifying where emphasis should be placed at this time.

Undertaking a requirements gathering process near the start of a project is necessary if it is to guide subsequent investment. However, it then meets an extra difficulty as many partners and individuals are new and are orienting themselves and building their own communication and decision-making networks. This was experienced and led to some delays. It also meant that some of the outcomes are not as authoritative and based on as extensive analysis as we might have hoped. Therefore, they should be checked before significant R&D investments are undertaken. Nevertheless, they are a significant and valuable achievement that meets the primary goal, and that makes a substantial contribution to all three subsidiary purposes.

The gathered requirements and the requirement gathering process are complementary to the use-case activity[55] that is also underway in ENVRIplus. The agile co-design and co-development undertaken for each use case will deepen and refine both requirements and technology review for their focused areas. The use cases will also develop and extend the communication paths, helping build a stronger asset powering collaboration—the first subsidiary goal.

The contributions to the three subsidiary purposes will be reviewed first. We then present an analysis of how well the primary goal was met.

Fostering communication: The intensive discussions between go-betweens and RI representatives formed many new interpersonal bridges. This was frequently a new connection and they have a good potential for sustained value throughout the ENVRIplus project and beyond. Although very little staff time was formally allocated to this activity, in many of the RIs, that initial communication frequently triggered further communication within the RI and among those who will undertake Theme 2 tasks. In most cases, the topic leader will also be leading the subsequent related task and they used this as an opportunity to start communications within their planned team.

Foundation for requirements refinement: This report has been derived from the wiki pages where the primary information about requirements were gathered[155]. This initial collection is already an asset for those planning implementation tasks and for those wanting to know how other RIs are addressing data challenges. It will provide an easily searched and easily updated framework as the understanding of requirements progresses. This should prove valuable even beyond the end of ENVRIplus provided the material is kept up-to-date.

Awareness raising and training: The requirements gathering, particularly the investigation of general issues and the analysis of community support needs, has identified areas where these needs are evident and relatively urgent. The differences between RIs' responses reveal more opportunities for developing these aspects of ENVRIplus.

This sets the scene for the analysis of the primary goal **validation of ENVRIplus's data-oriented ICT R&D**. In general terms every one of the planned lines of development were endorsed by the requirements gathering and no major omissions have been identified. However, a more detailed review does reveal some significant issues, which will be introduced below and collated in Sections 5.1 and 5.2. These will be pursued by first considering the overall process in conjunction with the general requirements gathering (Section 4.1.1 below), and then considered under the topic headings: *Identification and Citation*, *Curation*, *Cataloguing*, *Processing*, *Provenance*, *Optimisation* and *Community support* (Sections 4.1.2 - 4.1.8 below). These headings correspond to areas where significant effort will be invested in ENVRIplus. They are also informed by the reference model[29] developed in ENVRI and being further developed in ENVRIplus.

4.1.1 Process and general requirements

The detailed process was described in Section 2.1 on page 19. It ran as planned but it is worth reviewing its progress in terms of Table 4 on page 23. There it will be seen that there is substantial variability by RI and similar variability by topic. For every RI, a significant effort was made to develop communication and obtain information about requirements for all relevant topics. In some cases, a particularly strong relationship or existing knowledge enabled complete coverage. In some the RI was mature, in the sense that the RI or those involved in the work had been active in the particular domain for a significant number of years; the marine RIs that are already sharing data, such as Euro-ARGO and SeaDataNet are good examples. Such maturity leads to an appreciation of the complexities and significance of various requirements. In other cases, the RI concerned was in a consortium of interacting, often global, related communities that share data and hence appreciate many of the issues; EPOS is one example. For such RIs, it was possible to gather good input on virtually every topic. For all of the RIs, contact was made and information was gathered for at least the general requirements. In some cases, an RI deemed their interests were already covered by another RI known to be similar with which they worked closely.

The variation between topics is also a manifestation of maturity variation but this time combined with variations in the parts of the data lifecycle in which each RI is involved, as shown in Table 5. The topics such as *Identification and Citation*, *Cataloguing* and *Processing*, are encountered at

[155] https://wiki.envri.eu/display/EC/ENVRI+RI+Requirements

the early stages of developing an RI's work and at the early stages of the data lifecycle. Whereas, the value of *Curation* and *Provenance* become much more apparent after running a data gathering and sharing campaign for long periods or from being involved in the later stages of the data lifecycle. *Optimisation* is an extreme example of this effect; only when production and diverse users are demanding more resources than an RI can afford does optimisation become a priority; before that the focus is on delivering the breadth of functionality users require and gaining adoption. As we shall explain below, Section 4.1.7, these can be met by addressing different aspects of optimisation.

The outcome of gathering general requirements is analysed in Section 2.2.19 and summarised in a series of tables. These provide a summary of the information uncovered via each group of general questions; however, readers are referred to the relevant part of the wiki for all details[156]. The overall conclusion wold be that there are many opportunities for benefit from sharing ideas, methods and technologies between RIs, that there is much potential for using their data in combination and that there is a general need for awareness raising and training. However, these high-level consistencies have to be treated with great care; there are many lower level details where differences are significant. Future work will need to tease out which of those differences are fundamentally important and which are coincidental results from the path the participants have taken to date. Fundamental differences need recognition and support with well-developed methods for linking across them founded on scientific insights. The unforeseen differences may in time be overcome by incremental alignment; however, great care must be taken to avoid unnecessary disruption to working practices and functioning systems. This will require deeper investigation, e.g., through appropriate use cases and agile investigations[157].

4.1.2 Identification and citation requirements assessment

The *Identification and Citation* requirements are summarised in Section 2.3.1, which validates the need for this provision in ENVRIplus. However, the RIs showed significant diversity in their data-identification and data-citation practices and many were not aware of their importance in supporting data use. *Data Identification and Citation* are, however, key to reproducibility and quality in data-driven science and very often vital in persuading data creators of the value of contributing their data, data users of the need to recognise that contribution and funders to continue to support data gathering and curation.

The next steps will include:

1. ENVRIplus will consider a programme of awareness raising and practical training to alert those RIs that would benefit, and to raise the skills of practitioners in any RI, of the relevance of *Data Citation and Identification* issues and some of the available technologies that will help with solutions and rapid adoption of good practices. The EU EDISON project[158] has already worked on this for the cluster project CORBEL for the bio-medical RIs.
2. The conceptual and technical issues in *Data Citation and Identification* are strongly linked with best practice in *Curation* and practically linked with *Cataloguing* and *Provenance*. These will be considered together in order to provide consistent advice and solutions to RIs.
3. A key issue is adoption of appropriate steps in working practices. Where these are exploratory or innovative the citation of underpinning data may be crucial to others verifying the validity of the approach and to later packaging for repeated application. Once a working practice is established, it should be formalised, e.g., as a workflow, and packaged, e.g., through good user interfaces, so that as much of the underpinning

[156] https://wiki.envri.eu/display/EC/ENVRI+RI+Requirements
[157] https://wiki.envri.eu/display/EC/Use+Cases
[158] http://www.edison-project.eu

record keeping: e.g., *Citation*, *Cataloguing* and *Provenance* is automated. This has two positive effects, it enables the practitioners to focus on domain-specific issues without distracting record keeping chores, and it promotes a consistent solution to be incrementally refined. For these things to happen there have to be good technologies, services and tools supporting each part of these processes, e.g., data citations being automatically and correctly generated as suggested by Buneman *et al*. [Buneman 2016]. Similarly, constructing immediate payoffs for practitioners using citation, as suggested by Myers *et al.* [Myers 2015], will increase the chances of researchers engaging with identification at an earlier stage.

4. Many researchers today access and therefore consider citing individual files. This poses problems if the identified files may be changed, the issue of *fixity*. Many research results and outputs depend on very large numbers of files and simply enumerating them does not yield a comprehensible citation. Many derivatives depend on (computationally) selected parts of the input file(s). Many accesses to data are via time varying collections, e.g., catalogues or services, that may yield different results or contents on different occasions—generically referred to as *databases*. Some results will deal with continuous streaming data. Often citations should couple together the data sources, the queries that selected the data, the times at which those queries were applied, the workflows that processed these inputs and parameters or steering actions provided by the users (often during the application of the scientific method) that potentially influenced the result. All of these pose more sophisticated demands on the *Data Identification and Citation* systems. At present they should at least be considered during the awareness raising proposed above. In due course, those advanced aspects that would prove useful to one or more of the RI communities should be further analysed and supported. This is revisited in the technology review Section 3.2 and in Section 4.2.5.

4.1.3 Curation requirements assessment

The *Curation* requirements validate the need for ENVRIplus developing curation solutions but do not converge on particular requirements; see Section 2.3.2, which analyses the information supplied by seven RIs who responded to this topic; see the wiki page for details[159]. In the planned work of ENVRIplus this work is already conceptually and practically interrelated with *Cataloguing* and *Provenance* in WP8. As remarked above, it should also strongly couple with the work on *Data Identification and Citation*. Consequently, many of the issues that emerge are similar to those identified above. However, some further issues arise. These are enumerated below:

1. The appreciation of the needs for *Curation* is varied and often limited, one manifestation of this is the universal absence of data management plans[160]. Consequently, this topic again poses a requirement for an ENVRIplus programme of awareness raising and training. If that is conducted collaboratively then it will also help develop cross-disciplinary alliances that will benefit scientific outcomes, management decisions and long-term cost-benefit trade-offs.
2. The need for intellectual as well as ICT interworking between these closely related topics: *Identification and Citation*, *Curation*, *Cataloguing* and *Provenance* is already recognised. Their integration will need to be well supported by tools, services and processing workflows, used to accomplish the scientific methods and the *Curation* procedures. However, there was negligible awareness of the need to preserve software and the contextual information necessary to re-run it with identical effects. The need for this combination for reproducibility is identified by Belhajjame *et al*. with

[159] https://wiki.envri.eu/display/EC/Curation+requirements

[160] These may be latent in policy and management documents of each RI. Drawing them together into a formal DMP will take time. It might benefit from being collaborative, and from training such as that offered by the DCC, http://www.dcc.ac.uk/.

implementations automatically capturing the context and synthesising virtual environments [Belhajjame 2015].

3. As above, it is vital to support the day-to-day working practices and the innovation steps that occur in the context of *Curation* with appropriate automation and tools. This is critical both to make good use of the time and effort of those performing *Curation*, and to support innovators introducing new scientific methods with consequential *Curation* needs.
4. The challenge of handling all forms of data described in Section 4.1.2 for *Identification and Citation* is compounded with the need to properly capture diverse forms of software and a wide variety of, often distributed, computational contexts in order to fully support reproducibility.
5. Curation needs to address preservation and sustainability; carefully preserving key information to underwrite the quality and reproducibility of science requires that the information remains accessible for a sufficient time. This is not just the technical challenge of ensuring that the bits remain stored, interpretable and accessible. It is also the socio-political challenge of ensuring longevity of the information as communities' and funders' priorities vary. This is a significant step beyond archiving, which is addressed in EUDAT with the B2SAFE service[161].
6. One aspect of the approach to sustainable archiving is to form federations with others undertaking data curation, as suggested by OAIS[162]. Federation arrangements are also usually necessary in order that the many curated sources of data environmental scientists need to use are made conveniently accessible. Such *data-intensive federations* (DIF) underpin many forms of multi-disciplinary collaboration and supporting them well is a key step in achieving success. As each independently run data source may have its own priorities and usage policies, often imposed and modified by its funders, it is essential to set up and sustain an appropriate DIF for each community of users. Many of the RIs deliver such federations, *today without a common framework to help them*, and many of the ENVRIplus partners are members of multiple federations.

These issues are revisited in Sections 3.3 and 4.2. They lead to recommendations in Sections 5.1 and 5.2.

4.1.4 Cataloguing requirements assessment

As for the preceding topics, the analysis of requirements (see Section 2.3.3) validated the need for ENVRIplus help with *Cataloguing* solutions but current practice and understanding of precise needs was once again very varied. There are a wide variety of items that could be catalogued, from instruments and deployments at the data acquisition stage, right through every step of data processing and handling, including the people and systems responsible, up to the final data products and publications made available for others to use. Most responding RIs pick a small subset of interest, but it is possible that a whole network of artefacts need cataloguing to facilitate *Provenance*, and many of these would greatly help external and new users find and understand the research material they need. There is a similar variation in the kinds of information, metadata, provided about catalogue entries. Only EPOS has a systematic approach by using CERIF, though many have commonalities developing because of the INSPIRE directive [EU Parliament 2007]. So again we will consider a few implications:

1. A programme of awareness raising, training and boundary crossing events is urgently needed to help develop greater appreciation of the value of catalogues as an aid to research[163], lead to more precise requirements, initiate alliances and accelerate

[161] http://www.eudat.eu/b2safe
[162] http://wiki.dpconline.org/index.php?title=6-3_An_OAIS_Federation_Employing_a_Common_Catalog
[163] For example, in the IVOA context (see Section 1) machine-learning (ML) algorithms often run on the catalogues alone without recourse to the primary data. When a ML-based measurement or recognition method has become established its

adoption. As always, adoption will only happen if there is an evident benefit to researchers.

2. A critical factor that emerged in general requirements discussions was the need to easily access data. This clearly depends on good query systems that search the relevant catalogues and couple well with data handling and provenance recording. The query system is closely coupled with catalogue design and provision, but it also needs integration with other parts of the system. Euro-ARGO identified a particular version of data access—being able to specify a requirement for a repeating data feed.
3. Catalogues are a key element in providing convenient use of federations of resources. It is probably necessary to have a high-level catalogue that identifies members of the federation and the forms of interaction, preferably machine-to-machine, they support. Initially users may navigate this maze and handle each federation partner differently, but providing a coherent view and a single point of contact has huge productivity gains. It is a moot point whether this requires an integrated catalogue or query systems that delegate sub-queries appropriately. This is another example where effective automation can greatly improve the productivity of all the RI's practitioners; those that support the systems internally and maintain quality services, and those who use the products for research and decision making. It is anticipated that federations will grow incrementally and that the automation will advance to meet their growing complexity and to deliver a holistic and coherent research environment where the users enjoy enhanced productivity. This will depend on catalogues holding the information needed for that automation as well the information needed for RI management and end-user research.
4. Once again there may be some merit in making the advantages of catalogues evident in the short-term, e.g., by coupling catalogue use with operations that user want to perform, such as: having selected data via a catalogue, moving it or applying a method to each referenced item. Similarly, allowing the users some free-form additions and annotations to catalogue entries that help them pursue their own goals may be helpful.

Many of these issues are revisited in the context of the *Cataloguing* technology review (Section 3.4) and their implications are considered in Section 4.1.4.

4.1.5 Processing requirements assessment

Once again, the analysis of requirements (see Section 2.3.4) validated the need for ENVRIplus to help with *Processing* solutions. The wide scope of potential contexts in which processing could be applied: from quality assurance close to data acquisition to transformations for result presentation (and every research, data-management or curation step in between) makes this a complex factor to consider. User engagement with this topic also varies validly between two extremes: those who use a pre-packaged algorithm in a service almost unknowingly as part of a well-formalised, encapsulated, established method they use, to those who are engaged in creating and evaluating new algorithms for innovative ways of combining and interpreting data. Clearly, both continua are valid and any point in each continuum needs the best achievable support for the context and viewpoint. With such diversity it is clear that a one-size-fits-all approach is infeasible. This conclusion is further reinforced by the need to exploit the appropriate computational platforms (hardware architectures, middleware frameworks and provision business models) to match the properties of the computation, and the priorities of the users given their available resources. If such matching is not considered it is unlikely that all of the developing research practices will be sustainable in an affordable way. For example, too much energy may be used or the call on expert help to map to new platforms may prove unaffordable. Such issues hardly rise to the fore in the early stages of an RI or a project. So again,

results are often included when subsequent catalogues are built. This greatly accelerates the access to such measures and makes new science feasible.

we note forces that will cause the understanding and nature of requirements to evolve with time. This leads to the following follow-up observations:

1. A programme of awareness raising and training events will be tuned to different viewpoints of participants and also link up with the relevant target technologies and provider models. It is more likely that this will benefit the systems developers who are setting up processing services and the innovators who are creating new ways of using them. Bringing them together may trigger significantly beneficial mutual understanding and alliances.
2. The packaging of computations and the progressive refinement of scientific methods are key to productivity and to the quality of scientific conclusions. Consequently, as far as possible processing should be defined and accessed by high-level mechanisms. This allows a focus on the scientific domain issues and it leaves freedom for optimised mappings to multiple computational platforms. This protects scientific intellectual investment, as it then remains applicable as the computational platforms change. This will happen as their nature is driven by the much larger entertainment, media, leisure and business sectors. The higher-level models and notations for describing and organising processing also facilitate optimisation and automation of chores that otherwise will distract researchers and their supporters.
3. Providing support for innovation in this context is critical. Without innovation the science will not advance and will not successfully address today's societal challenges. It requires support for software development, testing, refinement, validation and deployment conducted by multi-site teams engaging a wide variety of viewpoints, skills and knowledge. For the complex data-intensive federations the environmental and Earth sciences are dealing with, this involves new intellectual and technological territory. Alliances involving multiple RIs and external cognate groups such as EUDAT, PRACE and EGI, may be the best way of gathering sufficient resources and building the required momentum.

Further consideration of these issues may be found in the *Processing* technology review, Section 3.5 and as suggested further actions in Section 4.1.5.

4.1.6 Provenance requirements assessment

At present, the need for and benefits of *Provenance* provision are only recognised by some RIs, Section 2.3.5. In abstract, we are sure that most scientists appreciate the value of provenance, but they tend to think of it as a painful chore they have to complete when they submit their final, selectively chosen data to curation. They often only do this when their funders or publishers demand it. That culture is inappropriate. For many RIs they are in the business of collecting and curating primary data and commonly required derivatives. Clearly, they want to accurately record the provenance of those data, as a foundation for subsequent use and to achieve accountable credit. For environmental and Earth scientists use of provenance throughout a research programme can have significant benefits. During method development it provides ready access to key diagnostic and performance data, and greatly reduces the effort required to organise exactly repeated re-runs; a frequent chore during development. As they move to method validation they have the key evidence to hand for others to review. When they declare a success and move to production, the provenance data informs the systems engineers about what is required and can be exploited by the optimisation system. Once results are produced using the new method these development-time provenance records underpin the provenance information collected during the production campaign. Of course, all of this depends on:

1. Users having control over which provenance data is generated and which is preserved;
2. The provenance system being fully automated so that no niggling chores intrude; and
3. Tools that exploit the provenance data and support all the innovation and production steps.

The RIs survey reported very different stages of adoption, and when there was adoption it did not use the same solutions or standards—this was almost always related to data acquisition rather than the use of data for research. The change in culture among researchers may be brought about by ENVRIplus through a programme of awareness raising and a well-integrated compendium of tools. The latter may be more feasible if the development of the active provenance framework is amortised over a consortium of RIs. This leads to similar observations to those given above:

1. An awareness arising and training programme should be considered to stimulate thinking about and use of provenance recording and exploitation.
2. Automation is essential for all aspects of provenance handling to avoid sapping productivity. But this needs to be under the control of the RIs and practitioners concerned.
3. Effective tools that show the value of provenance, e.g., those suggested be Spinuso *et al.* [Spinuso 2016], are key to changing the culture and encouraging early engagement with the quality of provenance [Myers 2015].

These issues are further considered when Provenance technology is reviewed, Section 3.5, and lead to suggestions in Section 4.1.6.

4.1.7 Optimisation requirements assessment

At present the *identified* set of optimisation requirements, analysed in Section 2.3.6, is relatively sparse. However, there is anticipated to be a demand, which will become manifest when production of research results ramps up, as RIs deliver continuous services and data feeds, or as the numbers and diversity of users grow. Experience shows that as data-handling organisations transition from pioneering to operations, many different reasons for worrying about optimisation emerge. These are addressed by a wide variety of techniques, so that investment in optimisation is usually best left until the following kinds of question can be answered:

1. What precisely does the RI or its user community want to be optimised?
2. What trade-offs would they find acceptable to achieve that optimisation?
3. How can this be formalised as a measurable cost function encapsulating the answers to the first two questions?

Very often there are significantly different answers from different members of a community. The RI's management may need to decide on compromises and priorities. For such reasons,

1. Awareness raising and training campaigns may be appropriate, though they may not be as urgent as they are for some of the other topics.
2. Optimisation needs to look beyond individuals and single organisations. When looking at overall costs or energy consumption in a group of RIs or the e-Infrastructures they use, tactics may consider the behaviour of a data-intensive federation. For example, when data is used from remote sites, or is prepared for a particular class of uses, the use of caching may save transport and re-preparation costs, and accelerate the delivery of results. However, the original provider organisation needs to have accountable evidence that their data is being used indirectly, and the caching organisation needs its compute and storage costs amortised over the wider community.
3. Thinking about optimisation tends to focus on technology and operational costs. However, the most valuable asset of an RI is almost certainly its community of practitioners:
    a. those who exploit the facilities, data, and services to pursue their research goals, and
    b. those who create, improve and operate the facilities the first subgroup uses.

Enabling these critical parts of the community to be as productive as possible can be viewed as an optimisation challenge. In the context of ENVRIplus, it is particularly important to consider empowering community members to collaborate effectively across boundaries:

4. How to provide automation that removes as many chores as possible from their routine work, while leaving them both the ability to understand and investigate what is going on and to apply controls where they are necessary.
5. How to provide tools, development environments and VREs that easy to use, particularly during innovation, without removing critical degrees of freedom from the innovation options.

These topics are revisited when the technology options for optimisation are investigated (see Section 3.7).

4.1.8 Community support requirements assessment

The requirements for community support are summarised and analysed in Section 2.3.7 from page 57 onwards. There you will find a shopping list of virtually all of the facilities for communication, information sharing, organisation and policy implementation that a distributed community of collaborating researchers and their support teams might expect – they normally expect those facilities to be well integrated and easily accessed wherever they are from a wide range of devices. However, care should be taken to consider the full spectrum of end users. A few may be at the forefront of technological innovations but the majority may be using very traditional methods, because they work for them. Investment is only worthwhile if it is adopted and benefits the greater majority of such communities, taking into account their actual preferences.

There may be two key elements missing in the context of ENVRIplus, which focuses on achieving the best handling and use of environmental data:

1. **Workspaces** that can be accessed from anywhere and are automatically managed, in which individuals or groups can store and organise the data concerned with their work in progress: e.g., test data sets, sample result sets, intermediate data sets, results pending validation, results pending publication. Since environmental researchers have to work in different places, such as in field sites, in different laboratories and institutions, they need to control these logical spaces, which may be distributed for optimisation or reliability reasons. These are predominantly used to support routine work but can also be used for innovation. This includes intelligent sensors requiring access to a variety of logical spaces for their operations.
2. **Development environments** that can be accessed from most workstations and laptops, and that facilitates collaborative innovation and refinement of the scientific methods and the data handling. Sharing among a distributed community, testing, management of versions and releases and deployment aids would be expected.

In ENVRIplus collaboration between various roles including citizen scientists, (across intellectual, organisational and academic-cultural boundaries) is a widespread requirement. We can illustrate this with the following roles. There are occasional heroes who span several of these roles, but predominantly we have to pool different skills from different roles to make breakthroughs or even to do the daily business.

1. **Domain specialists** who may be more or less computer literate, but who will develop and use new patterns of data and RI facilities use, at least characterised by parameter sets, and written procedures for themselves and others to follow. Such methods may be formalised as workflows, scripts and programs using frameworks and packages. Because scientific progress depends on sufficient repetition, perhaps 80% to 90% of their work is routine, repeating previous methods with refinement. But progress often depends on

their insights as they recognise new potential in the available data or new questions their domain should address.

2. **Data scientists** who are adept at and develop new statistical and machine learning strategies and analytic procedures for cleaning and preparing data, and for extracting derivatives or information from data.
3. **Data-Intensive engineers**, who set up and maintain a great deal of plumbing on top of standard e-Infrastructures to enable all of the data handling and storage required and the use of the data either through access interfaces for external processing or through local resources with which to process the data.
4. **Virtual Research Environment (VRE) designers and builders**, who shape the science gateways and improve the APIs and interactive services they offer.
5. **System administrators** who oversee the operation of the platforms and software, recognise impending resource shortages, plan and conduct procurement and installation of resource increments and keep the strength of security sufficiently high.
6. **Computational modellers and numerical analysts** who develop simulation systems and mechanisms for exploring their results.

To a lesser or greater extent virtually every RI will depend on such a mix of roles and viewpoints. Community support needs to recognise and engage with these multiple viewpoints as well as help them to work together. This is particularly challenging in the distributed environments and federated organisations underpinning many RIs. At least training and help desk organisation will need to take these factors into account. Productivity will come from each category being well supported. Significant breakthroughs will depend on the pooling of ideas and effort across category boundaries.

4.1.9 New requirements identified

The plan for work in ENVRIplus developed by the RIs and incorporated in the DoW has been validated and endorsed by the requirements gathering as is shown in Table 16. There are a few additional and not planned aspects of data management and user support that appear in the conversations and underpin some of the general issues. These have aspects of improving usability to improve the experience and productivity of users and the teams who support them. In part, they are better packaging of existing or planned facilities and in part they are intended to deliver immediate benefits to keep communities engaged and thereby, improve take up and adoption of ENVRIplus products.

1. **Boundary crossing** The participating communities experience boundaries between the different roles identified above (see Section 4.1.8), between disciplines, sub-disciplines and application domains, and between organisations. Many of today's pressing research questions and many of the federations addressing them (see Section 4.2.3) require teams to form and to think and work effectively together across those boundaries. ENVRIplus can stimulate this by:
    a. Organising *ad hoc* think tanks so that it brings together (virtually) participants from across the boundaries and stimulates them to think and work together on relevant topics, e.g., by bringing in suitable experts and setting up suitable practical challenges to be addressed during the course. This requires elapsed time, and allocation of both training effort and trainee time, so the target understanding that the course will deliver has to be carefully chosen.
    b. Establishing suitable agile development processes where people work intensely together on a common issue with a carefully set goal. Then assimilating the results and building on the networks provided.
    c. Delivering services and tools well suited to each role and organisational context.
    d. Arranging workspaces that facilitate such collaborative behaviour while ideas are being developed and formulated. This requires those involved to have

control over the release and sharing of the material they work on. Individuals may be involved in several groups, probably with different roles.

2. **Integrated communication facilities** The individual elements of communication for distributed participants in an RI need to be conveniently integrated. There are several potential solutions in this area. It may help if at least one well-integrated one were run to be available for RIs, project participants and ENVRIplus. This needs to present views that work well for each category of practitioner. Some of the selected use cases in Theme 2 may serve to achieve this.
3. **Exemplars and early benefits** The development of exemplars of effective methods and software or services that support them is key to spreading ideas, testing them in new contexts and developing buy in. This will be helpful in the training and outreach programme. It is also vital as part of the process of delivering as early as possible benefits to the active researchers and other practitioners. If we can deliver immediate benefits they will not have to struggle for so long investing unproductive time in tedious workarounds. An example follows.
4. **Data access interfaces** Researcher and others managing data-driven processes spend a great deal of time, identifying data they want, arranging to be permitted access, arranging transfers, arranging local storage, arranging onward shipment to computation resources if necessary and returning storage resources when they have finished. If this is packaged as a convenient operation their work is simplified and more productive. The parts of such a process are all being built, but delivering an integrated solution that just works would be a large benefit. It needs the provision of a user's or group's workspace. It needs a means of identifying the required data. Once deployed, it can be grown in small increments, taking the users along an improving path. They might prioritise some of the following:
    a. Identification using queries over associated metadata (in the identity registries or in catalogues (see Section 3.2)).
    b. Extension of the operations that are easily applied to the accessed data (we have found visualisation particularly relevant).
    c. Handling batches of data consistently at the same time (the tea tray metaphor).
    d. Handling intermediate (transient) results with various aids for handling them in bulk and for clearing up afterwards.
    e. Promoting selected results to properly identified and citable.
    f. Arranging for their data to be published or curated.

4.2 Assessment of technology review

There are a number of pervasive issues that impact all of the technology reviews. These are:

1. Nurturing boundary-crossing collaboration to address new challenges.
2. Harnessing both numerical models and data-driven statistical methods so that they work well in tandem.
3. Data-intensive federation as a foundation.
4. Software sustainability a critical long-term issue.
5. Lack of engagement from the ICT industry.

Each of these pervasive issues will be explained below, as they would otherwise reappear in many topics. Then progress with each technology review topic will be assessed.

Cultural diversity

The understanding of the technological options has to take into account the diverse cultures of the RIs and their communities. These prior investments and differing cultures have significant value, to an individual, an organisation, an RI or a scientific domain. The cultures are reinforced by educational and induction practices. These cultures, the ways in which disciplines work have been refined to work effectively. Disrupting such cultures should not be undertaken lightly.

However, these legacy and ingrained elements present serious barriers to more rapid adoption of consistent or interworking approaches. It is desirable to find a path whereby ENVRIplus and its cohort of RIs is an island of consistency and coherence for its own benefits and as a beacon to others. Section 3 presents a comprehensive review of the options and technologies. It identifies the key players moving data-driven research towards the nirvana of consistent data treatment. It is crucial to invest sufficiently in these causes, by (i) ensuring that there is very effective internal communication for awareness-raising, education and decision support, and (ii) by actively participating in a two-way channel between ENVRIplus and the key external organisations. Exemplary solutions and working practices, well supported by software worthy of future adoption, will be needed to evaluate options and to rally rapid and widespread adoption within ENVRIplus. Key use cases launch the work needed for those exemplars—agile development teams build them. Once exemplars exist they should be used within ENVRIplus for the education campaign, and for external outreach to help the adoption of common practices and standards reach critical mass (see Section 3.3.6 for an example). However, the solutions from individual use cases will need broadening to become more generic patterns with wider applicability.

4.2.1 Nurturing collaboration between different fields

Every major discipline already has challenges developing collaboration and communication between its subfields. The culture developed through higher education normally addresses this by having a common core that spans the fundamentals of the active approaches. Over time, this core of mutual understanding is whittled away as researchers' progress, specialise and develop their own skills and knowledge in a particular niche. As academic tutors and research leaders, we are often guilty of steering those we mentor into focusing on a particular topic so they may achieve promotion or be successful in gaining resources and leadership. *Such attitudes and traditional mentoring behaviour may be outmoded and we may need new behaviours and cultures to exploit today's research opportunities and to address today's pressing challenges*. That much is well recognised in contexts such as ENVRIplus, but what are we doing about it?

In ENVRIplus, with its context of RIs, the issue is broader in scope and more central. Many have reported that they wish to collaborate, learn from or harmonise with other groups. The issue is two dimensional; communication across domains, subdomains and infrastructures supporting those domains is one dimension. Another critical dimension is communication between roles. Collaboration across roles is critical [Atkinson 2013b], where they seek synergy across three viewpoints: domain experts, data-analytics experts and data-intensive engineers. Roles have been enumerated in Section 4.1.8; we revisit and group the roles in Table 19. Although this is inevitably a simplification it serves to show that there many more viewpoints than three. They have complex inter-relationships that need developing and nurturing if the e-Infrastructures underpinning RIs are to serve their communities well and be economically sustainable. It should be remembered that in every role most of the effort is invested in routine work (95% in some estimates) that underpins all science. But the remaining moments of invention and introduction of new methods or technologies, that leads to new advances is key for scientific progress. That innovation is much more dependent on cross-boundary collaboration. Improving the experience and productivity of routine work has direct payoffs as well as making innovation more likely. Managing the innovation so that it does not disrupt critical routine work is a requirement, and that requirement propagates to the steps we take to improve cross-boundary collaboration.

There are occasional heroes who combine mastery of a wide range of these viewpoints to lead campaigns and dramatic breakthroughs. But sustainable and affordable science cannot be predicated on a sufficient supply of heroes, and even they cannot develop sufficient depth in more than a few roles or fields. Consequently, we have to become more expert at combining independently developed knowledge and skills, from different minds and from different cultures. This is not an issue that can be tackled by ENVRIplus alone, but it is in a key position to give an important lead. Table 19 illustrates the diversity of viewpoints and skills needed to deliver successful research and to make breakthroughs with global and societal challenges. The RIs may

review they have the right kind of experts to meet their goals. While doing this they may also take into account skills they currently access from elsewhere, e.g., EUDAT, EGI and ENVRIplus, and consider the sustainability of those relationships in comparison with their target RI lifetimes.

TABLE 19: SOME OF THE ROLES KEY TO THE RIS SUSTAINABLE SUCCESS

Role	Description
Domain specialists	
Campaign leaders	Research leaders have a broad view of their domain and a commitment to a particular cause. They organise resources and steer effort, raise commitment and maintain focus on the goal and the quality of every step on the path to achieving it. As a result, they are usually adept at boundary crossing and may underestimate the challenge it poses for others.
Theoreticians	Theoreticians develop conceptually satisfying and in principle testable explanations of phenomena and observational patterns. These may address broad issues in a domain or relate to some specific aspect of interest to one or more domains.
Experimentalists	Experimentalists devise and conduct programmes of work to test a hypothesis. This may be lab based or field based. These days it is often in silico, i.e., it uses computation to run models, analyse observational data, or do both. Repetition is often necessary to marshal sufficient evidence.
Observers	Observers organise and conduct the collection of data that represents manifest properties of chosen phenomena or systems. They may commission instruments or establish coordination to obtain sufficient information, with sufficient reliability, for a sufficient sample of their target set of measurements or records.
Instrument builders	Instrument builders may draw on many sciences, engineering specialisms and technologies to construct instruments that collect measurements of the relevant properties and that operate in the required context.
Citizen scientists	These can contribute to any aspect of a campaign, e.g., data analysis and pattern recognition in Galxyzoo[164], or field observations of bird populations with eBird[165] [Kelling 2013].
Curators	Curators establish and run the collection, publication and preservation of selected reference information considered important by their community.
Data scientists	
Statisticians	Develop the mathematics and practical methods for inferring information latent in data, taking account of potential biases, such as: sampling, measurement and recording errors.
Machine-Learning experts	Machine-Learning experts deploy the statistical methods, such as strategies for handling missing data and statistical inference, and develop and deploy algorithms over large bodies of data to obtain derivatives that represent actionable information. That is, they are able to assess how reliable those derivatives are relative to target decision making or calibration goals.
Problem-solving kit builders	These kit builders develop libraries of packaged methods that work well together for doing data preparation, performing statistical steps and visualising results. They provide ways of using these so that users do not need to understand the details. They are often provided as problem-solving environments, e.g., mobile-app encapsulations or browser accessible tools, where the user can conduct and steer operations on their data, without having to explicitly manage data or resources. They often have a workflow or scripting notation to allow users to encapsulate repeated tasks as one task.

[164] http://www.galaxyzoo.org
[165] http://ebird.org/content/ebird

Role	Description
Visualisation experts	Visualisation experts develop ways of showing the significant aspects of data effectively, i.e., so that scientists and decision makers are best able to see and use the significant (to them) information. These techniques adapt to a range of output devices from smart phones to immersive video caves. They use dynamic viewpoint and presentation change controls to allow users to explore data and recognise salient features otherwise hidden.
ICT specialists	
Systems architects	The systems architects shape the overall structure: the choice of a series of software subsystems, layers and services, a choice of the hardware architectures that should support the software, the organisation and provision of data storage, the provisions for user interaction, security and operations management, and the distribution and interconnection of all of these across organisations and computational platforms. As in conventional architecture, a key responsibility is to tease out the actual requirements and planned modes of use, to highlight potential conflicts and risks and to reconcile aspirations with available budgets and resources. As in conventional architecture, considerable use is made of prior designs and pre-assembled systems.
Software engineers	Software engineers are responsible for the good engineering of software, ensuring that it is fit for purpose, i.e., delivers the functions and facilities required, is capable of being run in all of the contexts where it will be deployed, will prove dependable, i.e., not fail catastrophically and without diagnostic traces, that it will be continuously available and that it is maintainable. This is a complex engineering task, where given prototype software that already runs in its originator's context for their envisaged test cases, may take from ten to a hundred times the original effort to achieve full scalability and deployability with acceptable dependability and security. As software may often have a long lifetime investment in its quality and sustainability from inception to end of useful life is worthwhile for carefully selected software—see Section 4.2.4 or www.software.ac.uk. For open-source projects 30% of the effort goes into user support, for commercially supported software this is typically 50%. Ensuring effective mechanisms for following up all user issues is another software engineering responsibility.
Data-intensive engineers	These engineers take the algorithms that are developed by data scientists, the workflows developed by research campaigns, organise the data on appropriate storage media and map the algorithms onto appropriate hardware so that overall goals can be reached economically or quickly. There are a variety of rapidly evolving strategies.
HCI experts	The Human-Computer Interaction experts study and improve interaction at all levels, from the atomic elements of HCI communication, such as touch screen gestures, to the deep relationships of shared knowledge and skills that affect interpretations of responsibility and preparedness to take decisions.
VRE experts	Virtual Research Environments are logical foci enabling communities to see a collection of computational services, data and supported methods as a holistic integrated resource that is easy to use. The design and construction of these, undertaken by VRE experts, has the usual distributed systems, architectural and engineering issues. In most cases, it also has the challenges of drawing on diverse, independently owned, heterogeneous, autonomous resources—see Section 4.2.3.
Communications experts	Digital communications underpin every stage of the data lifecycle from acquisition, potentially in geographically isolated locations and with low power availability, to the data centres hosting curation. It connects all roles of user from their office, home or in the field to the full gamut of services. It builds on many digital transfer mechanisms with different modes of funding and management. Blazing and sustaining trails through this terrain is the key responsibility of these experts.

Role	Description
Storage & DB experts	Data volumes, rates of delivery and rates of access all have to be met by these experts, by mapping data to appropriate technologies, trading longevity of storage against speed of access. They meet these needs drawing on a wide range of technologies, from traditional bulk tape storage to the latest solid-state technologies, e.g., 3D Xpoint[166]; its applications for scientific computing are being investigated in the EU NextGenIO project[167] . They draw on a range of communication technologies, organise data placement and data movement. They construct algorithms and access models to accelerate the common requirements, such as content-based searches, parallel writes and co-location of computation and data storage, by mapping onto appropriate software and hardware platforms.
Simulation experts	Numerical analysts and simulation experts take the mathematical models developed by theoreticians, sometimes mathematically described and sometimes as preliminary implementations, and transform these into algorithms that run well on the available architectures and achieve the required precision.
Theoreticians	Computing science theoreticians formulate models of computational logic, of distributed systems, of algorithms, of hardware architectures, of data representations and semantics, of actor systems and so on. Many of these underpin the above work, e.g., the original description of map-reduce and its types by Milner and Plotkin. The theory is essential if transformations are to be undertaken to handle the scale and diversity encountered, e.g., between many metadata forms. It is not expected that ENVRIplus researchers will engage directly with such theory, but the ICT experts that they work with should certainly be tracking the relevant theories for their viewpoint.
Systems administrators	Once systems are built they need to be provisioned, the new versions of hardware and software need to be deployed and connected in and the arrangements to allow access to resources while protecting systems from misuse needs continuous vigilance. They are often involved in aspects of innovation support, such as configuring and deploying new subsystems and software platforms.
Hardware architects	Many advances by electrical engineering and production engineering lead to growing numbers of available logical, memory or data movement components for a given power, cost and volume. Harnessing these advances to yield more of the power that science needs has to be by combining these elements in new ways, as it is not possible to simply make these components run faster. The variety of combinations is potentially very large, but the ready-made units are largely shaped by the dominant internet and entertainment businesses. Therefore, the hardware architects develop ingenuity in delivering science platforms using general purpose hardware to save costs. Only exceptional systems, such as the HPC systems operated by PRACE, have architectures tuned for very large-scale numerical simulation. Understanding which aspects of science need and perform best on particular hardware architectures requires engagement with these architects.

Short-term and longer-term strategies in facilitating boundary crossing are recommended. In the short-term, i.e., within the lifetime of ENVRIplus the following steps should be taken:

1. Recognise the value and contribution of each domain and each role.
2. The ENVRI week already brings together participants from multiple domains and multiple roles. Ensure that it invests and encourages inter-role as well as inter-disciplinary communication.

[166] http://www.intel.com/content/www/us/en/architecture-and-technology/3d-xpoint-technology-animation.html
[167] http://www.nextgenio.eu

3. The Agile Task Force teams can span roles and domains. Where they do this, they form a crystal of cross-boundary understanding of developing depth. Asking them to do anything else while their campaigns are intense would inhibit the agile behaviour and thinking required. However, in periods between campaigns, sharing their experience with others, e.g., during ENVRI week or in training programmes, would help expand and generalise the impact of their work.
4. Collaborative training programmes should ideally engage participants spanning domains and spanning roles. These might be a succession of webinars, scaled and timed to not disrupt the routine work. If a motivational exemplar hands-on practical can support this, which requires collaboration across roles and/or domains this will be effective at building understanding. Dieter Kranzlmueller, director of the Leibniz-Rechenzentrum (LRZ), has identified the importance of training that brings together researchers in environmental sciences with those in computing sciences[168].
5. Deliver intellectual ramps to new technologies and methods of working. New facilities are invariably designed by experts who are no longer aware of how much they have learned and how many skills they have acquired. For others to adopt these new facilities it is essential that there is an easy-to-get-started mode of use that usually doesn't reveal the full power and flexibility and then there are incremental steps that form a path to the full power. This helps greatly in accelerating adoption, but it also helps cross-boundary exploration and integration.
6. Organise initial summer schools that deliberately cross both domain and role values – see below.

The longer-term strategies should include the following:

1. In conjunction with others, e.g., through partner institutions and international bodies help support the development and recognition of collaborative careers. For example, some of the following steps which are today in operation in a few places, but which need more widespread investment, could be encouraged and supported:
    a. Collaboration exercises across disciplines as part of undergraduate training.
    b. Collaboration exercises and projects that cross discipline and technical-role boundaries to achieve goals that show the value of interdisciplinary skills and perhaps virtual collaboration.
    c. PhD programmes that deliberately require supervisors from different disciplines and possibly different institutions, that develop deep understanding of a inter-disciplinary issue.
    d. Career selection and promotion procedures that recognise the value of interdisciplinary communication, collaboration and creativity.
2. In conjunction with others initiate summer schools that deliberately bring representatives from multiple disciplines and from multiple roles to develop and share understanding of how best to address these boundary-crossing challenges.

4.2.2 Numerical models and statistical methods in tandem

The paradigm of using mathematical models to capture our understanding of the phenomena we observe has certainly been with us since Newton's era. It has had a tremendous boost as computers have become progressively more powerful, and it certainly plays a key role in the environmental and Earth sciences; for example, in seismic inversion and modelling convection in the mantle. Szalay has pioneered better use of simulations based on such models [Szalay 2013].

Jim Gray thought of that as the third paradigm, after the observational and experimental paradigms. He coined the term "The *fourth* paradigm" as a new way of observing and characterising data [Gray 2007]. It is driven by the tremendous growth in digital data delivered

[168] http://www.envcomp.eu/ and http://www.nm.ifi.lmu.de/teaching/Vorlesungen/2013ws/UrgentComputing/

from instruments, from monitoring digital activity, from numerical simulations and from harnessing citizen science volunteers. The scale of data and progress with statistical methods such as machine learning, also exploiting the growing computing power, has led to new ways of recognising and describing patterns in the natural systems of interest.

This introduces new opportunities for science and its applications as these two approaches: numerical modelling and statistical analysis of observations, can be harnessed together to achieve breakthroughs, and develop new understanding and applications. However, achieving and successfully exploiting such combinations is very challenging; to quote geoscientists in a recent data science meeting at the Alan Turing Institute [Aston 2016]:

> "*Methodologically, there is a major gap between statistical modelling and machine learning on one side and numerical or physical modelling on the other. Hence a systematic approach to consistent data integration and model building is of highest value and priority.*"

The challenge is widely recognised at several levels: (i) the conceptual frameworks, (ii) the implementation and encoding as scientific methods, and (iii) the best ways of resourcing those implementations [Fox 2016].

Many environmental and Earth scientists will be encountering these challenges and seeking to reap the benefits of successfully harnessing the combination of statistical and numerical methods. ENVRIplus should seek ways of pooling intellectual and practical effort to reap these benefits. There are potentially theoretical issues. There are certainly ICT issues in how to describe and support such activity. There are organisational issues about how to support the working practices involved in a scalable and sustainable way. At the very least ENVRIplus should kick off a strategy that includes these combined approaches, even if they do not become a priority in its time.

4.2.3 Data-intensive federation foundations

There is a great need for data-intensive federations in the environmental and Earth sciences; particularly, as they study multifaceted global phenomena. There are many application domains where practitioners are trying to exploit a growing wealth of diverse and evolving data sources. It is imperative to provide an affordable and sustainable environment, which improves their productivity as they develop and use data-science methods. We refer to the network of data and resource sharing agreements as a *Data-Intensive Federation (DIF*). Data-intensive federations are virtual distributed environments that organise the *repeated* use of *dynamic* data from multiple sources owned and managed by *independent* organisations into a *holistic conceptual framework* that makes it *much easier* for multiple groups of practitioners to perform their *data-driven work*. As such, they are artefacts that involve the construction and maintenance of social, organisational and ICT infrastructure. They need to include: crossing boundaries, establishing and honouring agreements, supporting multiple work environments, tool sets, services and technologies. They must enable practitioners to undertake decision or policy support, information services, reference data and research, using their framework, as many of the participants are funded to do such things.

We argue that building each data-intensive federation incrementally, as a one-off and in isolation, is wasteful in effort and produces solutions which are not only less effective and efficient than 'state of the art' but also inhibit interoperation. We advocate investment in R&D to develop foundational principles and reusable frameworks (or data fabrics to use the RDA terminology) that can provide the 'core' of data-intensive systems for all domains and can be tailored for those aspects of each domain that are specific.

Data-intensive federations require the following features beyond the data-warehousing and data-lake strategies that are used to support commercial applications where the data can be corralled into one regime under a single data controller today:

1. *Data will not be wholly owned or under the control of one organisation*. Instead a set of data providers may actively participate in forming the federation, while other data providers will remain external. Standard services from the external providers may be used or special arrangements negotiated with them.
2. The federation will identify and develop a *framework in which the combined data can be more easily used*, i.e., each user does not have to negotiate access or deal with idiosyncrasies. Each user does not need to assemble the data they require into their own working space. This can benefit data owners as they do not have to deal directly with every usage request nor have to handle all data accesses provided that they can obtain usage records to justify investment in their services.
3. This framework will deliver a *holistic view of the federated data* that facilitates current well-recognised tasks. It will also support experimental development of advances in the use and interpretation of data. A selection of these will lead to advances in the holistic view.
4. A new category of experts, *data diplomats* (an extension to the roles in Table 19), need to be supported. They will represent the different organisations and negotiate the rules for forming, using and evolving a DIF. For example, robust mechanisms to track and cite the use of contributing organisations' data so that they can demonstrate value added by their work to their funders—these may be binding contracts for sensitive data. The data diplomats need to be able to negotiate, formalise and record these sharing rules, and then trust that the framework will honour them even when multiple stages of derivation and caching have been used. For sustainability, any data provider may require revisions to rules concerning the use of their resources. Tools should be provided to understand the interaction of rules, their impact on priority uses and their propagation. An example of such a framework being developed for sharing medical data under contracts that meet local regulations in Europe is given by Elliot *et al.* [Elliot 2015] that delivers data-integration based on logically described agents[169] and their interactions [Robertson 2016] and [Papapanagiotou 2016].
5. Many forms of *dynamic evolution* must be accommodated. Examples include, the addition of new data sources, improvements to the existing ones, and changes in the trade-offs in the underpinning digital platforms.
6. In addition to supporting collaborative development of analytic techniques, and scientific methods using these data, in combination with simulations, the framework must support the development and evolution of *data-integration recipes*. These will be re-playable to have equivalent semantics in new contexts, e.g., when dealing with a burst of new real-time data and social media inferences.
7. As the work environment provided to practitioners now has such scale and dynamics, *research and innovation must be well supported*. That is, experiment and exploration must be possible with few restrictions *as if in the production context* without jeopardising production priorities. When such innovation is successful there should be a *smooth path to deployment* ('translation') for production use at scale in the controlled and managed context.
8. *Procedures must be resilient to change* in the external environment. Otherwise they will fail at a critical time, when used in a new context or after a provider has updated their service, e.g., when help is needed during a natural disaster. Without such resilience those responsible for geo-information services have to be very conservative to avoid such failures. If the resilience is established, it will keep the old procedures viable while new approaches are pioneered. As far as possible most changes should propagate automatically until all their consequences are dealt with, otherwise sustaining a DIF will require unaffordable and unachievable effort from experts to maintain services.

[169] In the computer science sense of algorithms that interact with other agents and humans in order that a behaviour emerges without requiring a central point of control.

They also require features from other lines of research and development; namely trust and reputation, automated formation and management for virtual organisations (e.g., Patel 2006) and network-centric collaboration networks (e.g., [Camarinha-Matos 2006]).

There are a growing number of application areas where such requirements are manifest. Sharing the R&D for the underpinning architecture and novel functionality will be worthwhile. The environmental and Earth sciences are an ideal starter community to work with as they have a great deal of diverse data that is already accessible, and a tradition of sharing their data to tackle both deep science and societal challenges.

The facilities will include arrangements for practitioners to perform tasks such as:

1. Finding, understanding and obtaining data they require.
2. Specifying data integration criteria and data preparation pipelines.
3. Requesting delivery of data for their use, or to an analytic process.
4. Collection or storage of results, with provenance and diagnostic records automatically attached.
5. Submission of data, including such results, to the existing federation.
6. Establishment of temporary additions to the integrated data including the addition of new models for organising those data.
7. Encouraging and facilitating the free sharing of data and methods.
8. Facilitating users' proper attribution of data to the contributing federation members.

All of the above will be conducted using a high-level and abstract notation that avoids distraction by, or over-tight binding to, implementation and target platform details. Many of them overlap substantially with the requirements articulated and the developments ENVRIplus plans listed above.

We may consider a logical architecture for such DIF with the elements shown in Figure 15.

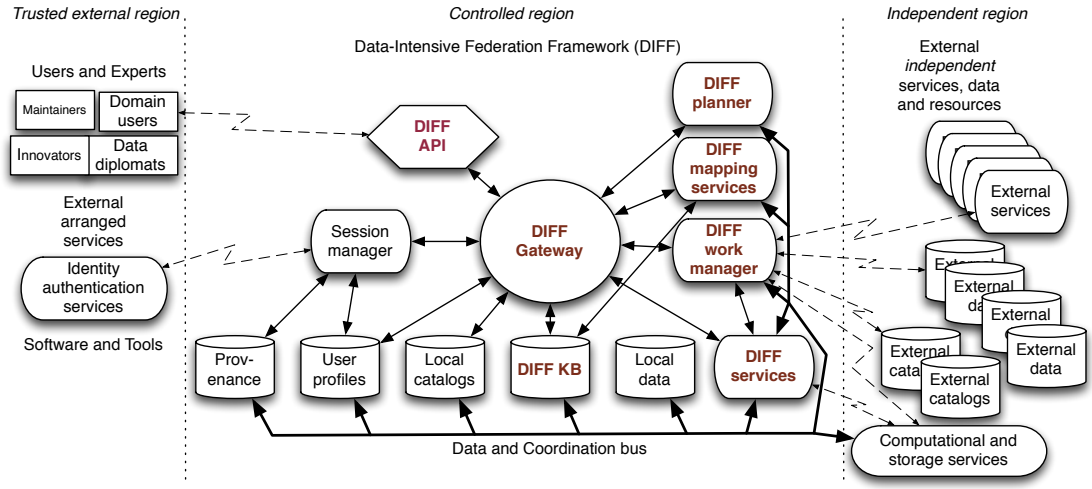

FIGURE 15: PROPOSED ARCHITECTURE FOR DATA-INTENSIVE FEDERATIONS

We see the world divided into three regions (divided by vertical dotted lines in the diagram; working from left to right:

1. *Trusted external region*: Organisations in this region provide services such as supplying credentials, software, e.g., tools and workflows, and crucial experts, particularly the data diplomats. Organisations in this region produce software that is trusted and can be trusted to assign authorisations to perform role appropriately. This varies from members of the public with few authorisations to rule specifiers with significant authority. The majority of the users are in this region and their work generates most of the calls on the services provided by the Data-Intensive Federation Framework (DIFF). Whilst this region is trusted to a lesser or greater extent depending on the credentials supplied, it is

entirely autonomous. The individuals and organisations in this region are involved in many activities. Some may depend heavily of a particular DIF and have regular patterns of interaction with it. Many will be engaged in more than one DIF or engage with the DIF for a fixed purpose or period.

2. *Controlled* region: The consortium taking direct control of the Data-*Intensive Federation Kernel* take full responsibility for this region. It is one administrative regime and governance system that is set up to steer the DIF's strategy, to resolve conflicting requests and requirements, to develop a collaborative ethos and to establish, run and support the DIFF for its DIF. It will need a business model to enable that role to be sustained and sufficient resources to provide the computations, storage and network connections needed for its internal operation and for all interactions with the two external regions.
3. *Independent region*: Here we find an extensive, potentially global, and heterogeneous collection of data and ICT resources that may grow incrementally when it is agreed that their resources are needed or they want the DIF to include their resources, e.g., to publish them to this DIF's communities. That growth will be subject to data-use agreements, either standard terms they offer or specially negotiated arrangements. The standard arrangements often have rules, such as usage rates. Special agreements may include an external organisation agreeing to notify the DIF each time it has a new release of its resource and specify the data content and representational changes. The DIF may then adapt its recipes to suit the new release and refresh the local cache avoid out-of-date version supply and to save repeated transport. In return the DIF will accurately report all data usage.

The DIFF should consist of re-usable subsystems that can be composed and configured across a distributed platform to: (a) meet the needs of the particular DIF, and (b) to monitor and maintain the various agreements that have been made to enable the use of external resources. Note that in some DIFs the quality and enforcement of these agreements has to meet privacy and ethical standards or meet commercial in confidence agreements. The governance of the DIF will determine its own policies and rules, which will also be captured and implemented via the DIFF. Many elements of the DIFF are just as ENVRIplus would build them, a few, identified by being coloured red in Figure 15. We provide a high-level description of those elements (in Table 20) that delivers a holistic and integrated view of an heterogeneous federation of data and computational resources while implementing and enforcing agreed inter-organisational relationships.

TABLE 20: THE ELEMENTS OF THE DATA-INTENSIVE FEDERATION FRAMEWORK

DIFF Subsystem	Functional description
DIFF API	A presentation via web services, often as microservices [Vianden 2014], and notification services of the data and facilities offered by the DIF. These will be organised as bundles, so that a user or tool developer community can often focus on just one bundle. The normal mode of use will deal with an abstraction of operations and data use that avoids technical detail. This allows development to take place outside the controlled region, in the trusted region.
DIFF Gateway	The **DIFFG** provides a stable API that supports tools and programmatic use in a consistent, coherent and stable manner. It directly initiates many of the functions that are specific to a DIF, such as recording agreements, providing the holistic view. It verifies that *all* operations are compliant with agreements, rules and contracts. In particular, it will provide interfaces for the work of *Federation diplomats* who negotiate agreements, recognise the established aspects of the holistic model, and devise mappings to deliver them, often drawing on recent innovations found via the *DIFFKB*. It will provide tools for recording encoded rules, for investigating the interaction between rules and for analysing provenance records to assess rule compliance. For rapid response to tasks entirely under the DIFF kernel's control it will directly call *DIFF services* or submit a description as to what is to be done to the *DIFF work manager*. For larger tasks

DIFF Subsystem	Functional description
	and all tasks that involve (external) services that may have changed since the task expansion template was developed it will refer them to the *DIFF mapping services* to be adjusted to the current context. The *DIFF mapping services* will then pass one or more workflows to *the DIFF work manager*. For large and demanding requests the *DIFFG* will delegate their organisation to the *DIFF planner*. This will analyse the requested workflow and decided whether it should be partitioned. It will then rearrange each partition taking into account the mapping by *the DIF mapping service* and information about target resources in *External services* and *Computational and storage services* and the performance of previous similar runs. The mapped and optimised partitions will then be delegated to the *DIF work manager*, which will call on specific local and external resources, according to the annotations provided by the planner.
DIFF Knowledge Base	The ***DIFFKB*** will contain information about the holistic model, its logical construction from the external and local sources and how it may be used. This will include explicit lists of allowable actions depending on a session initiator's authenticated identity, authorisation, current role and budget. The *DIFFKB* will have a viewable form that may be navigated or queried to support novices learning about the holistic model and federation, and to support experts extending their understanding and planning their actions. The *DIFFKB* will record how logical operations supported by the *DIFFAPI* are mapped to local and external services. The *DIFF Gateway* will use these mappings. They will be revised by *federation diplomats*, by automated optimisation and in response to changes in available data and services. The *DIFFKB* will include descriptions of the catalogues, the dynamic and static data, and local and external resources. These will eventually include the relationships between these organisation elements (constructed or discovered), the available operations and methods for using those data and how they should be used. User annotation will be encouraged.
DIFF Planner	Internally, every significant action on the holistic view of data will be represented in a suitably abstract workflow notation yet to be chosen. The planner will take a parameterised version of such a workflow, with its required data identified or embedding a mechanism for obtaining the input data. The *DIFF planner* will take into account the sources of the identified data (there may be multiple copies or they may be identified by queries over catalogues or data sources) and suitable target enactment services. It will transform the workflow to make it cost less according to an agreed or selected cost function. It will then arrange for the *DIFF mapper* to prepare each partition for execution, possibly in a coupled mode. The *DIFF planner* will record its treatment in the *DIFFKB* and will reuse that treatment when a similar request occurs unless the digital context has changed.
DIFF mapping services	The abstract workflows will need mapping for two reasons. The abstraction will omit many details, such as marshalling and moving data, implicit transformations, management and clean up of intermediate data, and target specifics. It will also handle changes that have occurred in the organisation or available facilities in the external independent federation partners and accessed independent services. These two forms of mapping are essential for sustainability. They deal with the inevitable and near continuous change in the digital environment. They retain freedom to revise choices of targets and computation arrangements, e.g., from using local resources to using an external resource or switching between Storm and Spark as a data analysis framework.
DIFF work manager	This takes requests for work to be done, either hand crafted for simple local cases of services offered via the *DIFF API* and submitted by the *DIFFG*, or as a result of the process described above to handle more complex work. The *DIFF work manager* makes final checks that the actions are authorised, comply with the rules and that the session owner has both the authority and allocated resources for the total request. It then finds the right mix of local and external resources to perform the task, recording at least the minimum records in the provenance store and sending results to the user—note that such response to a user may also have to comply with rules. Some of these requests or some stages within the request may be interactive.

DIFF Subsystem	Functional description
DIFF services	These are internal services to support all aspects of the *DIFFG*, particularly the work of federation facilitators, local data and catalogue management. This includes accommodating a wide variety of catalogues and of multi-faceted queries over them, if necessary generating actions on the data they reference. Services will also support rule definition, revision, testing and application. A local service will verify that a requested task, or a stage within a task is compliant with current rules. A multifaceted query over the provenance records may select a subset against which a rule can be tested, either to verify that a new rule now inhibits actions that were causing problems, or that a rule does not find fault with valid actions. The services should also support the development and testing of mappings and of requests that may be installed as available once they meet acceptance criteria.
External data	For a solid Earth DIF this might include data, such as the FDSN coordinated seismic trace archives, the LIDAR surveys, GPS streams, and the NASA and ESA satellite images, e.g., Copernicus and SAR, that are available. A DIF will choose, target and negotiate these. For example, this one may obtain data from fossil fuel and mineral extraction surveys, even though much of those data are commercially confidential. In some cases, this may require fairly strict rules about how those data may be used. Some data providers will also deliver identity services for their data, query services for selecting subsets, and host computations on their resources for computing derivatives. The may expect the accounting systems of the DIFF kernel to properly report use of their data, including reflection of consequent data derivatives and publications.
Local data	These are data directly contributed or collected by the federation users or by federation partners who choose to deposit directly. These will need to meet sufficient metadata standards that they can be used by other parts of the DIFF. Some automated tools for harvesting and validating such metadata will be developed as *DIFF services*. The local data will include caches for optimisation and user and group workspace. The handling of such data will depend on other services, such as those provided by EUDAT. The data – files, file collections, databases, and databases using a variety of models and representations and research objects – will all have suitable PIDs, so they may be referenced by methods and other parts of the DIFF without implying location and storage media. PID here means "*Persistent IDentifier*", so that it is unique within the required scope and persists for as long as it may be used. Not all of these need be permanent or publicly accessible. The framework may annotate local data to indicate such things as "locate with computation", "replicate for scale", "make durable", "archive", "transient".
Others	The other boxes in Figure 15 are the same as the corresponding functionality described in Sections 2 and 3

Data science is a fast growing field and research infrastructures have to be at the front line to best serve their users without bothering them with technical details. As such the RIs should have the right kind of expertise employed or otherwise hired. Following the initiative of the sister cluster project CORBEL to cooperate with the EDISON project[170], it is recommended that ENVRIplus also enters into such a cooperation to obtain a dedicated data-science training for their infrastructure staff.

4.2.4 Software sustainability a critical issue

Scientists, science and the applications of science are increasingly dependent on software. Consequently, this dependency has to be thought about as carefully as the dependency on instruments. When an instrument is designed, purchased, deployed and run for long periods relevant teams of experts are involved and if necessary trained at every stage. Extreme care is

[170] http://www.edison-project.eu

invested in engineering and production. Quality is of great concern and is assessed repeatedly. Upgrades during the operational lifetime take substantial planning and investment.

Software requires comparable care and similar engagement of appropriate expertise. As software is largely invisible and often acquired incrementally, today this attention is lacking. As more and more of the data-driven working practices depend on multi-layered stacks of software their continuity (ability to keep functioning) and quality depends on the underlying software being adequately sustained. Those meeting the challenge of deploying e-Infrastructure quickly or getting a new scientific method supported can be excused taking short-cuts and lashing together software components they find. However, those concerned with planning research infrastructures, their strategy and finance should recognise that this is building a potential software crisis.

If software is required to meet new functionality, e.g., new forms of analysis, or new capabilities, e.g., new sustained data rates, then adequate time must be invested in its design, development and testing. For example, the R&D campaign to develop data handling for the Large Hadron Collider (LHC) began in 2000[171], almost 10 years before the first particle collisions took place. When live data acquisition was delayed for over a year by a helium explosion leading to significant magnet damage, the team driving the Worldwide LHC Grid to production quality were relieved to gain extra time before full data rates as well as simulations had to be handled [Chalmers 2014]. Much research and development into smart data movement and optimal data distributions, as well as workload scheduling was need to reach the necessary operational quality. Much investment in developing skills and organisation was needed to achieve sustained running. Similarly, for the Square Kilometre Array (SKA)[172] the software R&D campaign to fully exploit the capacity of the synthesised aperture data acquisition via arrays of antennae forming the radio telescope has run in tandem with the physical telescope design. Prior R&D at LOFAR[173] forms a crucial input to this activity. ***Where major advances in data handling or data analysis, or modelling are required, adequate investment and time must be allocated to the software R&D***.

The **cost of software** is roughly 5% to 10% for its initial construction, and 90% to 95% for its lifetime maintenance. Maintenance involves three significant aspects:

1. **Bug fixing**: Dealing with errors that were not exposed by the initial testing (~18% of maintenance).
2. **Adapting to context changes**: The underlying layers of software, the inter-process communication and coordination frameworks, and the external services on which the software may depend, all change. The dominant drivers in the digital ecosystem are commercial. Science has little influence and must accept many changes. As described above (Section 4.2.3) most partners in a data-intensive federation are autonomous organizations, each driven by many pressures. They will change their services, data formats, choices of standards and so on without reference to others. The sources of standards a community chooses to follow will, when those sources refine their standards, cascade into software upgrade requirements. (~35% of maintenance).
3. **Enhancing functional and non-functional capabilities**: The moment a new working practice or method is introduced it stimulates ideas and change. There are already ambitions from the user communities and many external stimuli. These changes need to be resourced through the full software-engineering lifecycle from design to deployment if the science is to remain competitive. (~45% of maintenance).

In addition, commercial software vendors spend 50% of their costs on **customer support**, whereas, open-source software projects spend 30% of staff time on **customer support** [Swedlow

[171] European Data Grid (EDG), http://eu-datagrid.web.cern.ch/eu-datagrid/
[172] https://www.skatelescope.org/
[173] http://www.lofar.org/

2016]. Without that customer support, which includes courses and on line help, many users will fail to use the software successfully. This will either lead to them failing and not achieving their scientific goals or it will result in a stream of bug reports, exacerbating the maintenance costs.

Revisiting the comparison with an instrument, a prototype to prove that a detection method works, can be "*knocked up*" in a lab and tested without considering all of the engineering issues and lifetime calibration and maintenance tasks. Similarly, software can be built quickly to test an idea. All too often, it is then deployed into a production context without considering the lifetime costs were scientists to depend on it for their work. Of course, much software never becomes widely deployed. It is used for a short time (hence the context doesn't change); it is used by an individual or small group (and hence latent errors are not exposed) and then is forgotten (hence never needs upgrades). We should carefully provide an environment where scientists with a few co-workers can easily build, deploy and then pension off such software.

The focus of software sustainability, however is the subset of software that does persist; does become key to the culture and working practices of a community, and which therefore needs to be engineered with care as it has become a mission critical dependency. This subset should be carefully identified. There will be a continuous stream of candidates. Management must choose the subset very carefully, to:

1. *Not miss software on which their community depends*. Visible services, operating systems and compilers, etc. get bundled into the provided platforms by default[174]. Major simulation suites/codes have much longer lifetimes than hardware platforms and tend to have a community that is supported to invest in their engineering and maintenance. The challenge is to spot all of the subsystems and the "*glueware*" that assembles them to provide the research environment, often embedded in science gateways and in specific services, such as cataloguing and curation.
2. *Not expand the subset beyond the capacity of the software engineering resources they can muster*. A significant role of management in the software industry is to keep killing off software projects. Such projects often spring up from ideas and stimuli their teams have. Some of those ideas are vital for the next project or the future of the company; many are not. The same will be true for ideas emerging in an RI's community with respect to their future success. Pruning this subset so that the remaining software have enough engineering resources is a continuous and demanding battle.

The RIs and ENVRIplus should have in place the management effort and decision procedures to identify and maintain an explicit list of the software elements that are in the subset that needs to be well engineered and carefully maintained[175]. After the end of the ENVRIplus project this responsibility has to transfer to the ENVRI RIs community at large. There will be a significant list, which will be beyond the resources of ENVRIplus and the RIs alone. A strategy is needed to handle this mismatch. For more background material, and campaigns to raise this agenda with funders, readers are referred to the work of the Software Sustainability Institute (**SSI**)[176].

Three strategies are available for investing sufficient engineering effort in mission-critical software:

1. **Buy or co-develop software from a commercial vendor** that delivers the quality, maintenance and support. As vendors often have millions of customers, they can invest in quality engineering and develop teams of experts in user support. Often this will not precisely match the research community wants. Two ways around this are:
    a. **Engage the vendor in your cause.** This may not be difficult, as there are many commercial opportunities that emerge when the vendor has a new capability in

[174] Though there are traps here for the unwary, as software often depends on specific versions, that are not always available in the context.
[175] That is, which support their science mission and operations.
[176] http://www.software.ac.uk/

their product. A fine example is the interaction between the astrophysics community undertaking high-data volume sky surveys with Microsoft – see Section 1 and [Szalay 2008]. The handling of the data-intensive federation challenge is a already a pressing problem in many industries – for example, an accountant in one bank assessing risk has to deal with seven different loan systems each time as there is no holistic view and no prospect of reducing the number of loan systems [confidential communication 2015].

   b. **Meet most of your needs with the vendor's software.** Often a large part of the logic, or most of the functions may be obtained from the vendor's software. The extra functionality is written using these. The original requirement in the essential list is reduced to and replaced by this remaining wrapper. The engagement of the vendor, considered above, may still leave the ENVRIplus community with some such residual.

2. **Engage in an open source software campaign.** These are run under established regimes, such as the Apache Software Foundation[177]. The value of engaging (or leading) such an open source campaign is that it should gather a critical mass of contributors with sufficient software engineering skills to deliver and maintain the required product. This campaign must be well governed and recruit key experts, e.g., for the various target platforms, for the optimisations required and for systematic testing. It also needs the necessary user support team. Although the software licence may be free, participation implicit if you are a serious user, has to include contributions to the overall effort; at the very least user support for those with similar requirements. As for commercial software, the available open source products are unlikely to match the ENVRIplus requirement exactly. Again two options must be considered:
   a. **Engage in the Open Source campaign.** Working with the open-source community for this product, extend it to meet your needs. There is likely to be good support for all the stages of software engineering needed, as the active community will have established this. There may also be others in this open-source product's community who have a similar requirement. It will then be possible to pool your design thinking and software-engineering staff resources to deliver the new capability.
   b. **Use the open-source product with a locally supported extension**. This is the same as using commercial software to do as much as possible. Again the residual replaces the previous member of the critically important software set.
3. **Take full responsibility for a software requirement.** This is the last resort and the full range of maintenance tasks and user-support tasks must be resourced with suitably trained and skilled staff. Generally speaking, this is too demanding and expensive for a single organisation or small user community to take on. It is usually necessary to form alliances to obtain or extend the necessary resources—ENVRIplus is just such an alliance. This will then lead to bespoke software tuned to meet the identified purpose without unnecessary overheads or unnecessary interfaces with other software. *It will also result in a long-term responsibility for that software's maintenance*. After care in selecting the mission-critical subset, and reduction of costs using the above two strategic options, there will be a residual set of software items, with reduced functionality where possible, that needs to be built and maintained. A locally resourced team may open up new territory and yield a product that can move into the commercial sector or initiate an open-source campaign. In most cases it will not and all of the resources will need to be found and sustained locally.

For each member of the critical-software list the ENVRIplus and RIs will need to decide how best to treat it according to the above three strategic options, taking into account the RI-lifetime costs. They will then need to find, or pay for as outsourcing contracts, the necessary

[177] http://www.apache.org/

skills and software engineering capacity, and sustain that resource for as long as their user communities depend on their software.

There are traps that catch the unwary and these will need to be managed throughout the lifetimes of the RIs. For example:

- Depending on a vendor's product that is discontinued or changes its licence model.
- Depending on an open-source project that loses critical mass.
- A new middleware platform or toolset overtaking one that was previously best.
- Missing critical software because it was not high profile. No-one knew it was there or what it does.
- Inheriting, or choosing to build on, unmaintainable software.
- Failing to adopt the 'standard' that becomes widely adopted.

As well as taking responsibility for their own bespoke software, there are three forms of shared software maintenance that every RI community with an software dependency needs to be responsible for: a) their fair contribution to the multi-community software elements; b) the mappings to and integration between the common software elements to meet their specific needs; and c) on the hopefully rare occasions when a major element needs to be replaced by a thriving alternative, the integration of that alternative. Today this maintenance investment is only available for novelty items and recognised simulation codes. Many other software elements need maintenance for the investments in e-Infrastructure to survive and for the improved research environment to be sustained. Funders, research strategists, organisations providing platforms, e-Infrastructure builders and VRE developers need to form alliances to achieve this for the research infrastructures that are strategically important. The communities of researchers and others using these facilities may need to campaign for this to be included in the long-term agenda.

To summarise, Mattmann identified sustaining four research tracks as critically important for future data science, based on his many years of experience at NASA and at the Apache Software Foundation [Mattmann 2014].

- Rapid scientific algorithm integration.
- Intelligent data movement.
- Use of Cloud computing.
- Harnessing the power of open source in software development for science.

The first three of these depend on software, and in many cases it is built with a substantial component of open-source input identified in the fourth bullet. All of the routine user interaction is through VREs and portals that require many elements of software in and behind them. All of the data handling throughout the data lifecycle depends on software: tools, workflows and services. All of the innovation depends on shared development environments, IDEs and APIs, again totally dependent on software. The dependency on software is pervasive. *The commitment to sustaining all such critical software has to be equally pervasive. It must be sustained for the lifetimes of the RIs.*

4.2.5 Assessing the data identification and citation technology review

Section 3.2 gives very clear explanations of the value of good quality working practices for *identifying and referencing* all items of data that are, or may become, significant in research. It highlighted two pervasive challenges faced by all those who are engaged in stages of the data lifecycle or are using or producing data in their research or for decision support. One, there are diverse suggestions, but not agreed and widely adopted standards, underpinning the necessary actions, whether those actions are carried out by humans or software. Two, today there aren't good tools and technologies that make it easy for humans or software to perform these tasks efficiently. There is a great deal of work underway, and we can be optimistic about viable deployable support for data identification and citation becoming available within the next few

years. This poses another two challenges. One, how to identify and align with the software and methods that will be most widely supported and adopted, and two, how best to use the emerging software, metadata standards and proposed methods in the ENVRIplus context. That requires developing standard practices, metadata and protocols that allow interworking within and between the RIs and other organisations. That is an issue prevalent in *nearly all RIs for all technology topics*. Indeed, *cataloguing*, *curation* and *provenance* all need to make effective use of the functionality and facilities data identification and citation will provide. Conversely, the work on catalogues may provide facilities for PID registries with associated metadata.

Scientists in each field will need to associate their identified items with concepts in their view of the natural world. Terms for widely adopted *agreed* concepts may be identified by standardised vocabularies underpinned by formal ontologies – see Section 0. For such agreed concepts these external references provide identification and citation. However, scientists may take different views of the phenomena they observe, or they may be developing a conceptual framework for new phenomena, e.g., a new species, that they have identified. In this case, they need a framework for defining and citing the new concepts that they manage and develop. Presumably, this would use the data identification and citation machinery. When their contributions reach acceptance or are published these localised identities should easily migrate into the standard reference space of managed identities. Conversely, if they fail to establish evidence to back up their idea, their localised developments will not affect researchers other than those they are currently collaborating with.

Optimisation will interact with data identification for two reasons: both caching and the co-location of data, processing and derived data depend on precise data referencing. Integration into workflows of the functions required for data identification and citation is a crucial labour and error saving step. Processing will then need to execute the data-intensive workflows and call on data identification services.

The basic consistency for data identification and citation should be achievable within the ENVRIplus project's lifetime (see Sections 3.2.3 and 3.2.4). But, as in so many scientific contexts, this leads to further challenges. In this case dealing with the more complex, composite and time-varying data generated by the work of RIs and their research communities (see Section 3.2.5). Finding ways of succinctly, efficiently and precisely identifying the growing volumes and many subtleties of the data used by and produced by future data-driven research will always be a challenge. As one aspect is covered, increases in volumes, increases in rates, increases in diversity and researcher ingenuity will pose new ones, or break existing solutions. It is vital to be on the ladder addressing such issues, as that is key to international research leadership and to addressing societal challenges. There is clearly good reason to believe from the understanding and insights shared in Section 3.2, that data identification and citation will be progressing well up that ladder during ENVRIplus.

There are further considerations that may be addressed in the future. These are enumerated here in no particular order—many of them apply to subsequent technology review topics as well.

1. **Identification and cataloguing:** The relationship between data identification and cataloguing is very close. The identity record could also be a catalogue entry for the referenced data. The metadata required for the identity purposes could be a subset of the total catalogue metadata. An operation, possibly a standard query pattern, could yield the information required for the purposes of data identification and citation. It seems unlikely that independent development and support of identification and cataloguing will make long-term sense. However, there may be a distinguished subset of catalogues that register data identity and the identifiers these use would be used in other catalogues, provenance, processing and optimisation. The principles and procedures for minting adequate references would still need to be independently designed and agreed. But their implementation would employ appropriate catalogue functionality. At which point, non-functional issues, such as performance and availability

would come to the fore. Harmonised solutions would simplify interworking but they confront established cultures and investments.

2. **Roles for data identifiers:** It is obvious that cataloguing and provenance will need reliable data identifiers to refer to data from their records. It is desirable that almost all data processing refers to its inputs in terms of the data identifiers. This permits implementations to store copies of data replicated for availability and preservation, and then to choose the one with lowest contribution to the costs identified by optimisation. As the storage technologies, resource provisions and data-intensive middleware evolve, if the scientific methods are couched in terms of data identifiers rather than naming systems based on particular storage schemes, the mappings invoked during method enactment can adapt to those changes. Hard-wired naming means everyone who is involved in formulating the method has to be involved in adapting to the changed digital context—clearly an unsustainable policy. Indeed, the role for data identities in optimisation is much greater. If data are identified by a trustworthy mechanism, the optimiser can recognise when the same data are requested on two occasions (in the same workflow, by the same user running a different process, or by different users) and save work accessing, transferring and transforming it again.
3. **Raising the level of discourse:** The majority of discussions in requirements gathering and in technology review were couched in terms of practical and concrete implementation terms. Such delivery mechanisms are critical[178] but developing precise abstract models, then clarifying them through discussion and revision is of greater and long-term benefit[179]. The Reference Model (see Section 3.10) provides a vocabulary and context where such discussion takes place. Analysis and decisions couched in these higher-level terms are much less subject to the uncertainties of digital-technology evolution. Those decisions tend to have a rationale that is not based on the demands and issues of the current projects or current equipment and its software. Rather, it is based on the scientific and community goals. These need to be shaped and brought into as much harmonisation as possible. That is easier at a more abstract level. Once the goals are agreed, the mapping to implementations can develop and yield the best approximation to those goals given current circumstances. For example: Should the minting of a data identity be an atomic process? That is one that happens all at once leading to a complete and final record; i.e., the requestor would provide the data and all of the information required for metadata in one go. The identity system then allocates a PID, and builds all of the associated metadata, and makes a “permanent” record of the minted association. Or should a non-atomic, incremental sequence of operations be supported? For example:
    a. Either the workflow requests a data PID to allocate, or sends the data and gets a PID.
    b. Later the workflow can supply the data and required metadata.
    c. Later the workflow can say that it wants the PID promoted to a preserved status where it has been quality assured, or say that that binding between PID, data and metadata should be discarded.

4. **Transient identities:** Such incremental approaches might allow internal identities for workflow intermediates and potential result sets to be allocated quickly. Clearly, fixity is

[178] Specific observation networks have been doing a good job, refining their methods and delivering their data, for 20 years. However, engaging with this within the reference model framework will enable new data usage and identify opportunities to pool resources.

[179] This should not inhibit rapid developments of specific solutions and boundary crossing understanding achieved by agile design and development methods. There, a very focused approach breaks through barriers and creates an immediately adoptable prototype. Once the solution is understood in this particular context, standing back and casting it in higher-level terms enables others to draw on the understanding generated, and it enables the originators to plan the future path to general deployment and production engineering. This will be helped by a reference model expert joining in some of the agile campaigns.

not important at this time, and costs of supporting it and other metadata would slow a workflow. Many workflows fail or are under development, so their information should not be captured and curated, but their intermediate data may be highly relevant for diagnosis and for testing sub-tasks during development. What kinds of operation are allowed on a reference? Can a workflow or user retrieve the data associated with it? Can they retrieve (aspects of) the associated metadata? Can they formulate queries on the metadata that retrieve a set of identities? Can that set be used to perform a bundle of the above operations? And so on.

5. **Accommodating diversity:** Although the campaign for harmonisation is vital, it will never completely achieve conformity for two reasons: (i) there are many external forces, such as collaborative and global agreements that lock RIs and their communities into different standards and working practices, and (ii) researchers and their support teams are ingenious inventors of new methods—crucial innovation on which advances are built. We would not wish to inhibit this. The former can be accommodated by offering functions that map, sometimes with loss of precision, to a common interchange form, e.g., RDF triples conforming to the semantic web model [Berners-Lee 2001]. Is such a capability required for data identities and the operations on them? The user-driven variations can be accommodated by having some fields in the data-identity registry capable of holding any user defined material, e.g., JSON format records, RDF records, XML records, matrices, or plain text. The ability to use this information in queries would then be required [Spinuso 2016]. As most underlying database systems are capable of accommodating such flexibility (see below) this is potentially feasible. It is most likely to be required as the designated community becomes adept at using the existing facilities; hence, we would not expect it to have much prominence in the current round of requirements.
6. **Registry platforms:** It is necessary to build registries and other catalogues on top of high-quality database systems. There may be more of a catalogue framework above the database provided by the cataloguing campaign (see Sections 3.4 and 4.2.7) that delivers functionality that packages the underlying database semantics (so that, for example, the database delivery platform may be replaced). Database engineering has a huge investment and that helps address sustainability issues raised in Section 4.2.4. It also delivers scalability by use of multiple nodes, delivers accommodation of multiple data models, handles NoSQL and SQL distributed queries, and delivers mechanisms for high availability and mitigation of systems failure that might otherwise cause loss of data or inconsistent states. One approach to sustainability is to use *widely supported* Open Source projects, such as some of those under the Apache foundation, www.apache.org. Another is to depend on commercially supported software; there are several good database vendors.
7. **Temporal patterns:** Are all of the registries or catalogues built incrementally? For example, many IVOA catalogues are rebuilt with a specified periodicity. This has two advantages: (i) they act as a stable referenceable data source for that period, typically six months, and (ii) redesign and improvement of the catalogue building methods can progress during that interval and then be released. Are the visible registries of data identities always built incrementally? Are they ever rebuilt? What is the model for rebuilding them?
8. **Distribution patterns:** The registry of data identities for a particular community is almost certainly *logically* presented to them as a single entity. However, is it in fact a distributed entity, e.g., to handle transient and initial interactions locally and fast, but to promote persistent complete records to authority sites, replicated for durability and availability? Is this a recursive federation structure (see Section 4.2.3), where sub-communities within an RI pool their data identities and a group of RIs then pool these identity integrations? This may be more consistent with requirements for autonomy. It may also be easier from the viewpoint of incremental adoption. However, it almost certainly

increases the operational and implementation complexity. It may be a later adaptation; but that may be easier if it is anticipated.

9. **Jam tomorrow is not enough**[180]**:** Having visions of harmonised support for multiple RIs' data and communities and a well-planned path to deliver that goal is important. However, it will be a futile and wasted effort if the results are not adopted by the majority of relevant practitioners: the researchers, the technologists, engineers and managers who support them, and the users of the produced derivatives. Experience has shown that promising exciting advances but not taking the practitioners along with you leaves them finding alternatives to get today's work done. Once those alternatives have been developed, it takes a very long time and a lot of effort to re-recruit the community. Consequently, all of the important categories of practitioner have to see benefits as ENVRIplus progresses, i.e., they have to see **Jam today**[180]. This means finding immediate benefits for them, e.g., tools that help with their common tasks, data that they want, such as the usage of their data summarised, and automation of parts of working practices and scientific methods as soon as they become available, made accessible to them. This links back to raising the level of discourse above (3), as if we identify critical subtasks, e.g., request a data identity, issue a data identity, present a data identity to have action X applied to it, etc. then the discourse contains concepts that can be incrementally developed and can prove useful almost immediately. Their usability would be much helped by linking with communities of tool builders, e.g., the visualisation services in [Spinuso 2016]. Myers *et al.* [Myers 2015] identified the value of making the stages of curation incremental so that those stages yield benefits to practitioners as they work, e.g., they benefit from having some metadata associated with data in a catalogue, so they can find subsets of interest to them, and apply operations to all members of such subsets. This induces them to introduce metadata terms, such as dates and significant properties, because they are relevant for their current research campaign. Those same metadata will have been tested and improved by the time they are used for curation.

4.2.6 Assessing the data curation technology review

Data curation technology is reviewed in Section 3.3. Data curation is always important to allow independent review of scientific methods and of decision-support service output. It also offers a reliable repository for an open-ended set of researchers, experts or members of the public to access the data for any future research. There may be restrictions such that they require to show authority or that they are limited in their uses of the data obtained or in the resources that they may consume extracting or processing the data. In the RIs driving ENVRIplus they may collect observations of phenomena that will not be repeated. There is also growing pressure for curation from funding authorities. These all combine to make data curation essential for RIs in the longer term. Curation is more than archiving, it oversees the processes of deposit and access to maintain the quality of the collection and support its appropriate use.

Some RIs, such as ELIXIR, are already involved in long-established agreements for sharing the responsibility for curated life sciences data. Multiple organisations federating to curate collections distributes the cost of access and support, pools effort for quality oversight, and improves the protection against information loss, through multiple copies and multiple funders. In such contexts, the arrangements may be long-standing, e.g., for PDB [Berman 2008]. Similarly, many RIs are engaged in global commitments for curation, for example the data collected by Euro-ARGO needs to be curated and made accessible according to the global programme of ocean observation. Such long-term or collaborative arrangements set the scene for specific data-

[180] In English stories about poverty, when the children whine about their lack of food, mother promises "jam tomorrow!". Disillusionment, depression and crises loom when tomorrow never comes. See http://www.phrases.org.uk/meanings/jam-tomorrow.html.

curation campaigns. However, in the ENVRIplus community there are many who will gather or produce significant data without previously established models and practices. For these it would be beneficial to identify common practices, widely adopted and relevant standards, and supporting software so that they could have better prospects of their curated services interworking and benefit from shared implementation and support effort. Section 3.3 identifies some key standards and the coordinating standards development organisations, particularly RDA, producing the patterns for curation that may be widely adopted. The current state of RIs needs to be further understood and common solutions stimulated by a programme of awareness raising and training. By bringing together the RIs that are at a similar stage, possibly with experts from the Digital Curation Centre[181], and with potential providers, such as EUDAT's B2SAFE group[182], there will be a better chance of alliances forming, leading to common solutions.

Some of the considerations enumerated for data identification and citation – see Section 4.2.5 – also reapply here, unsurprisingly as curation almost certainly requires all of the steps of identification to have already been taken. Furthermore, the citable properties of data will probably be used for extending the set of metadata referencing the data, for referring to re-used type, format, interpretation and so on descriptions, and for forming related groups of data all of which may be the subjects of curation.

1. **Identification and curation:** The relationship between data identification and cataloguing includes the uses of identity in curation listed above. It is also possible that the mechanisms for longevity will draw on the same archival platforms and operational arrangements. Particularly, as the curated data would lose its value if the identities of the preserved data and of data referenced in relationships were lost. It is possible that when an archival process designates material *no longer curated* (eventually resource and relevance management may require this procedure), the curation workflow will place an "RIP" tombstone indicating the data's demise in the identification system, so that subsequent access requests can receive an informative error message. There is a similar relationship issue with cataloguing. Do the curation services use the same catalogue services and protocols as those that support other functions in the RI? Do the catalogue services depend on the curation services catalogues?
2. **Roles for curation:** It is clear that there is an immediate demand for straightforward data curation to meet science and decision making validation goals and to curate a reliable record of the state of observed and modelled systems as time progresses. This may also accommodate a reflection of the current state of understanding of the observed phenomena and the mechanisms behind them. This may be achieved by capturing the numerical models and their representations as simulation suites, by capturing the ways in which those systems are run, by capturing the ways observational and simulation data are then used (the formalised workflows) and the related documentation and publication. These may be equally important in validating approaches and in studying how the view of environmental and solid Earth systems progresses—see Section 4.2.2 for one reason why that may be an interesting focus. Curating the working practices, perhaps as summaries of provenance records, might complete the picture. This will allow the study of the changing populations of users of each facility, the changes in hot topics and the response to changes in the modelling and analytics arsenals.
3. **Raising the level of discourse:** The discussion of what should be curated, what properties it should have, of the responsibilities in that curation process and the responsibilities and functions of the system (platforms and organisations) needs to operate at the RM level, until there is clarity and precision. This should also include the ways in which the curation system may be accessed and used. Can it also be treated as

[181] http://www.dcc.ac.uk
[182] http://www.eudat.eu/b2safe

an active data repository from which data can be retrieved and used and into which results can be submitted. Clarifying such matters before getting into the engineering and technical detail is certainly necessary.

4. **Transient curation:** There is a case for time-limited publication of data to allow wide access and use of data that is expensive to obtain but that can be recomputed. The well-known examples are the results of large simulation runs, e.g., of fluid mechanics (mantle circulation), of seismic wave propagation, of cosmic and astrophysics events, such as the big bang, merging black holes and astro-seismicity, and so on) [Szalay 2013]. It need not be simulation, it could be a very extensive machine learning run over all available ecosystem data, to characterise relationship patterns. Using the curation system for publishing and accessing such results would align the access and resource mechanisms with those used for other data. In consequence, users would find their authorisations, resources and working practices will also work unchanged on these data. The temporary nature has three motivations:
    a. The results are often very high volume, e.g., time series of multi-dimensional data.
    b. The models and their mappings to simulation code are progressively improved, and the boundary condition data are also improved, the result is that sooner or later a re-run supersedes the previously curated run's data. But the curation of the fact that the previous results with their provenance existed must be curated and retained, so that the provenance of derivatives of the model make sense. IRIS[183] is acting as a repository for Earth models that each involve between 10^6 and 10^7 finite elements. But larger models that correspond to time series of states are far less tractable.
    c. Supporting work in progress—see (9) below.
5. **Accommodating diversity:** The campaign for harmonisation is particularly significant in the context of curation. Inevitably, it will never achieve complete coverage of facets of the metadata and representations, nor achieve universal adoption. The curation system normally has to have capabilities to cope with virtually any data that becomes important in the communities that it supports. This is very difficult to predict, so the curation model has to be open ended, though possibly optimised for the frequent and time critical activities.
6. **Curation platforms:** There are already very large-scale and long-running curation systems, e.g., those that support life-sciences reference data, e.g., PDB [Berman 2008] and those that support sky surveys [Szalay 2008]. The life sciences variety tend to have a number of professional curators overseeing quality and operations, even though they are highly mechanised. This may be mandated by their age. The Virtual Observatories accommodating sky surveys are invariably built on databases and involve less human labour, but very sophisticated workflows that rebuild the catalogues from the observations periodically. Organisations that provide national data curation and derived information services, such as BGS and INGV typically build on commercial database platforms, but develop sophisticated models and operational regimes on top of these. It is a moot point, the extent to which these could all be built on a common platform today, and amortised across more uses. The digital curation system iRODS[184] originally focused on digital documents, which being human generated are limited in size, has the aspiration to be such a general purpose underpinning technology. It has micro-services triggered by curation events, e.g., submitted data of a particular type or from a particular source that can implement any programmable rule. In principle, therefore, it is very flexible, and it is supported in Europe by EUDAT as B2SAFE. It may be limited in its

[183] http://www.iris.edu/hq
[184] http://irods.org/

capacity and ability to cope with diversity. For database solutions to such limits see this item in the list in Section 4.2.5.

7. **Temporal patterns:** Are the contents of a curated collection continuously evolving as each request for items to be curated comes in? Or is there a regime whereby updates are grouped to mitigate potential inconsistency problems and to provide periods of stability? There will also be periodic refreshes of the underlying machinery and software, presumably carrying forward the curated data faithfully.
8. **Distribution patterns:** The curation service and its implementation serving an RI or a group of cognate RIs, will normally be presented as a single logical organisation. At the very least it will have behind the scenes copies replicating information at multiple sites, delivering the LOCKSS principle[185]. There is potential for other, more significant partitions to keep data close to where it will be processed or where it was produced, in order to save movement costs or ownership concerns. One motivation for this is presented in the next item.
9. **Jam tomorrow is not enough**[180]**:** Delivering **jam today** to the curation user community is a challenge. Myers *et al.* [Myers 2015] report some success with this. They made the provision of workspaces and storage, and automation in its handling, a payoff for practitioners as they started using the system for their work in progress. This required them to incrementally provide metadata and quality assurance that would eventually be needed for curation. But they benefitted from help with managing the campaigns in their data-driven research, and in finding and using relevant subsets of data. Spinuso *et al.* [Spinuso 2016] have also shown that such tools, in this case running over provenance metadata, can visualise and mechanise organisational tasks that become a serious chore as the number of data items and research steps rise. Such integration of shared workspaces with the curation machinery would address one of the missed requirements – see Section 4.1.9 –but would require a distributed curation system, so that workspaces could be close to sources, sinks and compute resources that the particular user group is actively using.

4.2.7 Assessing the cataloguing technology review

Cataloguing (see Sections 2.3.3 and 3.4) plays a fundamental role in providing efficient indexes to accelerate the access to any items that the RIs and their communities choose to collect, collate, describe and organise. A catalogue associates an agreed description of each item, metadata, that summarises the item and specifies how it may be found, used and interpreted – the creators and users of the catalogue decide what the items should be, what the descriptions should contain and enable, and what can be left implicit or open ended – the engineers organising the implementation of the catalogue need to decide how the parts of each entry should be created and maintained with sufficient quality, and how the operations on a catalogue can best be implemented. The allowable operations have to include access by searching, but the query system specifying the search and implementing it is a design choice. The other operations often include:

1. Obtaining a reference token as a result of a query, or a *"no matching items"* result.
2. Obtaining all or part of the information associated with an entry, based on a query or token.
3. Obtaining all or part of the information for a subset of entries specified by a query.
4. Applying specified (from a supported set or user defined) operation to all the entries obtained by a query.
5. Adding information about an item.
6. Adding items.

[185] http://www.lockss.org

7. Using the catalogue as a primary data source to analyse properties of the population of items it contains.

Such catalogues provide a crucial resource around which a discipline may organise the collection and use of data; indeed, they were the initial focus of all discussions in IVOA (see Section 1). They are similarly important for the organisation of the storage and use of the data. Hence, they underpin Identification and Citation, Curation, Processing, Provenance and Optimisation, and may be crucial in many aspects of operations and management. *Their importance and central role cannot be overstated.*

In ENVRIplus and in many of the RIs, there are already many existing uses of catalogues –Section 3.4 – both within individual RIs and in some cases spanning a group of RIs. These draw on widely supported technology, such as CKAN, in many cases and often use a core of standards for metadata and its representation that is built on international campaigns for developing consistency.

Catalogues to hold and manage access to frequently used and critical data have been central to computing since the days when Alan Turing shaped the campaign to crack the Enigma code at Bletchley Park. Three critical properties are expected today:

1. They should be understandable, easily used and shaped by the user communities that commission them—a great deal of domain-led debate about exactly what should be in their catalogue and what operations should be well supported is a necessary investment by every discipline, sub-discipline, observational programme and experimental campaign. With serious intellectual investment by those pursuing the research goals the catalogue will become a key resource for communication within the discipline.
2. They can become very large—10^7 or 10^8 entries are not uncommon. This makes their computational support require careful engineering, to achieve availability, performance and reliability at an acceptable cost. Building on good platforms, such as high-quality database systems (see the two previous Sections 4.2.5 and 4.2.6) is therefore essential.
3. Whilst they are no longer a "*shopping list*" that can be browsed and self-managed, users still expect to interact with them. The predominant form of interaction is query access often embedded in workflows that then perform operations on all of the selected items.

The SkySurvey campaign was one of the triggers for the International Virtual Observatory Alliance (**IVOA**) formation, and certainly the stimulus that made Jim Gray propose the "*Fourth Paradigm*" [Gray 2007]. With ten years' experience it has valuable insights to offer on the design and use of catalogues [Raddick 2014 and 2014a] and [Budavari 2013]. As mentioned in the introduction, environmental and Earth sciences are more complex than astronomy, but the accumulated and published analysis of their workloads and user behaviours, will surely offer some benefits, and maybe some implementation strategies that are worth pursuing.

The ENVRIplus catalogue campaign will deliver mechanisms for holding the system-oriented and software-oriented aspects of an RI because the ICT experts know that they need this. The extent to which catalogues are built that handle domain-oriented items is less certain, but they are critical to the success of the environmental and Earth sciences. In some cases, such as LifeWatch[186] and ELIXIR, the maintenance and curation of information organised via catalogues is a primary role. Today, the underpinning platforms are probably independently chosen, engineered, maintained and operated in each context. For long-established catalogues and for very large communities that have the necessary resources, this is likely to continue. But for the many others, ENVRIplus working with other e-Infrastructure engineers should develop and deliver common solutions that are adopted. This would not only have economic and sustainability benefits; it will also facilitate cross-domain collaboration. It is clear from Section 3.4 that many are building on shared solutions, such as CKAN and drawing on core metadata

[186] Not a member of ENVRIplus.

standards. The extent to which such sharing and common standards choices is pervasive needs further investigation and consideration.

There are other considerations that may be addressed in the future. These are enumerated here in no particular order—many of them apply to other technology review topics as well.

1. **Cataloguing and other topics:** As discussed above and in the previous two sections (4.2.5 and 4.2.6), cataloguing has a key role to play providing the required technology for identification and citation, and for curation. It has several valuable roles to play in processing, for example:
    a. Finding the data (one or many items) on which the processes should be applied.
    b. Accumulating sets of results in derived data catalogues.
    c. Supporting efficient bulk operations on selected subsets of the catalogue.
    d. Providing users with prompts and lists of what operations, services, workflows, and tools are available that they can use to perform their processes.
    e. Cataloguing the formalisations of new and revised scientific methods that users may follow.
    f. Enumerating the structure, representation and interpretation of the ways in which data are or may be stored.
    g. Enumerating the co-workers and others who may have similar research interests or have skills they need in order to achieve a new process, subject of course to proper ethical and privacy rules.
    h. Listing the catalogues, what they contain and how they may be used.

    Similarly catalogues provide mechanisms that underpin the collection, search and use of provenance records. The optimisation mechanisms can mine information from catalogues, such as the numbers of items of a particular kind, and accumulate information in catalogues, such as data about previous runs, and previous mappings, in order to learn from these for future similar runs. In short, catalogues form a critical scaffolding both for the science and the technology of RIs.

2. **Roles for Cataloguing:** Catalogues are used consciously by most of the supported communities. They provide a conceptual framework for discussing what items are important, what properties of items are important, how these should be determined and what are acceptable quality controls. They then provide logistic support for assembling and using that information. As these are central roles, the community will refine their view of what should be catalogued and how it should be described. They, or their global standards, may also choose how the descriptive information is represented. Similarly, as described in the item above, the technological platform will draw heavily on catalogues. This will need to be shaped by efficiency and engineering concerns, e.g., lossless metadata compression [Arias 2013], as well as standards and consistency concerns. Advances will include the progressive addition of information to support automation, thereby reducing the time-consuming and error prone tasks that researchers and their support teams have to undertake.
3. **Raising the level of discourse:** The discussions about what should be catalogued, how those items should be described, i.e., the information content of metadata, and the operations to be supported should be defined precisely without recourse to properties of the underlying storage and software-platforms. This will deliver designs that stand the test of time, and that can deliver stability to practitioner communities. Re-engineering onto different platforms as the trade-offs change will then be feasible.
4. **Transient catalogues:** In most cases, the catalogues will be updated incrementally, but there may be reference catalogues that are rebuilt, in the manner of the sky surveys in IVOA regimes (See Section 1). There may be cases where scientists require the construction of a catalogue during a long running set of workflows or a research campaign. They then summarise or analyse its contents and discard it. The potential for and value of such transient catalogues may be investigated. Providing scientists with

tools as convenient as those which they have today on their personal machines to allow them to organise personal our group catalogues has been shown to be very helpful in medical imaging [Schuler 2014]. We suspect similar tools for environmental scientists would prove very beneficial for their research and its applications.

5. **Accommodating diversity:** There are two major subtopics here: diversity within a catalogue and diversity between catalogues. Some of the metadata within each catalogue will need to have a consistent and well-defined pattern, so that workflows and catalogue management tasks can be fully automated, and so that the fundamental modes of query and lookup work correctly. Section 3.4 identified four key groups of metadata standards that are already critical for the ENVRIplus catalogues. However, the metadata associated with each item may also accommodate user provided additions, relevant to work they are undertaking or testing the value of some new item attribute – this may be motivated by support for a new subdomain or observational system. This permits the innovation on which progress in science and its applications depends, as already explained in the previous sections. If an evaluated user-added attribute proves valuable, it can be promoted to a standard field in a later release of a catalogue. The range of types and representations of items for which catalogues may be produced is almost unbounded—it depends only on the ingenuity of researchers to recognise things to collect, list and process. Initially, a catalogue may only be used by those focused on the items it holds, but sooner or later it will be used with other catalogues, holding different items, described in different terms. This combined use will occur because the researchers posing such a multi-catalogue investigation have an understanding about how the two catalogues are related. If the relationship is based on some recognised properties, e.g., geo-location and time, then it is reasonable to expect that built-in transformations in the composite query system will select an appropriate set of candidate pairs, that may be tested for the relationship. If the relationship depends on a previously unrecognised features, e.g., blah in one catalogue and pling in the other, then ideally it should be possible for the user to supply a relationship test, e.g., related (blah, pling) and embed that in the composite query, so that the query machinery can more optimally return the subset. Several popular relationship tests will emerge and provoke catalogue redesign. Supporting such complex data access patterns via catalogues is one way in which they can help with boundary crossing between research communities or investigation viewpoints. The capability would need to be incrementally introduced.
6. **Cataloguing platforms:** It is evident, as described in this list item for Identification and Citation, and for Curation (arguments and references are presented in Sections 4.2.5 and 4.2.6), that the catalogue technology should be supported by a well-engineered and supported database system that delivers scalability, accessibility, recovery from failure, and supports multi-facetted range queries to select subsets and handle measurement error. Ideally, it should support a variety of models beyond relational, e.g., NoSQL such as MongoDB[187] or a scientific database (see Section 4.2.6), so that the freedom for users to explore extensions and the encodings communities use can at least also use XML and RDF. Scaling on nodes in one cluster is essential, but it is probably desirable for reliability to scale across clusters that are geographically distributed. This also enables queries and data accesses to be handled with less data transport costs and delays. Another expected attribute of the platform is that it handles time stamping of changes and can therefore run an old query at an old time, and retrieve the previous result. Helpful for curation, diagnostic investigations and partial re-runs.
7. **Temporal patterns:** There is a choice between building catalogues incrementally, which will probably happen in most cases, and in building them periodically. The latter strategy is particularly useful when the catalogue denotes all of the features or elements found in

[187] http://www.mongodb.org

a set of scanned resources. The periodicity can be based on the volatility of the sources, the costs of scanning them and the domain requirements for up-to-date data. A compromise is possible once composite queries are introduced; the main body can be refreshed periodically, to integrate new material, exploit improved information extraction algorithms and respond to request for changes in the catalogue design. A smaller and easily handled catalogue of recent observations can be maintained incrementally and reset to empty at the next periodic build. A composite query over the two catalogues then yields up-to-date information.

8. **Distribution patterns:** Modern environmental and Earth systems research is concerned with the whole globe and its interior. Individual research programmes and campaigns can combine information about almost any aspect of this complex system. Consequently, the domain scientists expect to be able to draw on data catalogued in many places and administered under different regimes (see Section 4.2.3). It will be very helpful when catalogue query systems have adopted sufficient standards and canonical representations that a single query can be automatically mapped to a distributed query set to the relevant set of catalogues. At each destination catalogue the canonical form, e.g., of parameter names and value ranges, are transformed into the local coordinate system and representations. The results from the multiple catalogues are then streamed back and assembled into a result of the form the users or software calling the query API expects. Almost certainly such developments will draw on more general-purpose, distributed query standards and protocols. The query system will typically be called from a science gateway or from workflows acting on behalf of users or data administrators. It is a moot point, how much of the data integration framework goes into the catalogue and query system and how much of it is encoded as data handling recipes (couched in a workflow language) that can be re-used in multiple scientific methods and management workflows.
9. **Jam tomorrow is not enough:** This is not so much of an issue for cataloguing, as catalogues are already widely used and proving their worth, i.e., catalogue users can afford jam. Two aspects of this issue may be considered:
    a. **B**y developing a common catalogue platform on reliable and affordable database platforms a wide range of catalogues needed for other data management parts, and a wide range of RIs' catalogues could be supported by common software. This will only happen if initial systems are available early enough that the other potential users don't feel they have to build their own. If they do, many opportunities for amortising costs, sharing understanding and making data accessible via consistent APIs will be lost.
    b. Another opportunity to help occurs when small groups or individual users want their own catalogues for a research programme, a project or an experiment. Offering an easily configured roll-your-own catalogue service and an easily downloadable catalogue platform (alternative routes via which they get help) and the associated training and support, would deliver what is required. Packaging this as a convenient service for individual researchers or a group to *easily* organise the data for their research has proved very popular for medical research data [Schuler 2014].

The key to cataloguing (and hence just about every other ICT aspect such as: provenance, curation, processing, identification and citation) is rich metadata with a canonical *'core'* and user-defined extensions. The metadata should come with matching and mapping specifications to other metadata *'standards'* and a set of convertors to permit the sort of homogeneous query over heterogeneous sources indicated above. Within the VRE4EIC project, such inter-conversion work is on-going between OIL-E from ENVRIPlus and CERIF from EPOS. There are already existing converters to/from CERIF with DC, DCAT, eGNS, ISO19115/INSPIRE and others.

4.2.8 Assessing the processing technology review

Processing, transforming, analysing and generating data, is a pervasive activity throughout the data lifecycle, that is often required at many stages and many iterations of scientific methods and their applications. It is already deeply embedded in the cultures and working practices of RIs, where it exhibits a great diversity: from time-critical and low powered quality monitoring and pattern detection close to data sources, through massive analyses to infer time-dependent behaviour over large regions with acceptable accuracy or simulation runs generating synthetic versions of a phenomenon's observable properties, to preparation of visualisations of significant results. Consequently, the activities referenced by "processing" are extensive, complex and often crucial parts of innovation and new achievements. The technologies concerned – see Section 3.5 – are themselves diverse, complex and critical to the missions of the RIs and their researchers. The multi-layer set of resources, from computational hardware and storage system platforms, through layers of software platforms that become progressively more specialised, to the means by which practitioners create, initiate, steer and manage computations. In most cases, the lowest layers are generic and standard equipment and software systems can be used. Such "standard systems" are greatly influenced by the commercial pressures, from entertainment, media and business, that dominate the ICT industry. There are a few cases, such as low-power sustained operation, HPC – see Section 4.2.2 – for running large simulations for non-localised interactions, and cross- correlations – derivations based on all-meets-all data comparisons[188] – where specialised hardware and provision is warranted. In most cases, common shared provision, using cloud or local cluster to amortise operations and management costs, is the appropriate platform.

Above these widespread and common layers, the layers of software systems incrementally shape the facility to match particular working practises and particular requirements. These include the programming languages and extensive widely used frameworks and method libraries that meet general or data-intensive requirements. These are often augmented by specialised libraries of functions required by each community, or by subgroups within those communities. Continuously running services for providing selective and transforming access to data, and to perform frequently required packaged functions also contribute processing power. Analytic tools, such as MatLab and R, scripting languages and workflows are used for composing these functions and services, to formalise and package repeatedly required processing combinations – virtually all scientific methods fall into this category as repeated runs are required during development and validation, and then repeated use is required to process each batch of data, e.g., data acquired during an observation period, or data acquired at each site, or data acquired for each region. Such formalisation, ultimately removes chores and opportunities for error. It enables experts from different sub-disciplines to refine parts of the method for which they have expertise, and it provides a framework for optimisation. These formalisations can soon become complex, sometimes involving millions of sub-stages. Hence they become difficult to work on even for experts. Tooling and diagnostic aids, often drawing on provenance records, are a great help. But tooling also needs to support the initial experiments—the first test of an idea about how to process some data. Consequently, tools or interfaces that enable the users to try out ideas using their own resources with minimum distracting technicalities are of paramount importance. Such development systems should keep careful provenance records to attribute credit, as many methods build on earlier methods; and as the provenance system – see Section 0– needs to be able to identify exactly which method was used. Fluent movement of the method formalisations between development and production contexts will reduce domain scientists' dependency on ICT experts, such as workflow and optimisation specialists, and thereby accelerate innovation and production. This will depend on fully automated selection of appropriate platforms and automated, optimised mappings from formalised methods to those platforms. Whereas the technologies for basic support of encoded methods in a number of scripting, programming and

[188] For example, the terracorrelator, http://www.wiki.ed.ac.uk/display/Terra/Terra-correlator+wiki, used to compare all pairs of seismic data stream and to compare observations with simulation results in the VERCE project [Atkinson 2015].

workflow languages is robust and ready for very demanding production use, e.g., workflows supported the recent discovery of gravity waves [Abbott 2016][189], the technologies to make the method development uncluttered by technical detail and to automate mapping exist for only a few notations and a few target platforms.

There is a strong mutual exclusivity between two modes of organising processing. In one, the user interacts, e.g., on their hand-held device or via a portal, to directly submit, control and monitor processing on their own resource or on a platform to which they have gained access. This may be through a problem-solving tool, through interactive programming, as in the iPython example given in Section 3.5, or through a portal providing some particular forms of analysis on some particular forms of data. These will often behind the scenes draw on the same repertoire for defining methods as we described above. This mode is appropriate for learning about systems, for testing and developing ideas, and when only modest repetition is required. In the second, the user, an event detector, or a scheduled time, initiates the request for processing, which is then submitted to a queuing and resource allocation system, and it then runs on the target platform. The time between initiation and response can vary from under a second, to days or even weeks, for very demanding jobs. Helping users monitor and organise such processing, particularly when they have many related requests in a research or derivative data generation campaign, is an essential element in addressing the scale of modern data. This links closely with the provenance system – see Section 0– driving the tools from provenance records, and delivering to the provenance all the information needed for its records. This also meets another issue in handling massive data volumes, after a partial failure, it automatically enables parts of work completed to be retained, and after clean up a restart to complete a complex method. Such issues will become important in RIs as they scale up and as their methods become more demanding.

There are further considerations that may need investigation. They are fewer in some ways, as processing is perhaps the best supported and most understood part of e-Infrastructure. However, it is good in parts, for example the frameworks and tools for building new data-intensive methods still demand deep understanding of technological issues that should be automated for reliability and to reduce chores. Some potential topics for further consideration are enumerated here in no particular order—many of them apply to other technology review topics as well.

1. **Processing and other topics:** Almost every other technology topic depends on processing. A few examples follow:
    a. Identity minting[190] – Section 3.2 – will need processing during minting to verify quality of metadata, communicate with reference sites and compute fixity data.
    b. Curation – Section 3.3 – similarly processes metadata to verify consistency. It also runs processes to make copies for preservation and to transform between media and contexts for longevity.
    c. Cataloguing – Section 3.4 – requires processing to populate catalogues from other data, to create automated metadata, to perform matches during queries, to assemble responses translating to target representations if necessary, and to preserve contents by mapping them to new forms as new versions and representations are adopted.
    d. Optimisation – Section 3.7 – requires processing to mine performance parameters from previous runs, to detect "seen before" requests, to analyse new requests and generate plans, to compute the cost of candidate plans, and to map the optimal plan to target platforms.

[189] https://pegasus.isi.edu/2016/02/11/pegasus-powers-ligo-gravitational-waves-detection-analysis/
[190] The word "minting" here is a metaphor from the process of making new coins, i.e., minting them in compliance with the rules that make them retain their value. Here refers to making a new identity with rules to ensure it retains its value of uniqueness and referring to the original entity.

e. Provenance – Section 0 – needs processing to transform incoming information from many platforms and subsystems into its standard form and to generate responses to requests for subsets of its records potentially in summary form. It may also transform those results into other standard representations.

2. **Roles for Processing:** It is called in regular patterns, e.g., every 10 seconds or every day, to support all of the routine operations of data management, such as QA, data shipment, data compression, etc. often as part of an automated and autonomic system. It is called to execute all of the steps in established scientific methods. These may be submitted directly or as a consequence of users interacting with a portal or analytic tools. It is called on an *ad hoc* basis as a user tests an idea, performs some quick transformation, generates a presentational form, or does a management task on their workspace data.
3. **Raising the level of discourse:** The ways in which users and communities discuss their processing are often mature and deeply embedded in their culture through training and practice. For example, one community may invariably use Python, another MatLab, another R, and another graphical representations of workflows. Most users will choose data and supply parameters to packaged methods to do their routine work. When they need to craft a specific process, most users will mainly compose and parameterise pre-existing functions and services. Specialists and innovators need access to all of the details. However, the framework that supports them and delivers processing can be considered without recourse to these details, in more abstract terms such as: create encoded method, reuse encoded method, parameterise encoded method, run parameterised encoded method, etc. Articulating provisions at this level may make commonalities recognisable and highlight particular cases that need special treatment.
4. **Transient processing:** Opportunities and facilities that aid creative thinking are very valuable in science. Often, at the start of an idea or invention of a new process, a user will want to experiment to test and explore their idea. Processing must support such computational experiments well. It should avoid distractions and impediments from heavy-weight machinery and administrative procedures. A fluent path from experiment to operational practice is desirable. For innovation to flourish that path should not depend on assembling a team of gurus. For sustainable science expert-guided optimisation as a necessary step on the path to production should become rare as automated mappings and data-intensive engineering improve.
5. **Accommodating diversity:** As described above, there are a great many programming languages and other method encoding technologies in use. These all need to be supported to avoid disruptive impacts. This is usually well done by systems for the actual execution of the methods. However, today there are significant issues to be overcome:
    a. Modern data-driven science depends on re-using and refining methods already developed. In many cases, boundary-crossing needs input into these methods from different cultures using different notations. Where those can be viewed as programs, today's workflow systems compose them. Where they are scripts, e.g., in R or MatLab, or workflows in different languages, very little help of production quality is available – SciBus is a good example [Kacsuk 2014].
    b. Developing such methods often requires good IDEs to organise all of the software engineering, and then it requires deployment systems that establish and run the new software in the target contexts. Neither of these stages, crucial to the development of RIs' e-Infrastructure, and essential for scientific innovation are well supported across boundaries.
6. **Processing platforms:** The lowest levels of the processing platform are dealt with in Section 3.11. These provide the computing cycles and storage, the networks, the operating systems, and in most cases the containment for processes. They also handle resource scheduling, computer=computer and process-process interactions and provide security. For most data-intensive or numerically-intensive processing there are other layers that he RI e-Infrastructure should exploit. For example, there are data-intensive

frameworks good for high and sustained throughput, such as Apache Storm and Apache Spark[191]. Similarly, MPI is a standard underpinning exploited by many, but not all, large scale simulations. Choosing a shared repertoire of these middleware frameworks that facilitate demanding computation, that handle reliability and availability and that are maintained (and often run) by mature global communities of engineers is a crucial decision for the RI e-Infrastructure builders. Whether suitable database technology that combines computation with data access in the query context, and with data delivery in the projection of results context, should be included in this platform is an open question. Sooner or later significant numbers of RIs will need such a platform – see Section 4.2.7.

7. **Temporal patterns:** What proportion of the workload can be decoupled from human activity and run when systems are lightly loaded? What proportion is time critical, e.g., for hazardous event recognition [Earle 2009]? What proportion of the processing needs human-interaction, e.g., for steering and development, and what proportion can be run in batch mode, some of which may still yield rapid responses?
8. **Distribution patterns:** How much of the envisaged processing needs to be done at specific sites for ownership or management reasons? How much of it will benefit from being co-located with the data it operates on to avoid bulk data movement, which consumes both time and energy? How much of the work can be run almost anywhere, and how much needs specialised platforms? How many scientific methods will there be that need to combine processing on two or more different kinds of specialist platforms? To what extent will the processing benefit from parallelisation across nodes within a cluster and between clusters? To what extent can the distribution of processing be automated and optimised?
9. **Jam tomorrow is not enough:** Virtually all practitioners in this context are adept at using computers. They all already enjoy reasonable access to platforms where they can run small scale programs on local data. Consequently, this is not as much of a pressing issue as it is for other technologies. However, the moment users move out of this comfort zone, e.g., preparing and encoding complex data-driven methods, running against remotely held or large-scale data, requiring substantial computational resources they get very little help. We should be building intellectual ramps and tools that support them for each of the directions in which a researcher can develop their processing needs[192]. In addition, providing effective interfaces for external tools, on hand-held devices and a range of interactive platforms would be helpful. For example, Sencha[193] used in the VERCE project [Atkinson 2015] or the Django[194] framework to provide an API to the data with angularJS[195] and semantic UI[196] to build the interactive framework; although ember[197] may be an improvement on angularJS [Taylor 2016].

4.2.9 Assessing the provenance technology review

For modern data-driven science there is a pressing need to capture and exploit good provenance data as explained eloquently in Section 0. Provenance, the records about how data was collected, derived or generated, is crucial for validating and improving scientific methods. It enables convenient and accurate replay or re-investigation. It provides the necessary underpinning when results are presented for judging the extent to which they should influence

[191] Apache Storm http://storm.apache.org/ and Apache Spark http://spark.apache.org/.
[192] For example, if the user wants to assemble a large number of observations and then run a series of large simulations on a PRACE facility, they should benefit from a framework that organises the movement of the large collection of inputs they have identified to the PRACE site at the right time and organises the return of results and provenance records automatically.
[193] https://www.sencha.com/customers/
[194] https://www.djangoproject.com/
[195] https://angularjs.org/
[196] http://semantic-ui.com/
[197] http://emberjs.com/

decisions, such as mitigating a natural hazard[198] to publishing a paper. It provides a foundation for many activities, such as: attributing credit to individuals and organisations, providing input to diagnostic investigations, providing records to assist with management and optimisation and preparing for curation. The RIs will need to perform these functions and consequently the e-Infrastructures they depend on will need to support provenance collection and use well. The interaction with *identification and citation*, and with *cataloguing* is made explicit.

Today, it is challenging to plan and deliver an implementation which is sustainable, i.e., sufficiently shared or dependant on a common widely supported platform – see Section 4.2.4, and which copes with the multiplicity of services and platforms, that typically do not adopt a common standard for provenance when they support provenance at all. The Section 0 provides:

1. A very clear list of the groups, standards bodies and research campaigns that are addressing provenance as it has been recognised by international bodies such as W3C as crucial as society depends more and more on information derived from data via long paths of inference, interpretation, filtering, transformation and integration.
2. An introduction and evaluation of the current candidates for shared approaches.

Though the analysis was deep and the coverage broad in Section 0 there are still opportunities to consider some issues further. These are enumerated here in no particular order—many of them apply to other technology review topics as well.

1. **Provenance and other topics**: taking the topics in the order they were presented in Section 3, we can illustrate significant interactions in the following way. *Identification and citation* are prerequisites of *provenance*. Identifying the artefacts, individuals, organisations and computational contexts may all be necessary for full provenance records. *Curation* will depend on *provenance* systems to provide automatically relevant metadata; otherwise, unrealistic amounts of human effort would be required. *Cataloguing* will almost certainly underpin the *provenance* system, as provenance requires searchable catalogues of its records. *Processing* has to support provenance collection, e.g., supply from all its services, platforms and workflows provenance information in a form that can be readily integrated. *Processing* may also use provenance records to resume after partial failures with a clean-up of partially complete (intermediate) results and re-use rather than re-computation of those that have been completed – a necessary provision as data scales and activity rates increase. The provenance records can be a great help to practitioners as they prepare and refine the encodings of their methods. Providing records to help with diagnosis and to easily organise exact re-runs after a putative repair. Providing records that can be analysed to review progress with a research campaign and to assess the coverage of a required field of evidence. Providing records that can be analysed to discover usage trends and to facilitate resource planning or to spot critical targets for *optimisation*.
2. **Roles for Provenance**: These have been covered under the previous heading to a great extent. However, traditional usage tends to focus on retrospective analysis and preparation for curation. As Spinuso has demonstrated [Spinuso 2016], and as other work reported in Section 0 shows, it is possible to use the provenance system actively as work progresses. This encourages uptake as it improves the productivity of the practitioners.
3. **Raising the level of discourse**: The facilities provided by the provenance system, and its requirements on other systems, such as processing, cataloguing, identification and curation, can be articulated at high level, mainly from the information viewpoint, by using the vocabulary and structures of the reference model – see Section 3.10. This will help clarify the requirements on other systems to supply information, and what are the temporal constraints, e.g., must a system support live provenance record transfers, so

[198] See for example the L'Aquila *post hoc* analysis, where lives were lost because building regulations had not taken into account seismic hazard maps, but the first court proceedings attributed blame to seismologists [Cartlidge 2012].

that user steering can be delivered? What functions that the provenance system can be expected to support, e.g., what categories of multi-faceted query should the provenance system support? Or what operations are provided to switch provenance recording on and off during processing? What mechanisms, if any, allow users to add their metadata to a provenance record? What kinds of provenance summaries should the provenance system support? What are the privacy versus sharing rules for provenance data and how are they specified?

4. **Transient provenance records**: At least an agreed synopsis of provenance records will be stored with provenance records. But what about the provenance records of intermediate data and of products generated during testing, refinement and innovation before production runs? Not having provenance on at such times is unwise, as it can support many forms of investigation helpful to the development team. It may be part of the evidence that a system is safe and of sufficient quality to deploy in production. However, indefinite storage and storage validating an encoding of a method that is now superseded by another validated method may be wasteful. How are provenance record lifetimes determined? What mechanisms should there be for users to affect the lifetimes of stored provenance records relating to them?
5. **Accommodating diversity**: As Section 0 identifies the current variations in standards for representing and working with provenance, make it very difficult to establish a single provenance system. The diversity of behaviours of subsystems, platforms and processing systems, such as workflows and simulation models, makes the provision of a uniform provenance system difficult. The users and RIs may also have different policies and practices that the common system needs to accommodate. To what extent can this be done with automated translation intermediaries?
6. **Provenance platforms**: It is difficult to identify a single provenance system that covers the whole data lifecycle and supports all of the provenance uses; Section 0 reviews five candidate systems that are developing, but none is yet ready for production use. It may be the case that standard representations and established workflows will need to be gathered around a supported subsystem, such as cataloguing that meets part of the task. Spinuso has used MongoDB to store PROV-compliant records for a range of provenance gathering and usage modes [Spinuso 2016], so building on a database platform is another approach. For sustainability – Section 4.2.4 – the approach eventually selected should either be a widely adopted supported system (none of which exist at present) or one supported by a suitable alliance as there many others needing such systems. It is possible that through RDA, or similar bodies, or through multi-community coordination alliances for environmental science, a consortium can emerge not only to agree the metadata and ontologies that are needed in environmental and Earth sciences, but also collaborate on tailoring a good platform to support those provenance data.
7. **Temporal patterns**: The consideration of whether provenance records are available during the run has been raised above. That can support monitoring and diagnostic tools and be used to trigger data management operations, such as collecting results. Some processing platforms prohibit such communication. If provenance data is fine grained its transport and storage may be a significant overhead. In which case summary records need to be derived sufficient for later purposes before discard of the details. User and community policies over the trade-off between complete details and overheads will need to be taken into account.
8. **Distribution patterns**: In most contexts, the data-intensive federation – Section 4.2.3 – will be using distributed data resources, which will hold their own provenance records. The provenance records for data derivatives will probably held at the site where the derivation was performed or at the site which initiated the derivation. That site will need to reference the provenance of the input data used as input from that derivation record. A tree of references to remote records may be the result. This may be even more complex if the derivation was performed by a workflow distributed across platforms. As

network traffic is expensive it may be best to postpone provenance record shipment to assemble the tree or chain of provenance records until they are needed. This may be achieved by a distributed query or a provenance workflow that runs in the background marshalling provenance data into collated sets. The users and resource providers will resist adoption if provenance gathering imposes a significant overhead – this may be most sensitive in distributed and real-time contexts. Yet it is precisely these contexts where proper attribution of responsibility and clarification of the evidential chain may become important. Without it, improving the quality of services and the information they offer decision makers may depend on guesswork.

9. **Jam tomorrow is not enough**: Section 0 already draws attention to the work, presented at our kick-off IT4RIs workshop, by Myers *et al.* encouraging the adoption of good metadata practices during the research processes [Myers 2015]. This clearly applies to the provenance element of that work, and early provision for ENVRIplus RIs should attempt to emulate their success. Some of the tools developed to use provenance records have excellent visible summaries of the information, e.g., those in [Spinuso 2016]. If possible, early work in ENVRIplus could develop exemplars that are easily understood and visually powerful, to use in outreach and training. This will help lower barriers to uptake and gain engagement. Fear of unacceptable overheads or demands on researchers' time must be addressed by showing the scale of the actual overheads and providing controls that allow researcher communities to switch them off when necessary. We hope of course they will not do that, but they do need this safety net!

4.2.10 Assessing the optimisation technology review

Optimisation, reviewed in Section 3.7 is important for every aspect of the e-Infrastructures and working practices ENVRIplus sets out to support. Making best use of people's time and minimising energy consumption are probably the most important goals – certainly long term. The shorter-term goals need sharpening with explicit cost functions, for example, clarifying the productivity of which roles need highest priority at a given period, the results that are time critical, and the cases where throughput is the highest priority, in each case, within the constraints of an energy or funding budget. These cost functions cannot be narrowly defined, e.g., we have seen that identification and citation, curation, cataloguing and provenance are deeply interconnected and they all depend on processing. Consequently, improvements in identification and citation that simply pass the costs on to cataloguing and curation would not have the intended value. There is therefore an argument for **making optimisation a cross-cutting concern**.

Two aspects of optimisation in large and sustained systems need a well-managed structure:

1. **Temporal partitions of decisions**: The awareness that an optimisation is needed may wait until a community or method is encountering problems. While this trigger may be a useful indicator of where to focus and what the priorities should in the short term, it will be difficult and expensive to address the issue if the required mechanisms are not already embedded in the technologies in use. Consequently, at the early stages of building or commissioning e-Infrastructures or productising new scientific methods, the capability to subsequently optimise must be evaluated. Simply throwing kit at the problem, a strategy often suggested by resource (e.g., Cloud) or technology (e.g., computational cluster) vendors, may prove unaffordable in the longer term. In the early years of ENVRIplus, the optimisation WP should be drawing attention to the selection of technologies which support optimisation well; many of which are introduced in Section 3.7 – it would be unwise to postpone this internal awareness raising, though most of the RIs, and all of the topic leaders are aware of it.
2. **Organisational partitioning**: As explained in Section 4.2.4, sustaining long-term support for research, for working practices and for he data lifecycle is best achieved by delegating to others substantial parts of the technology, particularly software, design

and maintenance. Alliances be forged to share the residual responsibility for a minimised list of software elements (and here their supporting hardware) that is particular to and essential for an RI's or group of Ris' research. By carefully choosing suitable software platforms, e.g., Apache Spark, and software systems, e.g., Mongo DB, three gains are made:

a. An **appropriate community of experts** is driving the design, construction and maintenance of those projects – they can gather and harness experts who understand very particular engineering and optimisation issues and hence optimise their subsystem far better than a self-build approach could ever hope to achieve. They also get pre-release access to new technologies, e.g., 3D Xpoint[199], and hence can deliver its very significant advantages more quickly than local teams.
b. A **wide and active community** of demanding users, e.g., in Apache Spark's case large companies, explores the functions and capabilities. The result is that latent errors that have evaded the release quality controls are likely to be encountered first by someone else. Similarly, capability issues, such as rates of data handling, are also likely to affect someone else first. Consequently, many potential problems are dealt with without any impact on your community.
c. As a result, your responsibilities are reduced and the technology supplied already has the required support for the specific optimisations that you need to do.

At present most of the optimisation considerations remain to be addressed in the future. These are enumerated here in no particular order—giving some more technical details about potential structure as we go.

1. **Optimisation and other technology topics**: As explained above, optimisation should really be a cross-cutting issue as all aspects of an e-Infrastructure and all uses of it can be subject to optimisation. For example, energy saving is normally a pervasive issue, a local reduction that increases energy consumption elsewhere would be valueless. The exception to this is remote field or marine deployment, where there are no local energy sources.
2. **Roles for optimisation**: Optimisation can *deliver improved services to researchers*, e.g., deliver data derivatives and analyses more promptly or deliver more data analyses with their budget or available resources. Note that these two optimisations are usually in conflict and the trade-off between them needs to be captured in a cost function that can be calculated, can be measured and therefore, can be minimised. As in many branches of engineering, measurement of progress towards a goal is key, as the complexity of the hardware and software systems defy modelling. But optimisation can also deliver productivity improvements for any of the roles concerned with setting up, deploying, running and managing the e-Infrastructure. For example, it can deliver significant labour savings to those forming and deploying the virtualised computing contexts that the operational system requires – see item 9 below.
3. **Raising the level of discourse**: The discussion about what potentially needs optimisation can be addressed in terms of higher-level concepts articulated in terms of the reference model – see Section 3.10. Similarly, an abstract characterisation of potential target cost functions to be minimised will facilitate discussion of their relative importance. Aspects such as the relative importance of energy consumption, response time and time to first operational deployment can then be discussed by those steering the RI and communicated clearly to those building and operating it. A clutter of technical details would not only inhibit discussion and obscure communication, it would also be fragile—as the digital context and scientific methods changed it would need to be revisited.

[199] See the EU project NextGenIO site for developments of its applications for science, http://www.nextgenio.eu.

4. **Transient optimisation:** Generally speaking, optimisation is a long-term investment with an even longer term payoff. However, there are occasions when there is a need to optimise a cost function that downplays other factors. For example, during a hands-on training session or a project review demonstration, response time may override balancing issues that normally apply, such as energy consumption, local resource costs, and other community member's throughput. Similarly, with a pressing submission deadline, the throughput to gain statistically significant evidence may have the highest premium.
5. **Accommodating diversity:** As indicated above, almost everyone, in all of the roles in Table 19, may make a case that if some aspect of their working practice is not improved the entire RI, e-Infrastructure or community will suffer, or they may collapse under the strains of their current tools, services and workarounds. As the RI progresses and its users become more expert, as data collection accelerates or gains resolution, as data analyses and models grow in complexity, as derived data and result production increases, as demand for data access from external data-intensive federation partners rises, or as curation volumes hit a maximum, anyone can identify and describe a pressing need. The software engineers will be overwhelmed by concurrent requests and unable to fulfil them all. It is vital that the governance system accepts these request for optimisation, and makes decisions about how they should be prioritised. They should consult the software engineers to assess the amounts of effort involved and the scope for improvement with the current platform and digital context. After an optimisation target has been set, the software engineers need to propagate appropriate shares of responsibility to those who are supplying and supporting subsystems—this is why they have the customer support mentioned in Section 4.2.4.
6. **Optimisation platforms**: As explained above, the underpinning software layers, platforms and subsystems will normally have significant capabilities for optimisation—this capability should be a significant criterion in their selection. But care has to be taken to understand the *interactions between layers*. For example, if Apache Spark is selected, then it does not distribute work over multiple clusters as many workflow systems do, e.g., Pegasus [Deelman 2015]. Therefore, it is limited to the capacity of the cluster on which it is running when work is submitted. There are further limitations that can limit its capacity on particular platforms; for example, it grows its RAM occupancies indefinitely if the analytic algorithm requires it. That leads to pathological page swapping. However, combining frameworks for building data-intensive applications with facilities to run them across heterogeneous platforms may offer a solution.

   Many platforms employ virtualisation to allow a predetermined computational context to be deployed and to deliver protection. There are two common forms, that which depends on a hypervisor and that which depends on a Linux Containers (LXC). The former intercepts all machine operations and maps them from the controlled image space into the external digital context, if they are legitimate. The LXC intercepts only all of the calls to the operating system, which is the way programs communicate with their external environment. The hypervisor mechanism to allow safe cohabitation underpins Cloud systems. The LXC mechanism underpins complex large-scale operations and was pioneered by Google. A popular version is Docker[200]. In our recent measurements the overhead for *data-intensive* workloads was much lower under LXC than in a cloud (AWS) context [Filgueira 2016].

   The choice of data storage and management systems, e.g., file systems, parallel file systems, e.g., HDF5[201], and databases (see Discussion in Sections 4.2.5 and 4.2.7 on database options, including scientific databases and NoSQL such as Mongo DB) is also a

[200] Docker: https://www.docker.com
[201] https://www.hdfgroup.org/HDF5/

critical issue for some classes of optimisation. Some of the platforms have pre-packaged choices – see Section 3.11.3.2.

7. **Temporal patterns**: Many optimisations are sensitively coupled to details of the underlying deployed platforms and the encoding of scientific methods, both of which change frequently. As indicated above there is usually concurrent demands for urgent rescue packages. Traditionally they were handcrafted by a performance engineer. As Section 3.7 points out, this is not sustainable if the experts also have to revisit previous optimisations when they have been invalidated by changes in context or encoded task. Automation is required that at the very least handles almost all of the changes and automatically revises the optimised mappings from required task to digital implementation. This self-same automation should perform much of the initial optimisation, but that requires (a) a formalisation of the target cost function, (b) sufficient metadata about the digital environment and about the task, obtained from a mixture of human input and automated analysis of task encodings and prior run provenance records. Thus the framework supporting optimisation and the information it draws on should improve with time. As remarked above, improvement will only be maintained if there are sufficiently precise measurements available.
8. **Distribution patterns**: After many years of experience at NASA, Mattmann identified "*Intelligent data movement*" as one of the four key factors in making data-driven science using big data a success [Mattmann 2014]. This applies at all scales, from the activities within a node and a cluster, to the location and movement of all data between sites. Optimal strategies for avoiding double handling, putting data where you will need it. Tactics to avoid going to backing store between subtasks in an encoded scientific or data handling method unless the intermediate is needed for diagnostic purposes [Filgueira 2016]. Tactics to use the right kind of storage when reading or writing to disc stores [Koltsidas 2008]. The new storage technology, 3D Xpoint, with significantly different properties is rapidly being brought into the storage framework. There is a deep tradition of developing a variety of hierarchical storage systems, from magnetic tape, to the solid-state disks. In some cases, the storage is a separate network-attached unit, coupled to one or more clusters on the same site. In other cases, the storage is distributed and associated with every computational node. The former may provide a good environment for curation, but the latter may be much faster for high-throughput data analytics. Thus the optimal strategy for data movement, data distribution, replication and preservation is complex when everything is located at one site. Modelling and analysing the options here is worthwhile. Within one platform or subsystem layer this will be done by the contributing organisation responsible for the provided platform. However, two issues still need attention:
    a. The provided platform will need configuring, e.g., informing it which data should be persistent and which should be active, which will be written and accessed by bulk serial transfers, and which will require random reads and writes.
    b. The interaction between the provided layers may be handled by the layer providers. On many occasions it will be the responsibility of the RI team and ENVRIplus.

   The management of distribution and data traffic becomes much more complex when we consider geographically dispersed sites, particularly if this is in the context of a data-intensive federation – see Section 4.2.3. Here the police and rules should ideally take into account the feasibility of optimising data placement and data transport. Different sites may have very different provisions and operational regimes, and so be well suited to particular aspects of the workloads. Different communication routes may have very different properties. Here there is an acute need for good descriptions of sites and networks so that as much as possible of the decisions can be automated. The rules and policy should be assessed against models of the anticipated inter-site and inter-organisation data traffic. Measurement is essential to monitor costs and to measure

progress. The provided e-Infrastructures that underpin this and bundle provision to many consumers, e.g., GÉANT – see Section 3.11.3.1, may appear to be free at the point of use. However, any attempt to organise data placement, data traffic and inter-organisation contracts, in order to minimise energy costs, will need to properly account for the energy involved in data movement, all the way along the path.

9. **Jam tomorrow is not enough**: Optimisation might not sound like an opportunity for **jam today**. However, if we analyse what are the pressing activities in the RIs at present, we may find they split into two groups:
    a. **The established group,** who may have immediate needs for optimisation. In this case choosing standards for the descriptions of some aspects of their e-Infrastructure and then providing help with collecting and setting up those descriptions might be a first step. By choosing a critical niche, e.g., data placement across sites, it may be possible to work with a relatively small amount of descriptive data. Then installing a system that helped place data optimally against some agreed cost function might yield rapid results, that are noticeably helpful.
    b. **The setting-up group**, who may at present be getting focussed on building and deploying the necessary platform layers and subsystems, and providing the computational contexts their community wants. By introducing tools for setting up a library of images, and of sharing them and deploying them easily, we would achieve at least three benefits. The library would encourage sharing. The scripted image build management would accelerate refinement and adaption. The hard pressed teams building and deploying the proto-production systems would have enhanced productivity. Examples are given in [Fox 2016] and the Pegasus team uses Docker scripts to construct and deploy images: the images are automatically built and stored in their docker hub repository[202]. So every time they upload a docker file, the image is automatically built and stored in the docker hub repository. These are then deployed ready for use using docker-compose files. Building up libraries relevant to commonly required deployments would be an advantage. Haydel *et al.* present examples where docker has accelerated their use of specialised hardware architectures [Haydel 2016].

4.2.11 Assessing the architectural approaches review

Architecture, the first of our cross-cutting themes, is reviewed in Section 0. Just as with the architecture of buildings, architecture for e-Infrastructures underpinning RIs, concerns balancing the complex and often competing pressures within a feasible budget (of software and systems engineering effort) in an acceptable time:

- Meeting the needs of the research communities and encouraging their expansion.
- Making continuous operation and maintenance of the data lifecycle feasible, sustainable and affordable.
- Using tried and tested designs wherever possible.
- Fitting into these innovations where necessary for function not for fashion.
- Using familiar and well-understood presentations of the expected and standard facilities.
- Ensuring that each aspect of engineering is properly analysed and taken into account.
- Using previously successful design patterns and implementation methods where possible.
- Exploiting as much pre-fabrication as can be effectively composed into a coherent whole.

[202] https://docs.docker.com/docker-hub/repos/

Again, as in the architecture of buildings, we do not have the luxury of a 'greenfield site'. The new facilities need to fit with, maybe extend and partly replace, existing investments so that they can be adopted with enthusiasm by the existing inhabitants, a complex community – see Section 4.2.1. They also need to fit with the surrounding context, which in most cases requires a complex network of agreements and operational interconnections – see Section 4.2.3. And finally, the construction needs 'planning permission' and to comply with 'building regulations'. For the former, we must include the relevant standards that already apply in the field, a significant bundle of which are incorporated in the INSPIRE directive [EU Parliament 2007] and many others apply. For building regulations, many are potentially in play, for example the rules concerning the use of nationally and regionally funded facilities, e.g., PRACE-centres' rules, the rules governing shared e-Infrastructure platforms, e.g., those imposed by GÉANT and cloud providers – see Section 3.11– and the EU H2020-backed EOSC – see Section 3.10.3. Once a building or e-Infrastructure has been constructed it needs maintenance to continue to serve its user community including adapting to their new requirements, but in the case of e-Infrastructure this is more difficult as the digital context is changing rapidly due to uncontrollable commercial and economic forces – see Section 4.2.4.

Because of these complexities, carefully considering architectural issues is critical. Astute design of the architecture can much improve matters:

- For the researchers and other data users, as they encounter a more comprehensible and coherent system that they find much easier to use.
- For the data providers and those managing and running the data-handling chains in the data lifecycle, who find that all aspects of their digital tools and services, fit together, meet accepted standards and can be trusted to have sustainability.
- For those setting up the e-Infrastructure they have well-identified pre-designed subsystems to import and guidance on how to fit them together with minimum extra construction and maintenance effort.
- For those providing funds and resources for the e-Infrastructure, they know that the maximum overlap with systems they already resource, and between systems that being introduced has been carefully considered.

The complexities of system architectures need good media for their discussion, recording and analysis. For building it used to be drawings and models; today it is predominantly computer-aided architectural designs (CAAD), with accompanying methods for generating drawings and models, for analysing engineering and regulatory requirements, and for feeding into construction planning and management. The first steps of an equivalent approach for large-scale distributed and multi-organisational computing systems have been built around the Open Distributed Processing (ODP) standard, which is used in ENVRIplus to represent and develop the Reference Model (RM) – see Section 3.10. This helps system designers and builders use a vocabulary and representation that can be interpreted unequivocally. It also helps address the complexity by establishing five viewpoints from which issues can be examined. Unfortunately, although well developed for human experts, it is not yet so well connected with simulation and evaluation tools, or automated coupling to construction planning and execution.

In summary, and effective architecture needs to meet the following principles in a way that can be communicated to all relevant parties:

1. **Acceptability**: It must be such that a deployed version will meet regulatory, ethical and political concerns necessary for approval, but also sufficient to satisfy broader public scrutiny. For example, if policy decisions draw on results from environmental and Earth science Ris, it must be possible for those who wish to question those policies to review and understand how the data-driven evidence was produced. At quite a different level, those working in the field need privacy protection, otherwise miscreants know they are away from home or where to find them to take them hostage or steal their equipment.

2. **Usability**: It must be feasible, where necessary, for researchers to develop experiments using familiar methods and workflows, subject to governance constraints, i.e., rules agreed by their community or federations in which their community has engaged, such as, maximum periods before which results are made shareable. If their methodological innovations prove valuable there should be a well-supported path to bring the new methods into the wider production repertoire.
3. **Sustainability**: As far as possible the software used must have either known support, e.g., from vendors, or a known *active* open source community. Where it is appropriate as part of the architecture to contribute to open source software then it is necessary to budget for contributing to that community and be sure there is a sufficient (global) community behind the open source that it can be anticipated to continue. Where it is appropriate to lead the development of architecture-specific software, this should be developed with the Software Sustainability Institute's model of sustainable software. The underpinning infrastructure, storage, digital communication and computational resources must also have sustainability that will meet the community needs, i.e., *for decades*.
4. **Flexibility**: The functionality of the APIs offered by the core or initial e-Infrastructure to other software should allow a wide range of future uses, that are as yet unpredictable, and should as far as practicable allow the architecture to accommodate much larger data volumes or more sophisticated models than initially encountered.
5. **Diversity**: The architecture should be capable of evolving to support the many varieties of data formats and analytic methods currently in use and frequently changing. Practitioners should be able to introduce, subject to an RI's policy, new forms of data, new structures interrelating data and new tool sets for performing operations over data.
6. **Scalability**: Although early instances of the architecture will rely primarily on human trust and associated manual operation of experiment workflow, as the volume of data and experimental demand increases it should be capable of adapting to include automation of workflow where this is consistent with other architectural and governance principles. For most Ris, such automation is already in place but it may not yet cover sufficient aspects of the data lifecycle and a sufficient range of research campaigns, particularly those that are interdisciplinary.
7. **Validity**: The management and governance of the implemented architecture should ensure that potential modes of failure and misuse will be progressively identified and enumerated. As each is identified, and prevented by combinations of rules and software, tests for the subsequent releases of the architecture should be introduced and overseen in a clearly defined governance structure that includes independent oversight.

The short-term requirements, well enumerated in Section 0 need to be addressed while taking these longer-term issues and principles into consideration. Candidate architectural strategies considered there include:

1. The use of the UML formalisations of the ODP reference model to separate concerns.
2. The use of a central core that supports portals and other uses.
3. The use of Model-Driven Development.

These are not mutually exclusive. For example, the central core, will almost certainly emerge or be a requirement, whichever way the system is modelled and however the construction is coordinated.

With each approach the challenge is to make best use of existing investments in Ris and to consider the many details needed in each context. The current investment in the ODP reference model taken further to develop the vocabulary and dialogue for dealing with the implementation details via the Engineering and Technology viewpoints is probably the best path – see Section 3.10. This will almost certainly develop an onion like structure, with a common core delivering the requirements that every RI requires and then layers that tailor it to meet the more specific requirements. Ultimately, the common core will probably meet most aspects of data-intensive

federations – see Section 4.2.4. However, initially it will almost certainly be simpler and less specific. For sustainability, inter-disciplinary harmonisation and for amortising costs – see Section 4.2.4, it will therefore draw heavily on the frameworks and models supported by resource providers – see Section 3.11 and EOSC.

Thus further issues that should be considered by the ENVRIplus community include:

1. The extent to which the **architecture should be shaped** by the provided services of resource providers versus the extent to which it needs to be crafted to meet a class of RI requirements, or a class of environmental requirements.
2. The **notation** used for formalising and agreeing architectural construction and maintenance contracts. Augmented and developed versions of the ODP reference model are one candidate. Critical factors will be the choices made by international alliances such as the RDA, the extent to which affordable tools and training become available, and the extent to which federated efforts commit to developing and refining the detail.
3. The **common functionality** which can be amortised across many Ris because they agree that they all want a kernel implementation of a functionality, e.g., scalable multi-site catalogues.
4. **Adaptation to data-intensive** aspects of data-driven science. The widely deployed platforms are well tuned to the ratios of data handling and computation cycles that underpin the dominant commercial fields and the sciences when simulation and small-scale data analyses was the only game in town. Today, our encoded scientific methods include substantial portions which are data intensive, i.e., they perform data input and output and data transfers to such an extent that data handling becomes a severely limiting factor. Various architectures and subsystems are emerging that can address this, but there is no one-size-fits-all solution. Consequently, how the e-Infrastructure best incorporates these contemporaneous advances is an open question.
5. **Managing software automatically** is an essential for sustainability and for large scale distribution. When it is new, it will always need fitting into the digital context the rest of the e-Infrastructure provides, either as it is imported or as it is developed *in situ*. After that introduction, it would be ideal if it were managed entirely automatically, e.g., deployed to appropriate platforms, scheduled, parallelised, and was fed appropriate input, and the uses of its output were understood by the system. Today, such automation is lacking except in limited cases and contexts, as there is no established and widely applicable model for describing software sufficiently precisely, as a result, there is insufficient payoff for investing the effort. The *Semantic Linking* research – see Section 0 – plans to investigate this in the context of the ENVRI reference model – see Section 3.10.

4.2.12 Assessing the semantic linking review

The *semantic linking* technology – Section 0 – will significantly contribute to issues such a cataloguing – see Section 3.4, curation –see Section 3.3, provenance – 0, architecture – see Section 0, and reference model – see Section 3.10, as well as being the only sustainable path to harmonisation and inter-RI integrated, coherent views of data and services. The reason is, that it seeks scalable strategies for coping with diversity, by handling different ways of describing and representing all of the concepts, data and software of interest. This has a pervasive and substantial impact because, as we have seen, there are many forms of data and metadata in almost every RI context, and certainly between them. Where there are standards, there are often more than one standard that could have been applied, and scope for variations within standards. This is not just a matter of representational variation for the same entities or properties of entities. It is a deeper variance where the conceptual space is named and partitioned differently, and organised via different structures.

Researchers have long invested effort in accommodating these variations. They can always hand-craft transformations for the data from each source they use into a form the next stage of their

work requires. It has been estimated that such data wrangling takes 80% of researchers' time. This shows that transformations preserving relevant information are possible. But we have moved into an era where one-off solutions are not acceptable, the data for data-driven science has to be organised and presented for multiple uses – a commitment every RI and the ENVRIplus project wholeheartedly endorses. To improve productivity, to accelerate discovery, to reduce errors and to improve the cost benefit ratios from investing in environmental RIs.

However, such an oft-repeated goal of harmonisation, is not easily achieved. The combinatorial space of forms of data and metadata is too large for hand-crafted solutions to handle. Indeed, it probably grows faster than the capacity of the experts in data integration can write algorithms to handle differences. Consequently, the semantic linking strategy is to assemble higher-level descriptions, that can then be used to automatically generate and revise the transformation algorithms. Section 0 offers a very thorough survey of much current R&D addressing this topic, including that ongoing in contemporary EU projects, such as the VRE4EIC project[110], and shows that there is great potential, at least for data and metadata, if the approach is structured using the ENVRI reference model.

We can therefore illustrate how this will pay off in various parts of ENVRIplus, and thence in the e-Infrastructures and working research environments of the RIs:

1. The contribution to cataloguing comes through two effects:
    a. When data and metadata are loaded into the catalogue, if they are in a different from the catalogue's standard, translation shims[203] should be available to convert all of the components into the standard stored form.
    b. When a user or workflow seeks to access information from the catalogue, terms in incoming queries can be transformed into the catalogue's notation and results can be transformed to meet the requestor's requirements.
2. The contribution from cataloguing will be the organisation of the evolving set of transformations, and maybe the higher-level descriptions that are used to generate those transformations. For example, the transformations themselves would be described in terms of the semantic category of their input, the semantic category of their output, and information about how these are represented in a standard established by ENVRIplus. For example, when a query mentioned a geological era, say the Pleistocene, and the data were stored as years before present, a time range could be generated. To find that transformation, the cataloguing system requiring it would need to query the catalogue of transformations for transformers that handle named eras as input and yield YBP ranges as output—there may be several of these to handle various binary encodings or to handle different views on the equivalences. If no transformer is found, then a request to have one made could be sent. The (automated) transformation generator might enquire of the catalogue of transformations whether there are partial solutions, handling the input, or delivering the output, that would contribute to the new transformer. One aspect of the semantic linking research will give the terms that catalogue transformers and deal with their variations.
3. The contribution to curation is very similar. Curation systems tend to require most of their metadata complying with their chosen standards. Transformations to this can be handled as described for catalogues. The same ingest treatment for data can also be applied. Subsequent requests will potentially offer transformations to required forms. The curation system often requires a binding to software and computational contexts that generated the data. The issue of software description, discussed below, then comes into play.

[203] In the days before computer-controlled machines, engineering tolerances often led to slight misfits. Small inserts that accommodated the error were called 'shims'. The Taverna group coined that word for the automatically inserted data transformations to make the output from one step of a workflow have the form required as input by the next step.

4. The relationship with provenance is very similar. Provenance systems are far more usable and useful if they achieve a uniform representation of provenance, but the subsystems called during the enactment of an encoded scientific method often support different representations or only offer logs. Again, if a catalogue of transformations is available, it can draw on these. It may also deliver transformations to the tools and systems that exploit provenance.
5. One of the primary purposes of the architecture is to aid communication about the systems, as they are planned, constructed, maintained and used. In particular, the architecture should identify all relevant software, probably the dominant topic of discussion, e.g., platforms, middleware layers, data-intensive frameworks, storage systems, workflow systems, other subsystems, programs, workflows, libraries of functions, and main services, etc. It is necessary to describe precisely how they fit together. At an abstract level this role is well-supported by the contributions of the ENVRI reference model. However, as construction and operation are considered considerably more detail is needed. For reasons given in Section 4.2.4, the majority of the subassemblies and frameworks will be the product of other independent organisations, which will describe their product in terms they choose. Most of the conceptual space describing such components lacks standards or lacks the adoption of standards. Consequently, the concepts, entities, nomenclature, and units used vary greatly; they are also encoded differently and are grounded on different ontologies, most of which will not be explicit. Consequently, there is a substantial role for semantic linking, helping to consistently describe the parts in use and the operational relationships between them. This may prove demanding, as Section 0 explains, because ontological and semantic standards for this space are partial or under development. See below for a separate discussion regarding software.
6. There is a close relationship between semantic linking and the reference model. The developed viewpoints and existing ontologies will be a vital starting point and will shape the direction of travel. Development of the existing viewpoints will provoke revision of some of these ontological foundations and will require extensions. The depth and precision of the engineering and technical viewpoints needed to guide the architecture and e-Infrastructure construction will certainly demand extensions. It expected that the semantic linking and the reference model will advance in tandem for these topics.
7. The **description and handling of software** is a major challenge. Section 0 identifies work in progress. The recent work of EU wf4Ever project [Belhajjame 2015] describes extensive work describing workflows. The work in the ADMIRE to describe data processing elements included properties of input and output handling rates to inform optimisers and validators [Martin 2014]. The work by the Taverna group describing the web services available to bio-scientists [Bhagat 2010], opened up this issue for a substantial range of web services. In that case, it required *separate sustained funding* to build a critical mass of descriptions; it is an open question whether investment in relevant descriptions is already budgeted for in ENVRIplus and the RI communities. The deeper challenge remains to establish a comprehensive way of describing software that covers the range of software and aspects relevant for ENVRIplus, e.g., the metadata descriptions in catalogues raised in Section 3.4, the software used to be recorded in provenance and curation with the data it produced (see Sections 0 and 3.3) or the clarification of issues in architecture – Section 0. It may be useful to first identify the categories of software needed in the reference model and architectural conversations, and to pick off these one by one as a way of dividing that challenge into feasible chunks.

4.2.13 Assessing the ENVRI reference model review

The reference model, described in Section 3.10, is as much about organisation of the design and construction of Research Infrastructures as it is about technologies, and the review properly considers such matters. The previous ENVRI project saw substantial development of three

viewpoints of the reference model using ODP. Changing the underpinning technology for representing the large, distributed and multi-organisational systems and development efforts is therefore not an option. ODP has proved a good approach for tackling the complexities of distributed systems with the scale and diversity needed for environmental RIs.

There are three issues, clarified in Section 3.10, regarding the reference model:

1. Assessing whether the new and larger community of RIs require revisions to the three already developed views in order to meet their needs. Some revision will be necessary but the requirements gathering in Section 2 has not revealed any reasons for radical change. Other refinements to the existing views, such as the clarification of the stages of the data lifecycle – see Figure 3, will need to be recorded in ODP and the ontologies.
2. Developing the reference model further, so that it can be used to formalise critical aspects of the architecture – see Sections 0 and 4.2.11 – and to plan and organise collaborative construction and maintenance, involving the multiple organisations within an RI for bespoke software, the organisations of a group of RIs for shared subsystems, and the organisations of resource providers and technology producers as sustainability is taken into account – see Section 4.2.4. This is precisely what ODP was designed for, but it requires the development of engineering and technology viewpoints, as planned in the ENVRI plus project.
3. Gaining sufficient traction that the reference model is used, not only at the strategic and planning level, but also to shape and coordinate the work of the engineers actually building the e-Infrastructure. This use is required to avoid the large-project disasters that are inevitable when the interfaces and inter-dependencies are not precisely agreed *and recorded*. Word of mouth and exchanges of documents, email, instant messages are all vital at arriving at a mutual understanding about what should be done at such a boundary. But they are completely inadequate when each software engineer is pressing on with their part of the implementation. Anchoring the agreements in a framework provided by the reference model and ODP is a good way of mitigating such disasters.

Finding resources for each of these lines of development simultaneously will be a challenge. The first requires dialogue with a broad range of practitioners in each RI. A good start was made during the requirements gathering – see Section 2, and there are good working relationships inherited from ENVRI for some RIs. Engaging the people in each RI who have the relevant information and then analysing and recording it in digested form will take effort on both sides and elapsed time for digestion. Yet the next two issues probably require immediate connections; can these be based just on the legacy from ENVRI?

The second, extending the reach to selection and assembly of subsystems and their integration by configuration, software interfaces and bespoke front ends, requires substantial input from those who are expert in the details of each candidate component, and those who fully understand the engineering trade-offs. Whilst ODP and reference model experts can guide this process, record findings and coordinate, that buy-in from experts in various parts of the system is essential; but these experts are usually hard pressed developing solutions in their own context. Often they are under a lot of pressure to support existing deployments and to deliver new functionality or capabilities. They will not allow themselves to be distracted by engaging in work on the reference model unless there is an obvious pay off. This takes us to the final parallel line of R&D required.

The reference model could, in principle, save much effort by successfully partitioning and coordinating design and construction tasks to avoid duplication and gaps, and to ensure the process of assembly works smoothly with *the parts working well together*. There are three preconditions to enable this to proceed. First, a sufficient proportion of the software engineers, "at the coal face" of importing subsystems and developing software, need to engage: using the reference model when they have questions, and improving it when they find the current answers insufficient. Second, enough of the context in which they are working has to be described at the level at which they work, i.e., at least information, engineering and technology viewpoints. Third,

the third parties providing systems, e.g., platforms, resource providers and technologies such as database and workflow systems, have to engage, describing their systems and conforming to agreements cast in the reference model.

All three of these lines of development would benefit from improved productivity yielded by good tools. These tools should facilitate authoring, refining, validating and interrogating the reference model. Ideally, they should also support automated generation of interface and framework code. Section 3.10 identifies two commercial tools, but does not find any open-source tools of comparable power. It is not economically feasible to get such commercial licenses and follow up training from these vendors in the context of ENVRIplus but in the context of the wider ENVRI community and the long-term lifetime of the environmental RIs this would be practical. Given the scale of investment needed to construct, operate and maintain RIs over their extended lifetimes, engineering tooling is a strategic issue in RIs management

The long-term utility of design, analysis and system assembly is enhanced if there are suitable high-level definitions that are independent of details of specific platform technologies. The reference model is potentially a good medium for this, but it requires so much investment that it may only pay off if it also meets the practical needs of those building the e-Infrastructure, but this is precisely what ODP was designed for. A wholehearted commitment is needed to reach the thresholds where its benefits are felt by all of those planning, designing, building and maintaining e-Infrastructure for the RIs. It is an open question whether this can be achieved with the ENVRIplus resources.

4.2.14 Assessing the review of compute, storage and network provision

Section 3.11 considers the provision of ICT resources is essential to enable every step of the data lifecycle, every part of scientific method development, from teleconferencing about the first idea to the final polished and optimised formalisation as a packaged workflow, for assuring sufficient persistence for all data, metadata, software and their relationships, and for supporting the human-computer interactions of all practitioners in the geographically distributed communities. These resources build on globally and nationally provided underpinnings, such as the Internet, and span all the way to the computers, laptops and mobile devices individuals use. Section 3.11 views this digital ecosystem environment – the platform on which we build Figure 4 – from the viewpoint of pan-European organisations and focuses on three aspects:

1. **Digital communication**: The support for are data movement, computer-computer interaction beyond very local connections, and human-computer interaction covering anything from remote field or marine observers to experts in presenting results in a 3D-video cave. A good proportion of this happens under the aegis of GÉANT, that coordinates the interconnection of national network provisions. Specialised arrangements are needed for the remainder. The combination needs to be delivered as a consistent functionality with standard interfaces, even though the non-functional parameters vary greatly. Fortunately, the established internet standards and protocols deliver this consistency almost everywhere.
2. **Computation**: may take place at every step on the data lifecycle, may be a routine process that is applied to every batch of data, or may be a demanding one-off simulation run or massive data analysis. The provision for many of these computations is met by home-institution resources provided to practitioners. The high-throughput computing is met by local or regional clusters or by the pan-European cloud providers listed in Section 3.11. Similarly, the HPC facilities may be local, regional or under the aegis of PRACE. Then specialised computational platforms are needed for some data-intensive workload patterns. For productivity of researchers, developers and data-handling experts, presenting these facilities in a coherent and consistent manner would be very helpful. That would also open up avenues of optimisation, selecting targets that reduced costs, and protect investment in computer-based methods, as they would no longer be tightly tied to the platforms that are evolving. Unfortunately, today, there do not appear to be

standards for job-submission, scheduling, interaction with running jobs and provenance collection, that will yield the required consistency. The protocols, intermediate software, authorisation and authentication, diagnostic interfaces and paths to get support vary significantly over this range. In some cases, for valid reasons.

3. **Storage**: Underpinning all of the lifecycle are arrangements for transient or persistent storage of all of the artefacts involved: metadata, data and software. The choices of provision are determined by factors such as: data volume, access patterns, target persistence duration, and speed of required operations, cost and energy consumption. Such storage only preserves the bits to an understood reliability. It needs to be interfaced by software that handles bundles of data, such as files, sets of files and (scientific) databases. It also needs to maintain relationships between such entities, e.g., a file name, a query or some metadata normally held in a database, with bundles, such as a set of files. The primary pan-European organisation delivering storage and covering many patterns of use is EUDAT. Unfortunately, it has far less maturity and assured longevity than the providers of the previous two essential ICT components: digital communications and computation. This is a strange accident of history. It is an acute concern in these days of data-driven science and particularly for ENVRIplus focusing on the data lifecycle. The predominant viewpoint of today's technological provision of storage systems is to focus on the file. This omits the opportunity to exploit significant structures, often latent within or between files. The scientific databases are attempting to retain the good properties of files, but also handle such structure well. Even if we only consider the file systems, it is unclear how consistently functionality is provided and supported across multiple providers.

These pan-European resource providers, many commercial providers and some of the major institutional providers, e.g., PRACE and other HPC sites and national environmental services, also contribute to other important factors: affordable sustainability and support being the primary example – see Section 4.2.4. This benefit derives from several contributions, for example:

1. **Provision or shared support of significant subsystems**: Significant subsystems may be set up, tuned to current loads, mapped to suitable lower-level platforms and administered by the collaborating resource provider. For example, the many file handling and metadata handling services provided by EUDAT or the job submission and workload management services supported by cloud providers. Incorporating and building on their supplied systems substantially reduces the RI and ENVRIplus specific software that needs to be maintained and supported. In the long-run learning about and understanding how to use subsystems provided by others can yield substantial savings. The maintenance costs are amortised over a larger community. If ENVRIplus coordinates the RIs, so that many of them are building on the same subsystem, this benefits both parties. The resource provider (or vendor) gains a large boost to its 'customer base'[204] from a modest amount of mutual adaption and customer understanding—ENVRIplus presents a set of RIs and their e-Infrastructure as one customer with a reduced cost for customer acquisition and initial training. The ENVRIplus community gains the advantage of maintenance and advice from an organisation experienced in maintaining and advising on the subsystem. This is an important step in the sustainability of e-Infrastructures – See Section 4.2.4– an absolute essential if the RI communities are to depend on the e-Infrastructure. Consequently, developing a list of suitable subsystems to adopt is a high-priority ENVRI-plus task – see List item 22 in Section 5.1. This should be approach incrementally and be guided by architectural – Section 0 – and reference model – see Section 3.10 – considerations as well as requirements –see Section 2.

[204] For many of these pan-European or commercial suppliers, such increases to their customer base for modest invest increase their viability and hence probable longevity. Thereby decreasing the risk of that subsystem ceasing to be actively supported.

2. **Engineering and integration**: For each subsystem or platform layer in use, a great deal of effort is needed choosing the lower layers, mapping and configuring the installations and making the parts work together. This takes a great deal of engineering skill and multiple categories of specialist knowledge and system administration skills, e.g., the providers of Apache Spark have already developed effective integration with many other commonly required systems, such as Python, Mongo DB, and R. As another example Pegasus and dispel4py work well together, exploit a mapping to Apache Storm and have deployment scripts that set this up by deploying Docker images on a cloud infrastructure [Filgueira 2016b]. External organisation who provide a major subsystem will have done this work drawing on skills and knowledge they have built through experience or being able to bring in the necessary expertise justified by their larger user or customer base. Once such systems are operational, they need tuning to carry the presented workloads, then need capacity planning, and they need optimisation to reduce energy consumption and operational costs. Again, the increased user base and organisational longevity makes this much more achievable than it is for the individual RI or small group of RIs.
3. **Maintenance**: As explained in Section 4.2.4 there are three drivers which make maintenance essential: changes in the digital context that must be accommodated or should be exploited, extensions to functionalities and capabilities, or correcting latent errors. The first two of these dominate, and will certainly be important for every RI. Again, the organisations with larger customer bases will be able to afford to do this well and benefit from a larger pool of skills and engineering knowledge. Conversely, if the RIs adopt this route they will benefit from advances other communities have initiated.
4. **User and developer support**: Good user and developer support requires training material: on-line notes, webinars, re-playable sessions, frequently asked questions, compelling exemplars, etc., that take significant effort to produce, refine and tailor for the various audiences in Table 19. A larger customer base makes this much more feasible and increases the chance of good quality for a particular group at a particular stage of engaging with the subsystem.
5. **Connection to services**: The set of subsystems and the base platform delivered by a provider organisation still needs to be used from non-local devices and to support a wide set of interactions from people and computers. Establishing frameworks for this with virtual research environments or virtual laboratories that suit RI communities may draw heavily on interaction and user arrangements the provider has developed. For example, the authentication and identification provisions developed by EGI or GÉANT may be sufficient for some categories of user community. Similarly, external connections with other services, e.g., for bulk data transport at low cost, or for synchronisation with minimum delay may already be an established arrangement. Some forms of automated mapping to deploy required multi-node computing environments, or to handle particular forms of workflow, may be re-usable and save e-Infrastructure or scientific method developers much work.

The above list shows that there are potentially substantial benefits from working with some of these suppliers and with using some of the subsystems they offer. But it is impractical to use too many in one e-Infrastructure, they may not fit well together and the resulting e-Infrastructure would be excessively complex. Analysis of the suitable compositions should proceed by developing the engineering and technology viewpoints in the ENVRI reference model, and then using this as a framework to select a candidate list of subsystem and provider bedfellows to best host RI requirements. In the interim, use cases should investigate specific collaborations in order to increase the background knowledge available when that selection is made.

Making critical decisions about software subsystems

The long-term impact from decisions about which platforms and subsystems to use as an RI's e-Infrastructure is designed and constructed are so significant that decisions should be taken very carefully. However, they are often taken coincidentally. An individual or agile development team starts using a technology because it is familiar, is already used in an example they are

developing from, or it is the first that comes to hand. This is appropriate during agile co-development and when try to get a prototype running quickly. However, that needs to be de-coupled from longer-term commitments. The complex set of aspects affecting such decisions are set out in Section **Error! Reference source not found.** as in major construction projects it is often the architect who has to identify such crucial questions and ensure that they are answered by suitably qualified and constituted groups representing the clients at present and in the future. An example was also illustrated at the end of Section 3.8.1. Factors such as comparing the up-front costs (financial, staff training, installation effort, disruption to current working practices) against the long-term costs (energy consumption, platform costs, staff time for maintenance and user support, licenses and service costs, etc.) depend very much on time scales, target availability and reliability, and required usability. These are policy matters, as is the judgement of the impact of ICT delays versus cost, or rapidity of processing and responding to a user versus utilisation of a platform. Of great concern is whether the user community will adopt features they could benefit from. Whether staff already performing many recurrent jobs will take on extra ones. These are policy issues that need clarification. They may only emerge when decisions need to be made. Consequently, there are at least the following factors that affect the quality of ICT decisions:

1. The quality and clarity of the **policy framework** and procedures for revising it if necessary.
2. The final **decision making body**: who oversees it? Who establishes its members and gives them authority? What is its scope? Does it have the right mix of experts? Does it represent all relevant constituencies? If it is on a per RI basis? How do the inter-RI factors get assessed? How prompt are its decisions? How binding are they? What resources does it have?
3. An appropriate **decision making procedure**. How are the important decisions recognised, brought to the fore and clarified? Who gets a chance to make input? What investment is made in evidence gathering through investigations, trials, benchmarks, and systematic comparisons? How easily can it obtain advice from experts? For example, can it employ consultants or make potential commitments to suppliers in exchange for them supporting the investigations? How are these information gathering exercises organised?

So much time of so many people: researchers and all the other roles (see Table 19) will be wasted or saved depending on the quality of these decisions that it is well worth investing significant staff time making the decisions carefully.

The decisions may be partitioned into tractable steps. These steps interact significantly, so they are potentially intractable if taken all together. Some of the providers of resources may offer a bundle of the choices, so that selecting them pushes you towards a particular choice on many points. We illustrate the idea of steps by working from the lowest levels of the platform upward:

1. **Computational provision**: The majority of this is provided by using standard computing nodes organised as clusters or clouds. Some provision is needed for specialised loads, such as those requiring high-speed interconnection between nodes, and those requiring very high data I/O, or those benefitting from very large shared memory. Questions arise about how this provision is organised and paid for? Whether its location is important for regulatory or data transport reasons? What operating systems, virtualisation methods, resource allocation methods and monitoring methods are supported? What file systems and database systems are included? How well is data transport on and off sites supported and what does it cost? What are the local storage provisions? What workload submission mechanisms does it support, e.g., for batch jobs, for complex graphs of tasks in workflows, for sustained services and for interactive use? What support is there for diagnostics, workload analysis, resource planning, elasticity and provenance capture? What forms of usage records and accounting does it support? Further discussions about some systems can be found in [Simmhan 2016] and [Fox 2016]. Energy costs should be taken into account; two examples are [Wilde 2015] and [Siew 2016]. The provision of

computation that matches environmental science requirements, at least at the HPC end of the spectrum, is considered by [Frank 2016].

2. **Storage provision**: What scales of storage and rates of data transport are provided? At what cost can these be expanded? What is the reliability of persistent storage? What underlying mix of technologies does the system use? Does it automatically handle this mix presenting simple stable interfaces? What data organisational systems are well mapped to these underpinnings: distributed file systems, scalable database systems, composite dta management services, e.g., those providing high volume catalogues of metadata bound to data, with or without PIDs? What forms of usage records and accounting does it support?
3. **Digital communications**: What range of geographical regions, categories of users and categories of organisation does the service cover? What models of protection and resource sharing does it use? How are gateways with other digital communication systems arranged? What protocols does it support? What are the ranges of bandwidth and latency it offers for the data trips relevant to the RIs concerned? What forms of usage records and accounting does it support? Some discussion can be found in [Masson 2016].
4. **Authentication, Authorisation, Accounting and Identification (AAAI)**: What range of users do the AAAI mechanisms support? For example, from members of the public to administrators controlling access to major or sensitive resources. What modes of identification does it handle? Does this cover all of the community's requirements with minimum disruption to existing practices? What range of trust is associated with the authorities allocating credentials? How many different trust relationships have to be taken into account? Can sensitive resources and roles be restricted to people granted authenticated identity by sufficiently trusted authority? To what extent are roles and groups managed? Can authorisation take into account properties of the data, such as its source or date of creation? Can permitted actions take into account the authority as well as role? For example, members of the public might be restricted to a maximum of 1 minute on 5 cores and 0.5 GB result delivery. Whereas, a project or RI authorised research might by default have a limit of 1 hour * 100 cores and a delivery of up to 0.5 TB of data. Exceptional competition winners, who had convinced peers of the value of their planned work, might submit 100 hour 10,1000 core jobs yielding 0.5 PB of data. Frameworks accommodating such a range may be needed.
5. **Data-intensive middleware**: What kind of middleware will best suit the workload patterns, e.g., Apache Storm, Apache Spark, Pilot Jobs [Turilli 2016] or data-intensive workflows, e.g., Pegasus handled the LIGO data for gravity wave discovery [Deelman 2015] or [Simmhan 2016]? How well will it scale given sufficient nodes? How well does it parallelise automatically?
6. **Database systems**: What DBMS does the community already use and how deeply are they bound to its idiosyncrasies? What models does a candidate DBMS support: relational, XML, SQL and NOSQL, RDF, scientific? How well does it support development and production? How does it scale? Haw well does it exploit new technologies for storage and new models of computation provision? Does it provide time-stamped query support? A participant, such as BGS uses Oracle databases to underpin many complex and very large catalogues and information services that meet demanding targets. The astrophysics community, IVOA see Section 1 and [Szalay 2008], pioneered very large scale sky catalogues that have to manage identities of many (more than 10^7 or 10^8 objects), with sophisticated metadata, high rates of ingest, and demanding global user communities. Their queries also involved significant computation. This has led to a whole community of database researchers and demanding users, that meets regularly at the eXtremely Large DataBase conference (XLDB)[205] that should be tracked by the

[205] http://www.xldb.org

ENVRIplus community. This campaign has triggered substantial modifications to traditional DBMS, for example, in the range of data, including files, that they handle and in providing non-atomic options for large updates. But more significant is the flowering of scientific databases that previously were a small research niche. These hold time series, matrices and bulk data such as images. They also accommodate the representation of uncertainty – value ranges – and support efficient queries over all of these data types. Examples are: SciDB[206], MonetDB[207], which is organised as a column store to accelerate retrieval, particularly of projections that are often used in science, and Rasdaman[208]. The latter is already well integrated with geospatial data and is being considered for EISCAT data. How ell does the system support distributed query and for what query languages?

7. **Workflow system**: What kind of workflow system is needed, e.g., task oriented c.f., Tavrena, Pegasus and KNIME, or stream oriented c.f., dispel4py? How well does it help scientists formalise, share and develop their methods? How well and how automatically does it handle all of the data management needed, e.g., movement of data between sites? How well does it adapt to load and context? How well does it handle partial failures: clean up and resumption with minimum re-computation? How optimally and automatically does it select platform targets and map onto them? How well does it deliver provenance? How good are its diagnostics? What forms of usage records and accounting does it support?
8. **Interaction system**: What modes and models of user interaction are supported? How easy are these to use? How easy are they to develop for? How many tool sets and apps already use them?
9. **Tools**: The crucial issue is productivity, e.g., how well do researchers get on when using this system? Providing them with tools that meet their needs is a key part of enabling their productivity. To what extent do the tools meet their needs? How hard is it to learn to use them, as an expert in some other tools, as a novice? How well will they integrate with the planned system? For example, reporting the information required by the provenance system in a form it can handle? Improving the productivity of those who design, develop, run and manage the e-Infrastructure is equally important, as that will enable them to deliver better facilities and more agile support. Consequently, investment in their tools, with similar considerations should be undertaken.

All of the above are illustrative of the appropriate partitioning and illustrative of the questions that should be asked for each partition. It may be a good question for a think tank: How should the RIs partition the decisions about which software systems to use? Ancillary questions are: To what extent can they do this collaboratively? And: How will it be resourced?

Once the partitioning and sequencing is agreed, each of the investigations should be launched in some optimised order. Each investigation may be sparked off by its own think tank, deciding on scope, constraints, key questions, experts needed and a plan for conducting the investigation and forming a conclusion. The same ancillary questions apply. The investigations should also consider sustainability factors and support issues, such as the extent to which the candidate product has an active user community with similar requirements, the resources and expertise the supplier has, and so on – see Section 4.2.4.

4.3 Characterisation of Task 5.1 outcomes and implications

The overall findings of requirements gathering and technology review are consistent with the Theme 2 plan, and indeed with other parts of ENVRIplus. Thus the position taken when the project was proposed is largely refreshed and endorsed. However, there are many detailed

[206] http://database.cs.brown.edu/projects/scidb
[207] http://www.monetdb.org
[208] http://www.rasdaman.com

findings that are collated below (Section 5). We introduce a number of categories immediately below and list under those headings the specific, tactical and organisational suggestions that should mainly be considered by Theme 2 in Section 5.1. The longer-term and strategic issues are collated in Section 5.2. These should concern those considering the future direction of RIs and of the environmental cluster. Some may have further reach.

We recognise here that many of the detailed recommendations emerging from Task 5.1 have a relatively short-term or localised relevance. Examples are enumerated in Section 5.1. These can be categorised into the following groups:

- **Making best use of Task 5.1 results**: The follow up needs to ensure the information represented by this document and the associated wikis is accessible, well presented, and effectively communicated and promoted. It will be worthwhile investing in maintaining the information to keep it up-to-date and in building on the communication networks and mechanisms already established, such as the ENVRI Community Platform[209]. It is a foundation worth building on, but it is also worth maintain. It should remain a live resource.
- **Universe of discourse**: The conceptual frameworks for supporting the research in each RI and the nature of underpinning technology is complex and full of detail. This presents barriers for immediate usability and implementation of functionality. However, the longer-term frameworks for working practices and the software that support them are better understood when they are discussed in terms of a more abstract viewpoint. The ENVRI Reference Model provides a vocabulary for discussion at this level. As far as possible, requirements and solutions should be expressed in such terms, to reveal potential commonalities and properties that change relatively slowly. The detail should then be built up, e.g., using the engineering and technology viewpoints – see Section 3.10, to deliver and support user requirements in a more sustainable way.
- **Awareness raising and training:** There is a widespread recognition of the need to improve and extend understanding and at the same time encourage cross-boundary communication – see Section 4.2.1. The role of Theme 2 (supported by Theme 5 Knowledge Transfer) in this internal campaign will include urgent transformation of results into material key for understanding potential and issues. The key asset of the RI communities and the ENVRIplus project is a great diversity of experience, viewpoints and skills. Every effort should be made to capitalise of this strength by pooling intellectual effort in a series of 'think tanks' that focus on priority issues and bring together experts from across the borders between disciplines and across the borders between technological viewpoints to create new strategies. A series of calls for proposals for such think tanks, should be organised.
- **Usability and take up:** The adoption of new data-handling technology will depend on how well it is packaged and made available. Several aspects of this should receive attention. This is crucial for sufficient take up to happen before harmonisation and resource or effort sharing will be achievable.
- **Shared subsystems and sustainability**: There are several important reasons why sharing substantial subsystems, often intermediate level software frameworks, is valuable:
    - The costs of building, installing, configuring and supporting these shared capabilities are then shared, bringing economies to the RI's when they set up, run and maintain their e-Infrastructure. Sustainability becomes more achievable.
    - Having relevant experts available to establish, tune and support these subsystems becomes achievable if there are fewer instances to consider.
    - Underlying software consistency makes harmonisation more feasible, and reduces the effort involved in boundary crossing, or moving to a new context.

[209] http://www.envri.eu/

As Section 4.2.4 explains, scientists and all the other practitioners associated with an RI and its community quickly become dependent on the software that enables their scientific methods and working practices. Consequently, loss of that software could be a severe blow. The ENVRIplus project and the RIs have budget for the steps necessary to sustain software –see Software Sustainability Institute (SSI) – typically 95% of software's lifetime costs. This motivates the need for great care in the choice of software already well supported, and limits to the additional software on which the RIs will depend.

Over the next four years, arising from political initiatives of the European Research Area, the ESFRI Forum, the e-IRG, the European Cloud Initiative, etc., the relationships between the RIs and the foundational infrastructure providers (EGI, EUDAT, PRACE) assume a great prominence. The RIs will be under considerable pressure to learn to outsource their IT needs and to work with these providers. These providers really have to learn to adapt and to be agile in meeting RI needs.

Groups of RIs with clearly articulated similar requirements will be in a much stronger position for negotiating and developing alliances with those providers, as well as with commercial IT consultancies, developers and suppliers. Opportunities for new markets potentially attractive to SME ICT suppliers (especially software suppliers) will be created through harmonisation when RIs act together.

We then draw attention to aspects that have longer term or more pervasive application. However, both are important. The short-term aspects have to be addressed to meet immediate needs so that practitioners in the affected domains can make progress in the short term. This then builds confidence in interdependencies and technologies that is essential for sustained investment and collaboration. Without this, researchers will not trust the emerging technology. They will avoid dependency and in consequence fail to reap its full potential. Once that confidence has been built the longer-term issues become critical. They address the strategic questions as to what routes to take to sustain and continue to advance the research without incurring unaffordable costs. Some suggestions of strategic issues that may be considered are listed in Section 5.2. They arise from the issues discussed in Sections 4.2.1 - 4.2.4.

5 Impact

Section 4.1 has drawn together the outcomes of the requirements gathering. Similarly, Section 4.2 introduced four general issues and then summarised the technology reviews. We now consider, based on the categories outlined in Section 4.3 the implications for the ENVRIplus project in Section 5.1 and longer-term issues that may concern the RIs or the wider community are considered in Section 5.2.

5.1 Impact on project

The short-term and focused results from Task 5.1 lead to a series of confirmations of current plans and a few issues that require attention and potentially could lead to modified plans. Items of concurrence are dealt with lightly or omitted if they have already been stated. Items provoking further thought and investigation are listed in the order that they are reported in the above work. In consequence they are not in any way prioritised[210]. The Theme 2 and ENVRIplus management should consider whether these need further attention and if so, how to prioritise and resource the follow-up activities.

[210] A numbered list appears here to allow cross reference and issue identification, not to imply any ordering due to time or importance.

Making best use of Task5.1 results

1. Ensure that the ENVRIplus Glossary contains all of the non-familiar terms used in this report. This action will be undertaken by the Project management and Task 5.1 teams after the Spring 2016 ENVRI week.
2. Establish the wikis used by this report in a properly supported context. This work is currently underway due to the help of EGI working in conjunction with Alex Hardisty, CU and Magdalena Brus, UHEL. The expected completion time is May 2016.
3. Much of the recent Task 5.1 work has focused on improving the content of this document. However, the primary reference material delivered by Task 5.1 will reside in the above wiki. It will be necessary to update the wiki with all the insights, useful presentations, tables and figures in this report. This report itself, once it is agreed for publication, should also be made accessible from that wiki in each of its spaces. This wiki-update is the responsibility of the Task 5.1 team and should be completed soon after the Spring 2016 ENVRIplus week.
4. The report contains an initial set of references that may be useful to the whole project and will certainly be needed for future documents written by Theme 2 members. Theme 2, or the whole of ENVRIplus, should set up arrangements for sharing the list; via organised wiki pages.
5. Attention should be drawn to the report and its conclusions via the ENVRIplus newsletter. A meeting to plan this has been scheduled for the 2016 Spring ENVRI week. Short documents (3 to 5 pages) summarising specific aspects of this report targeted at particular readerships should be developed, for example, one to be considered by each work package and one to be considered by strategic planners in the environmental RI cluster and beyond. The plan for this, choice of target audiences and commitment to the editorial effort required will be made during the Spring 2016 ENVRI plus week.
6. External attention and critical review of the conclusions should be obtained by publishing one or more derived papers. These should encourage open debate and refinement on the relevant wiki spaces. Target venues include: conferences such as: IEEE e-Science and WORKS, EU project events such as: EGI and EUDAT, and EU H2020 informatics events, journal: such as Future Generations of Computer Systems (FGCS).
7. This refinement of the requirements and technology wiki spaces should be moderated. If it is successful, the ENVRIplus project steering committee should consider how it can be supported and continued after the end of ENVRIplus.
8. One should not do anything to slow down or distract agile development teams tackling a use case; however, the gathered requirements and technology reviews should provide useful information as they start. If they have time they should report any deficiencies that are noticed. At the end of their campaign they will have deeper knowledge about requirements and technology. If they can contribute key points or a synopsis of this to the relevant wiki that would be very helpful. This will be part of keeping this material live and building on it.
9. The present requirements gathering and analysis is an initial solid effort in a fast changing field. In some topics there is room for improvement of coverage and precision, whereas in others, sharp definition has been achieved, but they could still be overtaken by external events in the digital ecosystem. In any case, the digital world is experiencing a period of rapid change that is perturbing the digital ecology and the patterns of work using its resources. As this change is substantial and rapid, the ENVRIplus project should commit to a focused effort of update, so that it is coordinated, has high impact and is efficient. We would suggest that this should happen between M24 and M30 of the project. This will need resources and possibly should be shaped by a think tank that ensures that it has substantial contributions from the RIs and practitioners working at the 'coal face' of developing and running e-Infrastructures. It may be appropriate to set up a think tank to establish its direction and to recruit a balance of viewpoints.
10. An integrated view of the technology reviews should be developed as a roadmap, for the next ten years of supporting data lifecycles in data-driven sciences, particularly those

falling within the ENVRIplus area of authority. This will be feasible if we take a high-level view informed by the reference model. After allowing time for the baseline information to be refined, this may also be a suitable topic for a think tank. It would require sustained effort to develop this. The think tank might clarify the scope, identify key questions, make the case for valuing such an effort, recruit resources and a balance of experts, and initiate the work. Thus, this think tank would need sufficient time and effort itself to delve deeply into the topics and to form a coherent integrated view of the potential roadmap.

11. Consider making ***Optimisation*** a cross-cutting concern as localised optimisation may simply move costs into different working practices, into different stages in the data lifecycle or into different subsystems within an e-Infrastructure.

Universe of discourse

12. The relationships between the requirements gathered and the reference model should be studied and clarified—the focus of Task 5.2.
13. The technology reviews should be input to initial versions of the engineering and technology viewpoints. This may reveal structural issues in the reference model, or areas of engineering and technology that need to be better understood—again, planned in Task 5.2.
14. The cross-cutting architectural review – see Section 0 (page 102) – should be analysed by casting critical parts of its proposals in terms of the RM, as the architectural review is intended to shape the systems built by RIs, or the kits from which they are built, and their interworking, and this is precisely the role of the reference model.
15. Review the use of workflow notations and APIs in the current and planned e-Infrastructures to ensure that as far as possible they are cast in higher-level terms to avoid undue lock-in to particular technologies and representation details.

Awareness raising and training

16. The requirements gathering discovered widespread interest in further awareness raising and training to help those engaged in ENVRIplus and in associated RIs. This would help participants better understand the existing plans and implementation strategies and their potential benefit for users and e-Infrastructure builders. Topics range across all of the stages of the data lifecycle and courses could target each of the roles —see Table 19. The R&D teams engaged in Theme 2 should actively engage in this, including communicating about topics that need explaining, as well as topics in which they are already expert. This should be conducted as a webinar series to reach the distributed community. It should also involve multiple disciplines and multiple roles as a means of stimulating boundary crossing. Webinars and/or their supporting materials should become available via the ENVRI Community Platform[209] and training platform. See also item 5 in this list above, which may identify initial foci and prepare some of the relevant material.
17. The understanding is greatly accelerated and participants are convinced, if there are working examples of good solutions for participants to try in remote, supported hands-on sessions. This means that effort should be invested in forming *easily presented and understood exemplars* of developing services and functionality, with a corresponding illustration as to how they benefit the research goals or those supporting the research communities. As above, these should become available via the training and community platforms. Visibility can be promoted by promoting *ad hoc* think tanks of engaged scientists and infrastructure operators, especially to initiate cross-border cooperation. Calls for proposals on topics for *ad hoc* think tanks should be considered. By giving selected think tank proposals a high publicity profile, their example role is expected to strongly contribute to convincing the wider community. It also is a mechanism to enhance the joint ENVRIplus cooperation, even after the end of this project.
18. The above two activities should provide an opportunity to refresh and deepen ENVRIplus's understanding of requirements and of all the practitioners working with

data in the RIs. It may help transform the initial results from each agile task force into more widely applicable solutions.

Usability and take up

19. Expectations about the quality of interfaces and tools for managing and analysing data are rising rapidly, due to commercial investment into good interaction frameworks for mobile devices and mobile workers. It is reasonable to expect that the researchers and the support teams can also have such good interfaces for their work with environmental and Earth-sciences data. APIs should be structured and shaped appropriately, e.g., as microservices offering suitable controls and formats, to use available interaction frameworks wherever relevant. This will help communicate the value of ENVRIplus data and products to a wider audience of potential adopters, as well as helping in the internal communication identified above.
20. In alliance with others, deliver with high priority the simply packaged functionality that many researchers need. For example (from Section 4.1.9) we may consider:
    a. A packaged service for accessing data, bringing together the various parts currently being developed in a simple to use manner. This could, for example feature a 'shopping basket' into which researchers place items (datasets) found in catalogues for later retrieval and use; and recommender features along the lines of 'users who retrieved this data, also retrieved …' or 'datasets frequently used together' (cf. the functionalities of typical online marketplaces, such as Amazon and Alibaba).
    b. Provision of user and group workspaces, for handling their data under their own control, but with easy download, e.g., via the access facility, upload to any ENVRIplus service, and export. Practical mechanisms for enabling researchers to easily manage their own data collections: selected and possibly copied data from any of the sources they have access to, data they have produced, intermediate results for methods they are developing, results that are pending publication. Making this easy has been shown to be very useful in other contexts. [Schuler 2014], [Wolstencroft 2015].
    c. A context for developing encodings of methods, e.g., as workflows, where users can easily experiment, develop, test and refine their ideas. Once they have validated this to their satisfaction they can submit their work unchanged to production scale facilities to progress towards production. They can move fluently between the experimental context, e.g., on their laptops, and the production context, e.g., a data-intensive or HPC cluster, easily and without changing their representation of the formalised method. This was provided in VERCE [Atkinson 2015]. It greatly accelerates the rate of innovation as researchers can make progress themselves and call on other experts only when they have explored their ideas to their own satisfaction and are now ready for repeated production use of their method, at which point some optimisation by experts may be warranted.
21. Improve the training, culture and tooling of software engineering that facilitates the work of the engineers building the e-Infrastructures for RIs, and the ICT experts helping fashion and formalise scientific methods. Similarly, for those overseeing and implementing stages of the data lifecycle, particularly curation and publication. Improving their work practices, working environment and productivity is as important as that of the end users in the long term. To achieve this, consider bringing in or at least sharing experience and pooling support effort for tooling. Such campaigns might benefit from close collaboration with organisations such as: EGI, EUDAT and PRACE, as well as with relevant commercial tool and training providers. It may be assisted by working groups of the RDA. It will need investment of effort by ENVRIplus to reduce costs over the longer term.

Shared subsystems and sustainability

22. Identify a small set of middleware components, such as data-intensive frameworks, inter-process communication and NoSQL or Scientific databases that should be part of the common shared e-Infrastructure being built for or supplied to RIs. For these vital underpinning software layers, locating them centrally in the Engineering and Technology viewpoints of the ENVRI Reference Model makes it clear to everyone the role they play in an RI (Section 3.10.3). Choices should take into account sharing responsibility with other organisations and the sustainability issues discussed in Section 4.2.4. Consider the priorities and scheduling needed to establish these in processing environments.
23. There is widespread need for workflow systems reported in many areas for automating and refining all of the data handling steps in the data lifecycle, and for implementing scalable and repeatable scientific methods for research and the applications of research. At present many different technologies and systems are in use. ENVRIplus should help the RIs and their communities focus on a small number of workflow systems, and then establish shared maintenance, development and support to sustain those for as long as required. They will need to work well with the layers chosen for item 22 above, and similar criteria apply. But there are additional criteria here, as it is necessary to cover the different patterns of data-intensive workload, and to seek good tooling for all of the roles involved. ENVRIplus should also help the research communities adopt a mechanism for sharing, such as myExperiment [De Roure 2009], so that encoding of methods and parts of methods get re-used rather than re-invented. They should also help with the introduction of appropriate curation for workflows, for example as suggested in [Belhajjame 2015].
24. Consider the extent to which the architectural proposals in Section 0 can take into account the longer term potential for common support for data-intensive federations – see Section 4.2.3. If practical, steer the architectural developments to facilitate replacement of relevant parts with a common DIFF kernel.
25. Consider the extent to which the architectural proposals in Section 0 and the more specific topics take into account sustainability issues identified in Section 4.2.4. Take these issues into account wherever and whenever it is feasible to do so, to save long-term maintenance and support costs.

5.2 Impact on stakeholders

This Section is aimed at the strategists in the Research Infrastructures, as the recommendations are longer-term. The impact of taking them into account may not be significant in the lifetime of ENVRIplus, i.e., until 2019, *but they will be significant for RIs* as they have planned lifetimes of 25 or more years, and their scientists will depend on their ability to sustain as well as develop capabilities. Shorter-term implications from Task 5.1 are dealt with in Section 5.1, and include initial steps preparing for these long-term strategic issues. Theme 2 will ensure that RI stakeholders are properly consulted as they consider any issues that have been raised there.

The impact on stakeholders is restricted to longer-term issues, so that they are able to consider these in their strategic planning. The topics raised in Sections 4.2.1 to 4.2.4 each lead to strategically significant issues which should concern the RI stakeholders during ENVRIplus and beyond. We conclude by raising the issue as to how decisions about ICT choices which will have a very long-term impact are made.

1. **Improving interdisciplinary collaboration**: The background to this is summarised in Section 4.2.1. The ENVRIplus community should help drive a wider programme to gain recognition of the value of people who are adept at working across boundaries. It should join in the growing campaign to develop training and education opportunities for those who want or need the ability to collaborate across traditional intellectual, cultural and role boundaries. This should include wide-reaching summer schools. It is probable that funding for such campaigns could be gained at this time, by association with the popular

data-science flag, but it should be steered to take a broader view of the issues. ENVRIplus leadership might at least stimulate proposals for such activity and endorse those that look well judged. A think tank might be convened to plot the strategy for an effective campaign. Early summer schools could clarify the critical issues and inject understanding into proposals to build a pan-European educational campaign to *substantially improve Europe's boundary-crossing collaboration capacity*.

2. **Leading the formation of a global environmental sounding board**: Just as for astrophysics – see Section 1 – the observation and understanding of environmental and Earth systems requires a global collaboration and significant contributions will come from all parts of the planet. Should ENVRIplus join forces with others globally to create and build an intellectual meeting place where this collaboration is nurtured and crystallised by agreeing *and adopting* the relevant standards globally? This would help with the boundary crossing needed, see Section 4.2.1, and in the longer term develop strong foundations for global campaigns to address pressing challenges. Some environmental disciplines have their own standard-setting international bodies (such as TDWG for ecology and biodiversity[211]) and sometimes have collaboration with similar bodies. ENVRIplus may want to initiate an exploratory meeting to consider developing a sufficiently broad and effective framework within one of them, or cooperate with similar organisations to steer and develop global alliances? Such an intellectual meeting place could act as a complement, from the RIs' perspectives, to the Belmont Forum[212] – the world collective of major and emerging funders of global environmental change research. Is the ENVRIplus Board of European Environmental Research Infrastructures (BEERi) the seed from which such a forum can grow? What can ENVRIplus best do to help?
3. **Combining both statistical and numerical methods**: The background to this is summarised in Section 4.2.2. Many disciplines and research groups in the environmental and Earth sciences are benefitting from a growing wealth of observational data. These data remain sparse and suffer from gaps where observation is difficult, high error rates in some cases, and in almost all cases a shortage of elapsed time for understanding dynamics. Statistical methods, such as machine learning are very powerful at gleaning information from such data. Numerical models are extremely powerful at representing current understanding and enabling researchers to explore its consequences. Bringing these two paradigms of studying nature into one framework often yields substantial progress – perhaps best illustrated in the climate modelling campaign [Edwards 2010]. ENVRIplus, or maybe even an extended cohort of RIs should lead the formation of a theoretically well-founded and systematic approach to their combination. It would be a nice example of a reason for establishing an ENVRIplus think tank. Possibly it should lead to proposals for new European collaborative projects developing the necessary methods and their implementation. A new culture is needed. The RIs of ENVRIplus could lead in its foundation.
4. **Sharing computationally expensive results**: As the scale and resolution of simulations are improved and as the scope and scale of statistical analyses increases, the time and energy needed to produce results grows substantially. Consequently, there are powerful reasons for sharing intermediate and final outputs from these computations: (i) that enables many researchers to explore and evaluate the data who would not be able to raise the resources to regenerate them, (ii) it encourages inspection and validation of the scientific method and its application in these circumstances, and (iii) when it avoids regeneration it saves substantial energy. To achieve this in fields where it isn't established practice requires:

[211] http://www.tdwg.org/
[212] https://www.belmontforum.org/

a. Agreement on the representation of the data and the required metadata so that others can interpret them correctly.
b. Mechanisms for storing, publishing and curating these data, with associated policies.
c. A business model to support the data storage, access and analysis for publically announced predetermined periods.

These should probably fall within the scope of arrangements for curation – see Sections 3.3 and 4.2.6. However, a new *modus operandi* and business model may be needed. For some expensive computations the data involved are massive. This means the movement of the data they produce is prohibitively expensive and analyses of them can also be expensive. So storage of the bulk of data may need to be close to its computational source, and the curation system may only hold the reference to it plus the metadata. For these large data cases the computations exploring and analysing the results also need to be co-located with those data – but funding that supports the simulations may not support the storage and subsequent analyses. As few moves as possible should occur during the lifecycle of these massive data sets. Finally, the duration of retention may be quite finite, e.g., 6 months, and often pre-specified. This is partially to reduce storage and access costs, but also because such result sets tend to be overtaken by re-runs with improved models or source data. There are many lessons to be drawn from numerical weather prediction and climate modelling in this regard [Edwards 2010].

5. **Data-Intensive Federation support**: The background to this is summarised in Section 4.2.3. Almost every RI is developing or depending on a data-intensive federation (DIF) bringing together heterogeneous data resources from independent and autonomous organisations. Some of these organisations are participating in many DIFs and many of the organisations have priorities and rules imposed by their governments, constitutions or funders, that mean they can only participate fully if those factors are properly considered and compliance with data and resource usage rules can be enforced or trusted. Consequently, frameworks that facilitate the establishment and operation of such DIFs would be widely used and very worthwhile. Within its project lifetime it is infeasible for ENVRIplus to conduct the R&D necessary to create a reusable and configurable DIF framework from collaboration rules to technological support. However, ENVRIplus could spark a campaign to get the EU to recognise the long-term importance of this issue for data-driven science and business. Building on prior work of relevance ([Camarinha-Matos 2006], [Patel 2006], [Atkinson 2013a]). It might then join or lead a research campaign to build the necessary foundations and to launch the necessary R&D with engagement from the IT industry.
6. **Software sustainability**: The background to this is summarised in Section 4.2.4. The decision to depend on software is as important as the decision to depend on an instrument and it should be taken equally carefully. This will lead to an identified list of mission-critical software. Each RI and ENVRIplus should establish mechanisms for determining that critical list. The list should be minimised by careful use of commercial and *well-supported* open-source software. The members of the residual list of software must be maintained or replaced throughout each RI's lifetime. This requires appropriate resources, particularly software engineering staff and processes with appropriate quality controls. Wherever possible these should be met through alliances.
7. **Promoting ICT harmonisation**: The background aspects to this have been covered in Sections 4.2.4 and 4.2.13. Both software sustainability and engineering tooling are identified as critical issues affecting productivity and overall lifetime costs of RIs. Software, and the means to engineer and maintain it is expected to assume significant and increasing proportion of overall costs. Indeed, the overall cost of the ICT needed to support the production, acquisition, curation, publishing and processing of data throughout its lifecycle should never be underestimated. There is significant economy of scale benefit from promoting ICT harmonisation across RIs wherever possible, from

interacting with foundational e-Infrastructure providers like EGI.eu, EUDAT, PRACE, etc., and from encouraging the adoption and use of commercial / industrial engineering tools for development and maintenance. Promoting ICT harmonisation has the added benefit of opening (a) new potential market(s) for software and services for research infrastructures, for research data lifecycle management, for scientific data collaboration, data-intensive federations, etc. (to suggest but a few) that can be of interest to and stimulate growth among SMEs, e.g., by engaging in co-development to handle the complexities of data diversity in federations of autonomous data owners[213].

8. **How should decisions be made?** The importance of making decisions about technical choices in the construction and maintenance of the e-Infrastructures for RIs has been discussed and illustrated at the end of Section 4.2.14 with the many aspects that should be considered and balanced presented in Section 4.2.11, as architects normally insist that their clients take such decisions carefully for buildings. Many of the background issues are presented in Section 4.2.4. It is clear that making wise decisions about technological ICT choices is very important. It can substantially affect costs, viability, sustainability and user experiences. It requires a clear policy context to guide the preferred balance between conflicting pressures. It requires investment in careful investigations. It then requires a decision by a group with the right mix of expertise and with a proper understanding of the community's requirements and priorities. Should this kind of decision making be undertaken by each RI, or will there be pooled effort making such decisions and taking into account potential cost sharing and interworking? At the very least, the ENVRIplus community and the RI communities should be agreeing on how such decisions and policies are best made. *Decision making is vitally important for software*. Software is essential for all data-driven science and every member of the research community, those supporting the communities, and those exploiting the resulting information, depend upon it for their daily work. Software outlives hardware and even the business models by which computational provision is made. RIs will live for a long time with the decisions they make today about software choices or need to invest heavily to replace software if it is deeply embedded.

6 REFERENCES

[Abbott 2016] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Observation of Gravitational Waves from a Binary Black Hole Merger*, Phys. Rev. Lett. 116, 061102 – February 2016.

[Aceto 2013] Aceto, Giuseppe, Alessio Botta, Walter De Donato, and Antonio Pescapè. "Cloud monitoring: A survey." Computer Networks 57.9 (2013): 2093-2115.

[Acs 2010] B.Acs,X.Llora`,L.Auvil,B.Capitanu,D.Tcheng,M.Haberman,L.Dong, T. Wentling, and M. Welge. A general approach to data-intensive computing using the Meandre component-based framework. In *Proc. 1st International Workshop on Workflow Approaches to New Data-centric Science*, WANDS '10, pages 8:1–8:12, New York, NY, USA, 2010. ACM

[Almas 2015] B. Almas, J. Bicarregui, A. Blatecky, S. Hill, L. Lannom, R. Pennington, R. Stotzka, A. Treloar, R. Wilkinson, P. Wittenburg and Z. Yunqiang, "Data Management Trends, Principles and Components – What Needs to be Done Next?" Report from the Research Data Alliance Data Fabric Interest Group, draft version (paris-doc-v6-1_0.docx) from September 2015. Available via http://hdl.handle.net/11304/f638f422-f619-11e4-ac7e-860aa0063d1f.

[Altintas 2006] Altintas, I., Barney, O. Jaeger-Frank, E.: *Provenance Collection Support in the Kepler Scientific Workflow System*. L. Moreau and I. Foster (Eds.): IPAW 2006, LNCS 4145, pp. 118-132, 2006.

[213] Such issues are prevalent in business, commerce, engineering, health care and governmental administration.

[Amaral 2014] Amaral, R.; Badia, R. M.; Blanquer, I.; Braga-Neto, R.; Candela, L.; Castelli, D.; Flann, C.; De Giovanni, R.; Gray, W. A.; Jones, A.; Lezzi, D.; Pagano, P.; Perez-Canhos, V.; Quevedo, F.; Rafanell, R.; Rebello, V.; Sousa-Baena, M. S. & Torres, E. Supporting biodiversity studies with the EUBrazilOpenBio Hybrid Data Infrastructure. Concurrency and Computation: Practice and Experience, Wiley, 2014, doi: 10.1002/cpe.3238

[Arias 2013] Arias, M., Corcho, O., Fernández, JD. and Suárez-Figueroa, MC. *Compressing Semantic Metadata for Efficient Multimedia Retrieval*, DOI: 10.1007/978-3-642-40643-0_2

[Aston 2016] Aston, J; Girolami, M; Hohl, D and Király, F; *Big Data in Geoscience*, ATI Scoping Workshop Evaluation Report, Ref.No. C80, Alan Turing Institute, British Library, London, 2016.

[Atkinson 2013] Atkinson, MP. (2013) Data-Intensive Thinking with DISPEL, in [Atkinson 2013a]. doi: 10.1002/9781118540343.ch4

[Atkinson 2013a] MP. Atkinson, R. Baxter, M. Galea, M. Parsons, P. Brezany, O. Corcho, J. van Hemert and D. Snelling (eds) *The DATA Bonanza: Improving Knowledge Discovery in Science, Engineering, and Business*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2013, doi: 10.1002/9781118540343

[Atkinson 2013b] MP Atkinson and MI Parsons, *The Digital-Data Challenge*, in [Atkinson 2013a] Chap 1

[Atkinson 2015] M. Atkinson, M. Carpené, E. Casarotti, S. Claus, R. Filgueira, A. Frank, M. Galea, T. Garth, A. Gemünd, H. Igel, I. Klampanos, A. Krause, L. Krischer, S. H. Leong, F. Magnoni, J. Matser, A. Michelini, A. Rietbrock, H. Schwichtenberg, A. Spinuso, and J.-P. Vilotte, "VERCE delivers a productive e-Science environment for seismology research," in Proc. IEEE eScience 2015, 2015.

[Baldine 2010] Ilia Baldine, Yufeng Xin, Anirban Mandal, Chris Heermann Renci, Jeff Chase, Varun Marupadi, Aydan Yumerefendi, and David Irwin. 2010. Networked cloud orchestration: a GENI perspective. In GLOBECOM Workshops (GC Wkshps), 2010 IEEE, pp. 573-578. IEEE.

[Barbacci et al. 1995] Barbacci, Mario, Mark H. Klein, Thomas A. Longstaff, and Charles B. Weinstock. *Quality Attributes*. No. CMU/SEI-95-TR-021. CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 1995.

[Beiskenr 2013] Stephan Beiskenr, Thorsten Meinl, Bernd Wiswedel, Luis F de Figueiredo, et al. KNIME-CDK: Workflow-driven cheminformatics, BMC Bioinformatics, 2013.

[Belhajjame 2015] K. Belhajjame, J. Zhao, D. Garijo, K. Hettne, R. Palma, O. Corcho, J.-M. Gómez-Pérez, S. Bechhofer, G. Klyne, and C. Goble, "*A Suite of Ontologies for Preserving Workflow- Centric Research Objects*," Journal of Web Semantics, 2015.

[Berman 2008] Berman, H. M. *"The Protein Data Bank: a historical perspective" (PDB). Acta Crystallographica Section A: Foundations of Crystallography* **A64** (1): 88–95 (Jan. 2008). doi:10.1107/S0108767307035623.

[Berners-Lee 2001] Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. The semantic web. Scientific American 284, no. 5, 28-37.

[Bhagat 2010] Jiten Bhagat, Franck Tanoh, Eric Nzuobontane, Thomas Laurent, Jerzy Orlowski, Marco Roos, Katy Wolstencroft, Sergejs Aleksejevs, Robert Stevens, Steve Pettifer, Rodrigo Lopez[2] and Carole A. Goble, *BioCatalogue: a universal catalogue of web services for the life sciences*, Nucleic Acids Research, Vol. 28(2)

[Bier 2013] C. Bier: *How usage control and provenance tracking get together – a data protection perspective.* Security and Privacy Workshops (SPW), 2013 IEEE, San Francisco, CA, 2013, pp. 13-17.

[Blair 2012] Blair, Gordon, and Paul Grace. "Emergent middleware: Tackling the interoperability problem." *IEEE Internet Computing* 1 (2012): 78-82.

[Blankenberg 2011] Daniel Blankenberg, Gregory V. Kuster, Nathaniel Coraor, et al. Galaxy: A Web-Based Genome Analysis Tool for Experimentalists, In Current Protocols in Molecular Biology, 2001.

[Bloomberg 2014] Bloomberg, J. *Enterprise Architecture: Don't be a fool with a tool*. Forbes 7th August 2014. http://www.forbes.com/sites/jasonbloomberg/2014/08/07/enterprise-architecture-dont-be-a-fool-with-a-tool/#497c9b4e45f1.

[Bordawekar 2014] R. Bordawekar, B. Blainey, C. Apte (2014) Analyzing analytics. *SIGMOD Rec.* 42, 4 (February 2014), 17-28. DOI=http://dx.doi.org/10.1145/2590989.2590993.

[Bose 2005] Bose, R., Frew, J. *Lineage Retrieval for Scientific Data Processing: A Survey*. ACM Computer Surveys, Vol. 37, No. 1, 2005.

[Boulanger 2014] Boulanger, Damien, Benoit Gautron, Valérie Thouret, Martin Schultz, Peter van Velthoven, Bjoern Broetz, Armin Rauthe-Schöch, and Guillaume Brissebrat. "Latest developments for the IAGOS database: Interoperability and metadata." In *EGU General Assembly Conference Abstracts*, vol. 16, p. 6510. 2014.

[Bröring 2011] Bröring, A., Echterhoff, J., Jirka, S., Simonis, I., Everding, T., Stasch, C., Liang, S., and Lemmens, R. *New Generation Sensor Web Enablement*. Sensors 2011, 11, 2652-2699; doi:10.3390/s110302652.

[Budavari 2013] Budavari, T., Dobos, L. and Szalay, AS., *SkyQuery: Federating Astronomy Archives,* Computing in Science and Engineering 15(3): 12-20, 2013.

[Buneman 2000] P. Buneman, Khanna, S., Tan, W.-C.: Data Provenance: Some Basic Issues. Lecture Notes in Computer Science, Volume 1974, Foundations of Software Technology and Theoretical Computer Science, (FST TCS 2000), pages 87-93.

[Buneman 2001] P. Buneman, S. Khanna, and T. Wang-Chiew, *Why and Where: A Characterization of Data Provenance*, in Database Theory —ICDT 2001, vol. 1973, J. Bussche and V. Vianu, Eds. Springer Berlin Heidelberg, 2001, pp 316-330.

[Buneman 2016] Buneman, P.; S. Davidson and J. Frew, *Why data citation is a computational problem*, to appear, CACM 2016.

[Burns 2014] Burns, R., Vogelstein, JT. and Szalay, AS., *From Cosmos to Connectomes: The Evolution of Data-Intensive Science*, Neuron 83, 2014 http://dx.doi.org/10.1016/j.neuron.2014.08.045.

[Bux 2013] Marc Bux and Ulf Leser. Parallelization in Scientific Workflow Management Systems, CoRR 2013

[Camarinha-Matos 2006] Luis. M Camarinha-Matos, Hamideh Afsarmanesh, Martin Ollus (eds). *Network-Centric Collaboration and Supporting Frameworks*. IFIP TC 5 WG 5.5, Seventh IFIP Working Conference on Virtual Enterprises, 25-27 September 2006, Helsinki, Finland. Vol. 224. Springer Science & Business Media, 2006. ISBN: 0-387-38266-6.

[Candela 2013] Candela, L.; Castelli, D.; Coro, G.; Pagano, P. & Sinibaldi, F. Species distribution modeling in the cloud. Concurrency and Computation: Practice and Experience, Wiley, 2013, pp. 289-301 doi: 10.1002/cpe.3030

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

[Candela 2014] Candela, L.; Castelli, D.; Coro, G.; Lelii, L.; Mangiacrapa, F.; Marioli, V.; Pagano, P. An Infrastructure-oriented Approach for supporting Biodiversity Research. Ecological Informatics, Elsevier, 2014, doi: 10.1016/j.ecoinf.2014.07.006.

[Cao 2009] Cao, B. et al: *Semantically Annotated Provenance in the Life Science Grid.* SWPM'09 Proceedings of the First International Conference on Semantic Web in Provenance Management. Vol. 526, pp 17-22.

[Cartlidge 2012] Cartlidge, E., *Convictions leave Italy's civil protection in chaos*, Science, Vol. 338, No. 6107, 589-590, 2012.

[Chalmers 2014] Matthew Chalmers. Large Hadron Collider: The big reboot. Nature 514 (2014), 158–160.

[Charalabidis 2012] Charalabidis, Yannis, Marijn Janssen, and Olivier Glassey. "Introduction to cloud infrastructures and interoperability minitrack." In *2012 45th Hawaii International Conference on System Sciences*, p. 2177. IEEE, 2012.

[Chatzistergiou 2015] Andreas Chatzistergiou, Marcelo Cintra and Stratis D. Viglas, REWIND: Recovery Write-Ahead System for In-Memory Non-Volatile Data-Structures, PVLDB 8(5):49 497-508, 2015.

[Cheney 2007] Cheney L., Chiticariu L., Tan W-C. *Provenance in Databases: Why, How and Where*. Foundations and Trends in Databases, Vol. 1, No. 4 (2007) 379-474.

[Churches 2006] David Churches, Gabor Gombas, Andrew Harrison, Jason Maassen, et al. Programming scientific and distributed workflow with Triana services: Research Articles. Concurr. Comput : Pract. Exper., 2006.

[COM (2015) 192 final] COM(2015) 192 final "Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A Digital Single Market Strategy for Europe". 6th May 2015. http://eur-lex.europa.eu/legal-content/EN/NOT/?uri=celex:52015DC0192. Accessed 28th April 2016.

[SWD(2015) 100 final] SWD(2015) 100 final "Commission Staff Working Document: A Digital Single Market Strategy for Europe - Analysis and Evidence Accompanying the document Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions A Digital Single Market Strategy for Europe". 6th May 2015. http://eur-lex.europa.eu/legal-content/EN/NOT/?uri=CELEX:52015SC0100. Accessed 28th April 2016.

[Coro 2014] Coro, G.; Candela, L.; Pagano, P.; Italiano, A.; Liccardo, L. Parallelizing the execution of native data mining algorithms for computational biology. Concurrency and Computation: Practice and Experience, Wiley, 2014, doi: 10.1002/cpe.3435.

[Davidson 2008] Davidson, S.B., Freire, J.: *Provenance and Scientific Workflows: Challenges and Opportunities*. SIGMOD'08, Vancouver, Canada.

[De Roure 2009] David De Roure, Carole Goble, Robert Stevens, The design and realisation of the Virtual Research Environment for social sharing of workflows, Future Generation Computer Systems, Volume 25, Issue 5, May 2009, Pages 561-567, ISSN 0167-739X, http://dx.doi.org/10.1016/j.future.2008.06.010

[Deelman 2009] Deelman, Ewa, Dennis Gannon, Matthew Shields, and Ian Taylor. "Workflows and e-Science: An overview of workflow system features and capabilities." *Future Generation Computer Systems* 25, no. 5 (2009): 528-540.

[Deelman 2015] Ewa Deelman, Karan Vahi, Gideon Juve, Mats Rynge, Scott Callaghan, et al. Pegasus, a workflow management system for science automation, Future Generation of Computing Systems, 2015.

[DFT WG – RDA 2015] DFT WG–RDA, *RDA Data Foundation and Terminology – DFT: Results RFC*. Eds. Gary Berg -Cross, Raphael Ritz, Peter Wittenburg. Date: 29/06/2015. Consulted on: 04/03/2016. Available at: https://rd-alliance.org/system/files/DFT%20Core%20Terms-and%20model-v1-6.pdf.

[Dodds 2014] L. Dodds, G. Phillips, T. Hapuarachchi, B. Bailey and A. Fletcher, "Creating Value with Identifiers in an Open Data World". Report from Open Data Institute and Thomson Reuters, October 2014. Available at http://innovation.thomsonreuters.com/content/dam/openweb/documents/pdf/corporate/Reports/creating-value-with-identifiers-in-an-open-data-world.pdf

[Duerr 2011] R.E. Duerr, R.R. Downs, C. Tilmes, B. Barkstrom, W.C. Lenhardt, J. Glassy, L.E. Bermudez and P. Slaughter, "On the utility of identification schemes for digital earth science data: an assessment and recommendations". Earth Science Informatics, vol 4, 2011, 139-160. Available at http://link.springer.com/content/pdf/10.1007%2Fs12145-011-0083-6.pdf.

[Earle 2009] P. S. Earle, D. J. Wald, K. S. Jaiswal, T. I. Allen, M. G. Hearne, K. D. Marano, A. J. Hotovec, and J. M. Fee, "Prompt Assessment of Global Earthquakes for Response (PAGER): A system for rapidly determining the impact of earthquakes worldwide," US Geological Survey, Tech. Rep., 2009.

[Edwards 2010] Edwards P. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. MIT Press 2010 ISBN 978-0-262-01392-5.

[e-IRG White Paper 2013] e-Infrastructure Reflection Group (e-IRG) Whitepaper 2013. http://e-irg.eu/documents/10920/11274/e-irg-white-paper-2013-final.pdf.

[Elliot 2015] Liz Elliot, Fausto Guinchiglia, Gabor Bella and Dave Robertson, *Healthcare Data Safe Haven: Overview and Logical Architecture*, October 2015. EU Healthcare Data Safe Havens project. Personal communication.

[Enoksson 2009] Enoksson, Fredrik, Matthias Palmér, and Ambjörn Naeve. "An RDF modification protocol, based on the needs of editing Tools." *Metadata and Semantics*. Springer US, 2009. 191- 199.

[ESFRI 2016] ESFRI, "European Strategy Report on Research Infrastructures: Roadmap 2016". *ISBN: 978-0-9574402-4-1*, Mar 2016. http://www.esfri.eu/roadmap-2016. Accessed 19th April 2016.

[EU Parliament 2007] EU Parliament, "*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*, vol. 50, no. L108, April 2007.

[Falt 2014] Zbyněk Falt, David Bednárek, Martin Kruliš, Jakub Yaghob, and Filip Zavoral. Bobolang: a language for parallel streaming applications, HPDC '14, 2014.

[Ferris 2014] Ferris, Virginia. "Beyond "Showing What We Have": Exploring Linked Data for Archival Description", School of Information and Library Science of the University of North Carolina at Chapel Hill, 2014.

[Field 2013] Field, Suhr *et al.*, *Realising the full potential of research data: common challenges in data management, sharing and integration across scientific disciplines*. doi: 10.5281/zenodo.7636.

[Filgueira 2015] R. Filgueira, A. Krause, M. Atkinson, I. Klampanos, A. Spinuso and S. Sanchez-Exposito, *dispel4py: An Agile Framework for Data-Intensive eScience*, e-Science (e-Science), 2015 IEEE 11th International Conference on, Munich, 2015, pp. 454-464.

[Filgueira 2016] Rosa Filgueira, Amrey Krause, Malcolm P. Atkinson, Iraklis Klampano, *et al*. dispel4py: A Python Framework for Data-Intensive Scientific Computing, IJHPCA 2016.

[Filgueira 2016b] Rosa Filgueira, Rafael Ferreira da Silva, Amrey Krause, Ewa Deelman and Malcolm P. Atkinson, *Container orchestration for designing, testing and running data-intensive workflows*, in preparation, 2016.

[Fox 2016] Fox, GC; Judy Qiu, Shantenu Jha, Saliya Ekanayake1 and Supun Kamburugamuve, Big Data, Simulations and HPC Convergence, Technical Report · January 2016, DOI: 10.13140/RG.2.1.1858.8566.

[Frank 2016] Frank, Anton, *In Need of Partnerships: Environmental Computing and European e-Infrastructures*, publisher Romanian Academy of Science, 2016.

[French 2015] French, SW. and Romanowicz, B., *Broad plumes rooted at the base of the Earth's mantle beneath major hotspots*, Nature vol. 525, 95, 2015. doi:10.1038/nature14876.

[Garijo 2014] Garijo, Daniel, Óscar Corcho, Yolanda Gil, Meredith N. Braskie, Derrek P. Hibar, Xue Hua, Neda Jahanshad, Paul M. Thompson and Arthur W. Toga. "Workflow Reuse in Practice: A Study of Neuroimaging Pipeline Users." eScience (2014).

[Garijo 2014a] Garijo, D., Gil, Y., Corcho O.: *Towards Workflow Ecosystems through semantic and standard representations*. Proceedings of the Ninth Workshop on Workflows in Support of Large-Scale Science (WORKS), held in conjunction with SC 2104, New Orleans, LA, November 16, 2014.

[Gallagher 2015] J. Gallagher, J. Orcutt, P. Simpson, D. Wright, J. Pearlman and L. Raymond, "Facilitating open exchange of data and information". Earth Science Informatics, Volume 8, Issue 4, pp 721-739, December 2015. Available via http://dx.doi.org/10.1007/s12145-014-0202-2.

[Ghijsen 2013] Ghijsen, Mattijs, Jeroen Van Der Ham, Paola Grosso, Cosmin Dumitru, Hao Zhu, Zhiming Zhao, and Cees De Laat. 2013. A semantic-web approach for modeling computing infrastructures. Computers & Electrical Engineering 39, no. 8, 2553-2565.

[Gölitz 2007] Gölitz, Olaf. "Distributed Query Processing for Federated RDF Data Management", PhD thesis, Universitat Koblenz-Landau, 2007.

[Gray 2007] J Gray, *eScience a transformed scientific method*, in [Hey 2009] pages xix to xxxiii.

[Hardisty 2016] Hardisty, A and Nieva de la Hidalga, A*., How the ENVRI Reference Model helps to design research infrastructures.* ENVRIplus project newsletter No.2, April 2016. http://www.envriplus.eu/wp-content/uploads/2016/04/ENVRI-Reference-Model.pdf

[Hardisty 2015] Hardisty, A., *Reference models: What are they and why do we need them?* https://alexhardisty.wordpress.com/2015/07/08/reference-models-what-are-they-and-why-do-we-need-them/. Accessed 18 April 2016.

[Hartig 2009] Hartig, O., *Provenance information in the web of data*, in Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009), 2009.

[Haydel 2016} Haydel, N., Madey, G., Gesing, S., Abdul Dakkak, Simon Garcia de Gonzalo, Taylor, I. and Wen-mei W. Hwu, *Enhancing the Usability and Utilization of Accelerated Architectures via Docker*, 2016. DOI: 10.1109/UCC.2015.57

[Hey 2009] T. Hey, S. Tansley, K. Tolle (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[Higgins 2006] Higgins, S. (2006). "Using OAIS for Curation". DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Handle: 1842/3354. Available online: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation

[Huber 2013] R. Huber, A. Asmi, J. Buck, J.M. de Luca, D. Diepenbroek, A. Michelini, and participants of the Bremen PID workshop, "Data citation and digital identification for time series data & environmental research infrastructures", report from a joint COPEUS-ENVRI-EUDAT workshop in Bremen, June 25-26, 2013. Available via http://dx.doi.org/10.6084/m9.figshare.1285728

[ISO 1997] International Telecommunications Union. 1997. ITU-T X.641, information technology—quality of service: framework.

ISO/IEC 10746-1:1998 Information technology -- Open Distributed Processing -- Reference model: Overview.

ISO/IEC 10746-2:2009 Information technology -- Open distributed processing -- Reference model: Foundations.

ISO/IEC 10746-3:2009 Information technology -- Open distributed processing -- Reference model: Architecture.

ISO/IEC 10746-4:1998 Information technology -- Open Distributed Processing -- Reference Model: Architectural semantics.

ISO/IEC 19793:2015 Information technology -- Open Distributed Processing -- Use of UML for ODP system specifications.

[ISO 2007] ISO. 2007. Geographic information—Metadata—XML schema implementation. ISO 19139:2007.

[ISO 2008] International Telecommunications Union. 2008. ITU-T E.800, definitions of terms related to quality of service.

[ISO 2009] ISO. 2009. Information and documentation—The Dublin Core metadata element set. ISO 15836:2009.

[ISO 2011] ISO. 2011. Geographic information—Observations and measurements. ISO 19156:2011.

[ISO 2014] ISO. 2014. Geographic information—Metadata. ISO 19115:2014.

[Jeffery 2006] Keith G Jeffery, Anne Asserson: '*Supporting the Research Process with a CRIS*' in Anne Gams Steine Asserson, Eduard J Simons (Eds) '*Enabling Interaction and Quality: Beyond the Hanseatic League*'; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp 121-130 Leuven University Press ISBN 978 90 5867 536 1

[Jeffery 2014] Jeffery, Keith, Nikos Houssos, Brigitte Jörg, and Anne Asserson. "Research information management: the CERIF approach." *International Journal of Metadata, Semantics and Ontologies* 9, no. 1 (2014): 5-14.

[Kacsuk 2014] P. Kacsuk, Ed., *Science Gateways for Distributed Computing Infrastructures: Development framework and exploitation by scientific user communities*. Springer International Publishing, 2014.

[Kelling 2013] Kelling, S, Fink, D, Hochachka, W, Rosenberg, K, Cook, R, Damoulas, T, Silva, C. and Minchener, W., *Estimating species distributions—across space, through time and with features of the environment*, in [Atkinson 2013a] Chapter 22, pp 441-458.

[Khalil 2013] Tawfiq Khalil, Ching-Seh (Mike) Wu, "Link Patterns in the World Wide Web", International Journal of Information Technology & Management Information System (IJITMIS), Volume 4, Issue 3, 2013.

[Klump 2015] J. Klump, R. Huber and M. Diepenbroek, "DOI for geoscience data - how early practices shape present perceptions". Earth Science Informatics Volume 9, Issue 1, pp 123-136, March 2016. Available via http://dx.doi.org/10.1007/s12145-015-0231-5.

[Kokkinaki 2016] Kokkinaki, A., Buck J. , and Darroch L., A semantically rich and standardised approach enhancing discovery of sensor data and metadata, EGU2016-12970, April 2016, http://meetingorganizer.copernicus.org/EGU2016/EGU2016-12970.pdf

[Koltsidas 2008] Ioannis Koltsidas and Stratis Viglas, Flashing up the storage layer, PVLDB 1(1), 514-525, 2008.

[Kozlovszky 2014] Miklos Kozlovszky, Krisztián Karóczkai, István Márton, Péter Kacsuk and Tibor Gottdank. DCI Bridge: Executing WS-PGRADE Workflows in Distributed Computing Infrastructures, Book Chapter 4, Science Gateways for Distributed Computing Infrastructures, 2014.

[Kyriazis 2008] Kyriazis, Dimosthenis, Konstantinos Tserpes, Andreas Menychtas, Antonis Litke, and Theodora Varvarigou. 2008. An innovative workflow mapping mechanism for grids in the frame of quality of service. Future Generation Computer Systems 24, no. 6, 498-511.

[Lebo 2014] Lebo, T., West, P., McGuiness, D.L.: *Walking into the Future with PROV Pingback: An Application to OPeNDAP using Prizms,* Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014.

[Li 2012] Li, Zheng, Liam O'Brien, Rainbow Cai, and He Zhang. "Towards a taxonomy of performance evaluation of commercial Cloud services." In *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*, pp. 344-351. IEEE, 2012.

[Li 2014] Feng Li, Beng Chin Ooi, M. Tamer Özsu, and Sai Wu. 2014. Distributed data management using MapReduce. *ACM Comput. Surv.* 46, 3, Article 31 (January 2014), 42 pages. DOI=http://dx.doi.org/10.1145/2503009

[Liew 2016] Chee Sun Liew, Malcolm P. Atkinson, Michelle Galea, Paul Martin,et al. Scientific Workflow Management Systems: Moving Across Paradigms, ACM Surveys, 2016.

[Lim 2010] C. Lim, S. Lu, A. Chebotko and F. Fotouhi, *Prospective and Retrospective Provenance Collection in Scientific Workflow Environments*, Services Computing (SCC), 2010 IEEE International Conference on, Miami, FL, 2010, pp. 449-456.

[Liu 2015] Ji Liu, Esther Pacitti, Patrick Valduriez and Marta Mattoso. A Survey of Data-Intensive Scientifc Worklow Management, Journal of Grid Computing 2015.

[Lopez 2009] Lopez, D. M., and Blobel, B., *A Development Framework for Semantically Interoperable Health Information Systems*. International Journal of Medical Informatics, Volume 78, Issue 2, Pages 83-103, February 2009. doi: 10.1016/j.ijmedinf.2008.05.009.

[Manvi 2014] Manvi, Sunilkumar S., and Gopal Krishna Shyam. "Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey." *Journal of Network and Computer Applications* 41 (2014): 424-440.

[Marru 2011] Suresh Marru, Lahiru Gunathilake, Chathura Herath, Patanachai Tangchaisin Marlon Pierce, et al. Apache airavata: a framework for distributed applications and computational workflows, ACM workshop on Gateway computing environments (GCE '11), 2011.

[Martin 2015] Martin, Paul, Paola Grosso, Barbara Magagna, Herbert Schentz, Yin Chen, Alex Hardisty, Wouter Los, Keith Jeffery, Cees de Laat, and Zhiming Zhao. "Open Information Linking for Environmental Research Infrastructures." In *e-Science (e-Science), 2015 IEEE 11th International Conference on*, pp. 513-520. IEEE, 2015.

[Martin 2013] Martin, P. and Yaikom, G., *Definition of the DISPEL language*, in [Atkinson 2013a] Chapter 10, 237-250, 2013.

[Martone 2014] M. Martone, ed., "Joint Declaration of Data Citation Principles", Data Citation Synthesis Group and FORCE11, San Diego CA, 2014. Available at https://www.force11.org/group/joint-declaration-data-citation-principles-final.

[Masson 2016] Jason Masson et al. \On the Complexities of Utilizing Large-Scale Lightpath-ConnectedDistributed Cyberinfrastructure", Concurrency and Computation: Practice and Experience, (2016) in press.

[Mattmann 2014] C. A. Mattmann, "*Cultivating a research agenda for data science*," Journal of Big Data, vol. 1, 2014.

[Mayr 2004] Mayr, Ernst (2004). *What Makes Biology Unique?.* Cambridge: Cambridge University Press. ISBN 0-521-84114-3

[Menychtas 2009] Menychtas, Andreas, Dimosthenis Kyriazis, and Konstantinos Tserpes. 2009. Real-time reconfiguration for guaranteeing QoS provisioning levels in Grid environments. Future Generation Computer Systems, 25(7), 779-784.

[Miled 2001] Miled, Zina Ben, Srinivasan Sikkupparbathyam, Omran Bukhres, Kishan Nagendra, Eric Lynch, Marcelo Areal, Lola Olsen et al. "Global change master directory: object-oriented active asynchronous transaction management in a federated environment using data agents." In *Proceedings of the 2001 ACM symposium on Applied computing*, pp. 207-214. ACM, 2001.

[Moreau 2008a] Moreau, L. et al: *Special Issue: The first provenance challenge*. Concurrency and computation: practice and experience. 2008: 20, 409-418.

[Moreau 2008] L. Moreau, P. Groth, S. Miles, J. Vazques-Salceda, J. Ibbotson, S. Jiang, S. Munroe, O. Rana, A. Schreiber, V. Tan and L. Varga, "The Provenance of Electronic Data". Communications of the Association for Computing Machinery (ACM), volume 51, number 4, April 2008. Available at http://faculty.utpa.edu/fowler/csci6174/papers/Reilly_provenaceCACM.pdf.

[Mork 2015] Ryan Mork, Paul Martin and Zhiming Zhao. Contemporary challenges for data-intensive scientific workflow management systems, WORKS '15, 2015.

[Motik 2006] Motik, B, I. Horrocks, R. Rosati, and U. Sattler, "Can OWL and Logic Programming Live Together Happily Ever After?", Proceedings 5th International Semantic Web Conference, 2006.

[Myers 2015] Myers, J; M. Hedstrom; D. Akmon; S. Payette; B. A. Plale; I. Kouper ; S. McCaulay; R. McDonald; I. Suriarachchi; A. Varadharaju; P. Kumar; M. Elag; J. Lee; R. Kooper and L. Marini, *Towards sustainable curation and preservation*, in Proc. IEEE eScience Conf. 2015, 526-535.

[Ngan 2011] Ngan, Le Duy, Yuzhang Feng, Seungmin Rho, and Rajaraman Kanagasabai. "Enabling interoperability across heterogeneous semantic web services with OWL-S based mediation." In *Services Computing Conference (APSCC), 2011 IEEE Asia-Pacific*, pp. 471-476. IEEE, 2011.

[Ortiz 2011] Ortiz, Sixto. 2011. The problem with cloud-computing standardization. Computer 7, 13-16.

[Park 2008] Park, U., Heidemann, J.: *Provenance in Sensornet Republishing*. J. Freire, D. Koop and L. Moreau (Eds.): IPAW 2008, LNCS 5272, pp. 280-292, 2008.

[Papapanagiotou 2016] Papapanagiotou, P, Dave Murray-Rust, and Dave Robertson, *Evolution of the Lightweight Coordination Calculus Using Formal Analysis*, in preparation and personal communication April 2016.

[Parsons 2010] M.A. Parsons, R.E. Duerr and J.-B. Minster, ”Data citation and peer review”, EOS, Transactions of the American Geophysical Union vol 91, no 34, 24 August 2010, 297-304. Available at http://modb.oce.ulg.ac.be/wiki/upload/Alex/EOS_data_citation.pdf.

[Patel 2006] J Patel, WTL Leacy, NR Jennings, M Luck, S Chalmers, N Oren, TJ Norman, A Preece, PMD Gray, G Shercliff, PJ Stockreisser, J Shao, WA Gray, NJ Fiddian, S Thompson. “CONOISE-G: Agent-based virtual organisations” , AAMAS '06 Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems pp1459-1460, 2006. doi: 10.1145/1160633.1160914

[Pérez 2007] Fernando Pérez, Brian E. Granger, *IPython: A System for Interactive Scientific Computing*, Computing in Science and Engineering, vol. 9, no. 3, pp. 21-29, May/June 2007, doi:10.1109/MCSE.2007.53. URL: http://ipython.org.

[Raddick 2014] Raddick, J., Thackar, AR., Szalay, AS. and Santos, RDC., *Ten years of SkyServer I: Tracking Web and SQL e-Science usage,* Computing in Science and Engineering 16(4): 22-31, 2014.

[Raddick 2014a] Raddick, J., Thackar, AR., Szalay, AS. and Santos, RDC., *Ten years of SkyServer II: How astronomers and the public have embraced e-Science*, Computing in Science and Engineering 16(4): 32-40, 2014.

[Rauber 2015] A. Rauber et al., “Data citation of evolving data. Recommendations of the Working Group on Data Citation (WGDC)”. Preliminary report from 20 Oct 2015. Available at https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf.

[Rauber 2016] A. Rauber, A. Asmi, D. van Uytvanck and S. Pröll, ”Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use”. Bulletin of IEEE Technical Committee on Digital Libraries, Special Issue on Data Citation, 2016. (In press; pre-print available from author.)

[Riedel 2009] Riedel, Morris, Erwin Laure, Th Soddemann, Laurence Field, John-Paul Navarro, James Casey, Maarten Litmaath et al. "Interoperation of world-wide production e-Science infrastructures." *Concurrency and Computation: Practice and Experience* 21, no. 8 (2009): 961-990.

[Robertson 2016] David Robertson, Luc Moreau, Dave Murray-Rust and Kieron O’Hara, An *Open System for Social Computation,* SOCIAM 2016.

[Robertson 2014] D. Robertson, L. Moreau, D. Murray-Rust and K. O’Hara. *An Open System for Social Computation*. In O’Hara, Nguyen and Hayes editors *Digital Enlightenment Yearbook: Social Networks and social machines, Surveillance and Empowerment*. IOS Press, 2014. ISBN 978-1-61499-449-7.

[Sahoo 2011] Sahoo S.S. et al: *A unified framework for managing provenance information in translational research*. Bioinformatics, 2001, 12: 461.

[Santana-Perez 2015] Idafen Santana-Perez and María S. Pérez-Hernández. Towards Reproducibility in Scientific Workflows: An Infrastructure-Based Approach, Scientific Programming, 2015.

[Santana-Perez 2016] Idafen Santana Perez PhD thesis “*Conservation of Computational Scientific Execution Environments for Workflow-based Experiments Using Ontologies”,* January 2016, at UPM (Madrid, Spain). http://idafensp.github.io/ResearchObjects/WICUS_Makeflow_Blast/ and http://www.sciencedirect.com/science/article/pii/S0167739X16000029

[Schwardmann 2015] U. Schwardmann, “ePIC Persistent Identifiers for eResearch” Presentation at the joint DataCite-ePIC workshop *Persistent Identifiers: Enabling Services for Data Intensive Research*, Paris, 21 Sept 2015. Available at https://zenodo.org/record/31785.

[Schuler 2014] Robert E. Schuler, Carl Kesselman, Karl Czajkowski, Digital asset management for heterogeneous biomedical data in an era of data-intensive science. BIBM 2014: 588-592.

[Siew 2016] Siew Hoon Leong, Antonio Parodi and Dieter Kranzlmüller, *A Robust Reliable Energy-Aware Urgent Computing Resource Allocation for Flash Flood Ensemble Forecasting on HPC Infrastructures for Decision Support*, under review.

[Simmhan 2005] Simmhan, J. L., Plale, B., Gannon, D.: *A survey of Data Provenance in e-Science*, SIGMOD Record, Vol. 34, No. 3, Sept. 2005.

[Simmhan 2016] Simmhan, Y., Ramakrishnan, L., Antoniu, G. and Goble, CA., *Cloud computing for data-driven science and engineering*, Conc. And Comp. Practice and Experience, 28(4):947-949, 2016.

[Simmhan 2009] Y. Simmhan and *et al.* Building the Trident Scientific Workflow Workbench for Data Management in the Cloud. In *ADVCOMP*. IEEE, October 2009.

[Socha 2013] Y.M. Socha, ed., “Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data”. Data Science Journal vol. 12, 13 Sept 2013. Available at https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_pdf.

[Spinuso 2016] Spinuso, A; Fligueira, R., Atkinson, M., Gemuend, A. *Visualisation methods for large provenance collections in data-intensive collaborative platforms*. In *EGU General Assembly 2016, - Information in earth sciences: visualization techniques and communication of uncertainty.* http://meetingorganizer.copernicus.org/EGU2016/EGU2016-14793-1.pdf.

[Stehouwer 2014] H. Stehouwer and P. Wittenburg, eds. "Second year report on RDA Europe Analysis Programme: Survey of EU Data Architectures", Deliverable D2.5 from the RDA Europe project (FP7-INFRASTRUCTURES-2012-1), 2015. Available at https://rd-alliance.org/sites/default/files/Survey%20of%20data%20mangement%20needs.docx.

[Swedlow 2016] Swedlow, J. "The challenges of leading The Open Microscopy Environment **(**OME**)** open-source project", Private communication, March 2016. http://www.openmicroscopy.org/site

[Szalay 2013] Szalay, AS., *From Large Simulations to Interactive Numerical Laboratories*, IEEE Data Eng. Bull. 36(4): 41-53, 2013.

[Szalay 2008] Szalay, AS: *The Sloan Digital Sky Survey and beyond*, SIGMOD Record, June 2008, Vol. 37, No. 2.

[Tan 2007] Tan, WC: *Provenance in Databases: Past, Current, and Future*, IEEE Data Eng. Bull.

[Taylor 2016] Taylor, I., *Experience building realiybid*, personal communication, Cardiff University, April 2016

[Taylor 2006] Ian J. Taylor, Ewa Deelman, Dennis B. Gannon, and Matthew Shields. Workflows for E-Science: Scientific Workflows for Grids. Springer-Verlag 2006.

[Tilmes 2010] C. Tilmes, Y. Yesha and M. Halem, "Tracking provenance of earth science data". Earth Science Informatics 3:59-65, Volume 3, Issue 1, pp 59-65, June 2010. Available via http://dx.doi.org/10.1007/s12145-010-0046-3.

[Turilli 2016] Matteo Turilli, Mark Santcroos, and Shantenu Jha. A comprehensive perspective on pilot-jobs, 2016. (under review) http://arxiv.org/abs/1508.04180.

[Uhlir 2012] P.F. Uhlir, rapporteur, "For Attribution - Developing Data Attribution and Citation Practices and Standards". Summary of an international workshop (August 2011), National Research Council, 2012. Available at http://www.nap.edu/openbook.php?record_id=13564.

[Vahi 2013] K. Vahi, M. Rynge, G. Juve, R. Mayani, and E. Deelman. Rethinking Data Management for Big Data Scientific Workflows. In *Workshop on Big Data and Science: Infrastructure and Services*, 2013

[Vianden 2014] Vianden, M., Lichter, H. and Steffens, A., *Experience on a Microservice-based reference architecture for measurement systems*, 21st Asia-Pacific Software Engineering Conf., 183-190, 2014.

[Weigel 2014] T. Weigel, T. DiLauro and T. Zastrow, "RDA PID Information Types Working Group: Final Report", Final report from the Research Data Alliance PID Information Types (PIT) Working Group, released on 2014-11-25, 25pp, http://dx.doi.org/10.15497/FDAA09D5-5ED0-403D-B97A-2675E1EBE786.

[White 2012] White, Laura, Norman Wilde, Thomas Reichherzer, Eman El-Sheikh, George Goehring, Arthur Baskin, Ben Hartmann, and Mircea Manea. "Understanding interoperable systems: Challenges for the maintenance of SOA applications." In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pp. 2199-2206. IEEE, 2012.

[Wilde 2015] Torsten Wilde, Tanja Clees, Hayk Shoukourian, Nils Hornung, Michael Schnell, Inna Torgovitskaia, Eric Lluch Alvarez, Detlef Labrenz, Horst Schwichtenberg: *Increasing Data Center Energy Efficiency via Simulation and Optimization of Cooling Circuits - A Practical Approach***.** D-A-CH EI 2015:208-221.

[Wilde 2011] Michael Wilde, Mihael Hategan, Justin M. Wozniak, Ben Clifford, Daniel S. Katz, and Ian Foster. Swift: A language for distributed parallel scripting, Parallel Computing 2011.

[Wolstencroft 2013] Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, et al. The Tavernaworkflowsuite: designing and executing workflows of Web Services on the desktop, web or in the cloud, PubMed 2013.

[Wolstencroft 2015] Wolstencroft, K., Owen, S., Krebs, O., Nguyen, Q., Stanford, N. J., Golebiewski, M., ... & Snoep, J. L. (2015). SEEK: a systems biology data and model management platform. BMC systems biology, 9(1), 33.

[Yu 2005] J. Yu and R. Buyya (2005) A taxonomy of scientific workflow systems for grid computing. *SIGMOD Rec.* 34, 3 (September 2005), 44-49. DOI=http://dx.doi.org/10.1145/1084805.1084814

[Zhao 2006] Zhao, Zhiming, Suresh Booms, Adam Belloum, Cees de Laat, and Bob Hertzberger. "Vle-wfbus: a scientific workflow bus for multi e-science domains." In *e-Science and Grid Computing, 2006. e-Science'06. Second IEEE International Conference on*, pp. 11-11. IEEE, 2006.

[Zhao 2010] Zhao, Zhiming, Paola Grosso, Ralph Koning, Jeroen Van Der Ham, and Cees De Laat. 2010. An agent based planner for including network QoS in scientific workflows. In Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on, pp. 231-238. IEEE.

[Zhao 2011] Zhao, Zhiming, Paola Grosso, Jeroen van der Ham, Ralph Koning, and Cees de Laat. 2011. An agent based network resource planner for workflow applications. Multiagent and Grid Systems 7, no. 6, 187-202.

[Zhao 2014] Zhao, Z, P. Grosso, C. de Laat, B. Magagna, H. Schentz, Y. Chen, A. Hardisty, P. Martin, and M. Atkinson. (2014) Interoperability framework for linked computational, network and storage infrastructures, version 2. Accessed: 2015-07-21. [Online]. Available: http://envri.eu/.