COMPUTATIONAL RESEARCH INFRASTRUCTURES FOR GLOBAL ENVIRONMENTAL CHALLENGES

Paul Martin[a], Yin Chen[b], Alex Hardisty[c], Keith Jeffery[d] and Zhiming Zhao[a,1]

[a]*System and Network Engineering, University of Amsterdam, Netherlands*

[b]*European Grid Infrastructure, EGI.eu, Netherlands*

[c]*Cardiff University, UK*

[d]*Natural Environment Research Council, UK*

[1] Corresponding author: Zhiming Zhao, z.zhao@uva.nl.

TABLE OF CONTENTS

ABSTRACT

Environmental science research is increasingly dependent on the collection and analysis of large volumes of data gathered via wide-scale deployments of sensors and other observation sources. Meanwhile researchers are being called upon to address global societal challenges that are inextricably tied to the stability of our native ecosystems. These challenges are intrinsically interdisciplinary in nature, forcing scientists to collaborate across traditional disciplinary boundaries. The role of research infrastructure in this context is to support researchers in their interactions with a host of different data sources and analytical tools, as well as with each other, but no single environmental research infrastructure can hope to fully encompass the entire research ecosystem that has arisen to support the study of environmental science. The challenge therefore is for new environmental research infrastructures to exhibit sufficient technical interoperability between the different services they offer so as to permit researchers to freely and effectively interact with the full range of research assets potentially available to them, allowing them to collaborate and conduct innovative interdisciplinary research regardless of the particular research community to which they belong. Realising this ideal however requires a broad understanding of the fundamental commonalities of environmental science research infrastructure services as well as the development and wide adoption of common foundational services. It also requires a pragmatic bridging between the different standards and controlled vocabularies currently in use or preparation by different scientific communities, a process that can be expedited by the use of a standard reference model and the use of a formal framework for semantically linking similar concepts in different contexts.

1. INTRODUCTION

Environmental science addresses both our understanding of the physical world and our relationship with that world. As humanity continues to expand and place increasing pressure on rapidly diminishing natural resources, a number of societal challenges have arisen that can only be addressed by innovative new science that combines knowledge and expertise drawn from across all environmental domains—whether they focus on the earth, oceans, atmosphere or biosphere. These challenges include managing and adapting to climate change, overpopulation, food security, disaster prevention and relief, and maintaining biodiversity in fragile but vital ecosystems.

One characteristic shared by all of these societal challenges is that they concern large, inter-connected systems of considerable complexity. Carbon emissions in one corner of the world can affect the global climate, changing rainfall patterns in another part of the world that then has impact on crop yields. The loss of species biodiversity causes a dramatic change in local ecosystems, resulting in mass die-offs and gradual desertification of wide regions. The shrinking of glaciers changes the temperature and composition of the oceans, affecting its interaction with the atmosphere. Natural disasters (e.g. rising sea levels, earthquakes or tsunamis) cause immediate chaos and death, but also affect trade networks and thus the global economy, which can lead to resource shortages, geopolitical instability, or even change how people interact with the land as they are forced to change livelihood, feeding additional ecological impacts. Forecasting or even simply extrapolating the potential consequences of these scenarios requires interdisciplinary, data-intensive science—science that interleaves theory, models and data in order to describe a combination of complex, closely-related systems that in the past have had to be considered, at least to some degree, in isolation.

It is only recently that the computational resources and data collection facilities have existed to permit large-scale analysis and integration of significant volumes of data from multiple sources in real time. Nevertheless, the environmental systems under study remain extremely complex, and the interactions across different environmental systems are still not fully understood. However, the problem is not simply one of scientific understanding, but also of how to support the practical integration of data and methods needed to develop this understanding. One distinction between the environmental sciences and some of the other data-rich fields such as astronomy and high energy physics is the diversity and spread of data sources, as well as the variety of forms that data can take. Whereas in those other fields the number of data sources are often (relatively) few in number and concentrated around large-scale research facilities, albeit with extremely rich data yields, the environmental sciences typically have a greater number of different data sources, many (but not all) yielding modest quantities of data individually, but huge amounts in aggregate. The challenge then becomes how best to integrate this data, and to provide the infrastructure necessary to do so. Environmental science has long been based on the collection and analysis of empirical data, but the quantities and scope of data being gathered via dedicated instruments and observations now outstrip the capacity of classical research methods. It is therefore necessary to support the development of a range of tools and services to be made available to researchers who wish to explore new sources of data. Scientists are also being challenged to collaborate on a global scale across traditional domain boundaries to discover and interact with data from many different, sometimes unfamiliar, research contexts. It is also necessary therefore to provide new research environments that support cross-disciplinary collaboration, allowing researchers to share new approaches to data analysis and integration, and take advantage of their peers' expertise and technical knowledge.

Environmental science research infrastructures aggregate technical infrastructure with standardised practices for data handling and experimentation in order to support a particular range of environmental science research activities. Research infrastructures typically integrate large-scale sensor and observer networks with dedicated data curation facilities, data dissemination and analysis tools, and other research assets. Examples of research infrastructures in Europe include LifeWatch (for biodiversity)[2], EPOS (solid earth sciences, including seismology, volcanology and geodesy)[3], ICOS (carbon science, in terrestrial, ocean and atmospheric domains)[4], and EMSO (ocean/marine science)[5]. These infrastructures are being developed specifically to become important pillars of research and fulfil specific roles within their respective user communities. However to fully address global environmental challenges, it is important that all research activities be well integrated in order to enable data-intensive system-level science (Foster and Kesselman 2006). This requires there to be common policies, protocols and standards in order to realise the optimal coordination, harmonization, and integration of data, applications, and other services shared between research infrastructures. However, the complex nature of environmental science often results in the development of isolated environmental research infrastructures that meet only the immediate requirements and needs of a

[2] LifeWatch: http://www.lifewatch.eu/.

[3] European Plate Observing System: http://www.epos-eu.org/.

[4] Integrated Carbon Observation System: https://www.icos-ri.eu/.

[5] European Marine and Seafloor Observatory: http://www.emso-eu.org/.

specific research community, with very limited *interoperability* of data, data access mechanisms, and data processing tools. Interoperability is key to streamlining the process of interdisciplinary research—ensuring that common standards and interfaces are used as widely as possible allows datasets, tools and services to be composed in innovative and unexpected ways with the minimum of additional engineering. Such technological inter-compatibility encourages inventive research by removing unnecessary technical barriers. This why forums such as ENVRI[6], EarthCube[7], RDA[8], Coopeus[9] and others have been established to identify and support certain common operations shared by different research infrastructures in the different domains. By accelerating the construction of standardised solutions for technical problems common to research infrastructure (such as to do with data identification, citation, cataloguing, and curation), the hope is that there will be a maturation of common policy and a wider adoption of standards that enhance technical interoperability between different infrastructure initiatives.

The integration of standards and best practices requires a strong formal understanding of the architecture, norms and processes of research infrastructure at the social, physical and technical levels. The focus here is on ‘computing’ infrastructure—data archives, online services, networks, etc. The construction of a standard model of computational research infrastructure—a ‘reference model’—should provide a shared taxonomy of concepts by which to understand different aspects of such infrastructure, and also provide a common basis for understanding the different standards, specifications and schemas currently used to describe data, services, processes and policies relating to environmental science research. Controlled vocabulary is essential if we wish to describe and implement interoperable services. Therefore, we need to consider how to construct a ‘semantic linking framework’ that can guide the process of integrating different controlled vocabularies and translating between them where necessary. This chapter discusses some of the essential characterisation of research infrastructures from the technological, computational perspective. It identifies some of the key technical services needed to realise technological interoperability between different research

[6] Common Operations of Environmental Research Infrastructures: http://envriplus.eu/.
[7] EarthCube: http://earthcube.org/.
[8] Research Data Alliance: http://rd-alliance.org/.
[9] Coopeus: http:///www.coopeus.eu/.

infrastructures. It describes how the construction of a standard reference model for environmental science research infrastructures might assist current and future infrastructure developments. It also describes how the development of a semantic linking framework can be used to enhance interoperability by bridging the semantic gap between the many different vocabularies used in different scientific disciplines to characterise environmental data and processes, and we discuss how such a framework might also help the development of the interoperable services identified earlier.

2. CHARACTERISING RESEARCH INFRASTRUCTURES

Modern scientists interact with a host of resources in order to do their work, including instruments, databases, analytical tools and simulation platforms. Regardless of the different methods that they might apply, their research efforts can be thought of in terms of a series of interactions between different actors and resources. The role of research infrastructures then is to support researchers in the conduct of their research by materially supporting a subset of these interactions.

'Research infrastructure' is thus a term that can be used in a broad range of contexts—technically speaking, a research infrastructure is simply a deployment of technologies or practices that support a set of research activities conducted by a group of researchers. At a more practical level however, research infrastructure commonly refers to the technical integration of large-scale data collection with data curation and data processing facilities behind a unified service interface (e.g. a single data portal for accessing datasets). They exist to organise the facilities and technologies needed to provide researchers with the means to interact with a particular collection of data, tools and services, as well to strengthen the community that exists around it.

From the societal perspective, research infrastructures augment or extend the primarily interpersonal research networks that exist among researchers, laboratories and other organisations, typically by connecting researchers to useful computational services or data. Our concern here is mainly with *computational* infrastructure—infrastructure that provides computational tools and services and associated informatics that e.g. support the discovery of online datasets, the execution of data mining processes and the transfer of data over electronic networks. Such infrastructure comes in many forms:

fundamental ‘e-infrastructure’ for computation, storage and networking; domain-specific infrastructure providing services and tools of interest to specific scientific disciplines; and virtual research environments for improving coordination and collaboration among researchers. However, all such technological infrastructure exists in the context of pre-existing social structures that form the basis for the very collaborations that lead to development of technological infrastructure in the first place. Llewellyn-Smith (2011) shows an increase in global collaboration and networking in research since the turn of the last century and makes the case for further enhancing transnational cooperation. Wagner (2009) argues that the scientific world is now best characterised by self-organising networks of researchers who collaborate “not because they are told to but because they want to”. These social research networks are often invisible to policy makers, their impact not always formally recognised, but nonetheless represent a critical informal structure for global research and knowledge sharing. With that in mind, it is foolish to ignore (or worse interfere with) existing social networks when identifying the research interactions that a research infrastructure might enable.

Likewise, certain research actions can be automated, but certain other actions can only currently be conducted with human expertise—for example, the orchestration and deployment of data analyses on computational hardware is inherently amenable to automation, but the expert selection of specific statistical analyses can generally only be accomplished through the guidance and intuition of a trained scientist. Thus there is an inherent limit to the extent to which technology can be integrated into research practice. Nevertheless, to the extent that technology *can* be integrated, it behoves us to consider how best to synthesise computational research infrastructure to best support research communities.

[Figure 1—Subsystems in environmental research infrastructures.]

There are a number of different ‘configurations’ of environmental science research infrastructures, depending on the particular kinds of research activity constitute their primary focus. However, there are common elements found with varying degrees of emphasis in most infrastructures. In (Chen et al. 2013a) it is posited that environmental science research infrastructures can be functionally decomposed into five distinct subsystems of data acquisition, data curation, data access, data processing and community support. Figure 1 illustrates the relationship between the five subsystems and the broader user community, the underlying technical resources enlisted by the infrastructure, and the accumulation of observations and measurements in the field or in laboratories. (Chen et al. 2013a) goes on to make a distinction between **large-scale observatory systems**, which focus on acquisition and curation of data from a specific collection of instruments or other observation sources, and then on how best to allow research communities to access that data, and **comprehensive integration infrastructures**, which focus on providing unified platforms for data processing and community support, often on behalf of a number of observatory systems within a specific domain. This distinction embodies two different (but overlapping) perspectives on research infrastructure: infrastructure that is constructed around *instruments* (sources of scientific observations and measurements), and infrastructure that is constructed around *services* (platforms for data discovery, analysis and integration). The challenges for the former include how to handle the real-time ingestion of data from the instrument network, its packaging and its curation, and how to provide timely access to that data and its derived products to its chosen user community. The challenges of the latter focus on how to provide access to computing resources and code, how to facilitate more complex experimental workflows, and how to help users disseminate results both formally (e.g. in academic publications) and among their peers (for collaboration purposes) in a manner that supports verifiability and reproducibility. Both models are concerned with provenance (where data came from and what was done with them on the way), availability (short and long term) and discoverability (data and services exist to be used, and that requires researchers to know about them).

Some large-scale observatory systems focus on providing dedicated support for a 'single' (albeit sometimes physically distributed) instrument, very similar to examples in astronomy (e.g. LOFAR[10]) or high-energy physics (e.g. the Large Hadron Collider[11]). The principal challenge for these infrastructures is to handle the influx of data from the main instrument, and to pass it on (modulo

[10] Low-Frequency Array for Radio astronomy: http://www.lofar.org/.

[11] CERN: http://home.cern/

quality checking) to its target community. For example the EISCAT_3D infrastructure[12], currently in construction by the European Incoherent Scatter (EISCAT) Scientific Association at the time of writing, provides a three-dimensional incoherent scatter research radar to study the upper atmosphere and near-Earth space. The infrastructure itself consists of the antenna arrays, the signal processing system, the network, and the data distribution system. The beam-formed sample data, together with data from the interferometry system and some high-volume data from other supporting instruments, are streamed to a large ring buffer designed to hold several days worth of data, after which the data will be overwritten. The ring buffer serves to store the raw data for long enough to allow it to be interpreted and processed. The interpreted data can then be transferred to a permanent data archive. Simultaneously, a second copy of the incoherent scatter data is separately passed through default signal processing in order to produce preliminary datasets needed for real-time experiments. The infrastructure must be able to cope with a data ingestion rate of at least 18 Gb/s per site, preferably scaling up to somewhere between 50-100 Gb/s per site.

Other large-scale observatory systems are more highly distributed. The challenge is much the same as for single instrument infrastructures, but in addition there is the difficulty of handling multiple sites and determining where it is most practical to process the raw data. Argo[13] is a global ocean observing system comprised of a large network of robot floats distributed across the world's oceans. The robot floats serve to monitor heat, salt transport, ocean circulation, and the ability of the ocean to absorb excess carbon dioxide from the atmosphere. Euro-Argo[14] is the European contribution to Argo. Euro-Argo as an infrastructure supports an array of around 800 floats (roughly a quarter of the global Argo deployment), providing enhanced coverage of oceanic conditions in European seas and providing researchers access to quality-controlled data via client services such as offered by Copernicus. Data collection in Euro-Argo is based on periodic communication with its robot float network. Every ten days, a float dives 2000 metres then rises to the surface to transmit data by satellite link. More than 200 such cycles can be performed during the float's four year lifespan. Data assembly (basically the packaging of raw information into useful, self-describing datasets) is performed at designated centres,

[12] EISCAT_3D: https://eiscat3d.se/

[13] Argo: http://www.argo.ucsd.edu/.

[14] Euro-Argo: http://www.euro-argo.eu/.

which receive data from satellite operators and perform automatic quality control. Corrected datasets are passed onwards and made available to selected researchers somewhere between 24 and 48 hours of original transmission, with general availability within 6 to 12 months of transmission. Data are also delivered to other regional sites, which perform more comprehensive and specialised analysis, including integration with other data sources.

Some large-scale observatory systems are both distributed and have a broad research focus, the unifying principle being the specific deployment of instruments providing the raw scientific data. What data to collect, and what to do with the data becomes as important as how to handle the data as it emerges (for the scientists themselves, probably more so). The European Multidisciplinary Seafloor and water-column Observatory (EMSO) is a European network of ocean observatories for the long term monitoring of environmental processes relating to climate change, ecosystems and general geo-hazards. The objective of the EMSO research community is to provide a sustainable framework for the investigation of the interaction between the geosphere, biosphere and hydrosphere based on (near) real time data transmission. In each EMSO observatory a common set of sensors for core measurements (including seismometers, hydrophones, magnetometers, gravity metres and pressure sensors) will be deployed together with a number of additional sensors for specific purposes as proposed by researchers. An additional source of data will be laboratory studies performed on material (e.g. sediment cores) collected at observatory sites by sampling devices. EMSO data is collected at a number of regional sites, locally stored and organised into catalogues. Some data will be harvested, archived and made available at a few specific data centres (e.g. PANGAEA[15]), with a single common portal providing access to data at all three sites. All data is harmonised according to standards provided by SeaDataNet[16], a network (and itself a research infrastructure) for providing integrated access to marine databases.

An example of a comprehensive integration infrastructure is LifeWatch. The difficulty for comprehensive integration infrastructures is identifying what facilities are available (fundamentally determined by the experimental sites contributing resources to the infrastructure) and defining a suite

[15] PANGAEA: http://www.pangaea.de/.

[16] SeaDataNet: http://www.seadatanet.org/.

of common functions or services to be supported across all facilities—for example supporting the execution of data analyses via a specific scripting language, supporting the browsing of data libraries via a specific protocol, or supporting the composition of executable workflows for doing multiple tasks in sequence or parallel. LifeWatch is an infrastructure for biodiversity and ecosystem research that provides standard data processing facilities over a range of existing data centres. The main objective of LifeWatch is to put in place essential infrastructure needed to provide an analytical platform for new and existing biodiversity data. The emphasis is on a distributed network of services providing secure access across multiple organisations and providing relevant analytical and modelling tools to research collaborations. Another example of a comprehensive integration infrastructure is AnaEE[17]. AnaEE places emphasis on providing high quality facilities for facilitating biodiversity and ecosystem experiments at selected physical sites (rather than a virtual research environment for analysing data per se), though it also seeks to provide platforms for computational research. In particular, AnaEE is intended to provide a unified framework for four major phases of environmental experimentation—in natura, in vitro, analytical, and modelling—so as to support deep investigation of data within a single context.

The distinction between large-scale observatory systems and comprehensive integration platforms is imprecise—many infrastructures combine aspects of both in practice. In the case of the European Plate Observing System (EPOS), which is intended to provide a unified research infrastructure for the solid earth sciences in Europe, the research infrastructure is essentially a comprehensive integration platform that provides a unified set of core services by which to interact with a confederation of large-scale observatory systems, or client infrastructures, in the form of existing national and international data centres and experimental facilities. The objective of the EPOS research community is to integrate the existing research infrastructures in seismology, volcanology, geodesy and other solid earth sciences in order to increase the accessibility and usability of multidisciplinary data—as such, EPOS encompasses a range of different data products. In seismology for example, real time seismic

[17] Analysis and Experimentation on Ecosystems: http://www.anaee.com/.

waveform data from more than 500 broadband stations in Europe are collected by VEBSN[18] and maintained within the European Integrated Data Archive (EIDA). A number of data centres (including ORFEUS[19] and EMSC[20]) provide quality control and archiving. All data is made available to researchers via a variety of means including web services and direct access. EPOS is intended to build upon such existing data acquisition and curation facilities by adding a new layer of interoperability between constituent infrastructures and presenting a standard interface to researchers—essentially trying to realise many of the goals of interoperable infrastructure within a single federated structure.

For comprehensive integration infrastructures, there is usually an existing set of large-scale observatory systems upon which additional research infrastructure is being constructed. Typically, these observatory systems are organisationally independent, participating voluntarily in the greater infrastructure in accordance with some agreed set of policies. It is therefore be useful to distinguish between 'integrated' infrastructures where all core resources are administered under one central authority, and infrastructures founded on collaboration among different authorities.

A common scenario where multiple organisations cooperate to build a research infrastructure is where key elements of the core infrastructure's operations are delegated to specialised sub-infrastructures, such as delegating long term data archival to a specialist data centre. Another scenario is where a dedicated data or computational infrastructure is enlisted to handle more immediate data curation and access, including services for cataloguing and data identification (and therefore citation resolution). For example, EUDAT[21] is an initiative concerned with the integration of data sources for research. It provides a number of key data management services to research communities, essentially allowing research infrastructures to delegate some data curation and access functions to EUDAT-affiliated institutions. Meanwhile EPOS is an example of an environmental science research infrastructure that is looking closely at how to use data infrastructure such as EUDAT to manage some of its long term data needs.

[18] Virtual European Broadband Seismograph Network: http://www.orfeus-eu.org/data/vebsn.html.
[19] Observatories and Research Facilities for European Seismology: http://www.orfeus-eu.org/.
[20] European-Mediterranean Seismological Centre: http://www.emsc-csem.org/.
[21] EUDAT: http://eudat.eu/.

A number of infrastructures demonstrate a tendency to separate operations by sub-discipline, especially where this reflects existing standards of practice. The Integrated Carbon Observation System (ICOS) is a research infrastructure specialised in quantifying and understanding greenhouse gas fluxes, its objective to harmonise the measurement of greenhouse gases across sites in Europe and increase the availability and accessibility of the data being collected. ICOS divides its operations by 'themes' of atmosphere, ecosystem and ocean, meaning that it can be thought of as an example of a federated infrastructure, depending on how much operational independence the individual themes are given. Each theme has its own dedicated thematic centre, which acts as coordinating site and common data centre for the network of observation sites contributing to the theme. EPOS, again, does a similar division between seismology, volcanology and several other specific solid earth science domains. This pragmatic decision to create a federated infrastructure allows independent segments of their respective research communities to plan and develop their activities and infrastructure based on their related but distinct needs, and ensures that infrastructure development is not unduly held back by issues and conflicts that can be localised to certain components of the overall architecture. On the other hand, this does mean that the need for standardisation of services and interfaces is particularly important in order to ensure interoperability within the infrastructure and to present a unified research infrastructure to the outside world.

Research infrastructures do not exist in isolation; rather, research infrastructure is an aggregation of services and networks, both technical and social, that augment the activities of researchers. Many self-identified research infrastructures overlap in function and resources with other sources of infrastructure to the extent that many data centres and research sites actively contribute to many different initiatives, each of which might present themselves as a 'research infrastructure'. For example, GBIF[22] is a distributed open data infrastructure for accessing species-level data about global biodiversity. EMSO contributes to GBIF, as does EMBRC[23], SeaDataNet (via the Archive for Marine Species and Habitats Data), Lifewatch, and many others. GBIF in turn, also contributes to LifeWatch. What is important then is not 'which initiatives do and do not constitute research infrastructures', but

[22] Global Biodiversity Information Facility: http://www.gbif.org/.

[23] European Marine Biological Resource Centre: http://www.embrc.eu/

how these different initiatives cooperate materially to support the ambitions of researchers, regardless of the organisational umbrellas under which particular resources belong.

As such, when we speak of ‘interoperable’ research infrastructure, what we are really interested in is the system of services that all of these different initiatives and facilities collectively provide, and the technical barriers that might prevent them from being easily exploited by a wider community. These barriers might be privileged access, poor documentation or customisability, or an inability to selectively retrieve specific results, but one of the most pernicious barriers is a lack of standardisation—that certain tools and services in one context simply do not work in the same way as equivalent tools and services in a different context, and as such require the would-be user to have to learn a new set of protocols and adapt their working practices in order to take advantage of them. Identifying commonalities between different research infrastructures and developing services according to a standard model where possible, or at least establishing intermediary brokers that can translate between different contexts, would significantly improve the technical interoperability of infrastructure.

3. REQUIREMENTS FOR INTEROPERABLE SERVICES

There are a number of research activities that a research infrastructure can support; the most important of these regard the curation of scientific data and tools, ensuring their long term availability and immediate accessibility, and with providing a full accounting of their provenance (basically where they came from) and use (both internally in order to evaluate their importance, and externally to foster trust in their validity as useful research assets). New environmental science research infrastructures need to provide a number of services to their respective communities if they want to maximise their usefulness, including but not limited to:

- **Resource discovery**. Every research asset, whether they be datasets, code, documentation, instruments or tools, should be catalogued, and the contents of these catalogues should be made available to researchers and be searchable via a variety of means.

- **Data harmonisation**. Data are recorded in different formats in different levels of detail, and are then catalogued in different ways. One of the roles of research infrastructure is to harmonise how data are collected, and provide a unified model of interaction with those data.

- **Data preservation**. Data stored should be kept in good condition, ensuring that the data is made accessible over the longer term. This entails not only maintaining backups of the data, but also maintaining documentation regarding the use and purpose of the data, and ensuring that the tools needed to interact with the data remain themselves available.

- **Provenance recording**. In order for researchers to correctly evaluate the relevance of different data to their research, it is necessary to be able to answer questions about where data have come from, how they were gathered, and what has been done to them (e.g. in terms of quality control and error correction).

- **Workflow composition**. There exist a variety of models, algorithms, code and services for processing data that a researcher might employ. Every computational experiment has a workflow that can be partially or wholly automated by research infrastructure. The use of automated workflows also assists in the reproduction of experiments, a key problem in verifying the results of peers.

- **Computational task deployment**. Where the infrastructure is used for processing, it is necessary to optimise the use of limited computing resources (including supercomputers, network bandwidth and temporary storage) for the tasks assigned. This requires information about both the application needs and the execution environments available.

- **Publication of research**. Publishing research outputs (whether in the form of papers, documentation, models, code or datasets) must be done in a way that permits citation, supports corroboration of experimental results, and is persistent. Research infrastructure can support this by hosting essential assets in a stable environment.

- **Accounting**. It is important to know how data and tools are used and by whom, making sure that the researchers who benefit from access to resources are actually using them, and ensuring that the benefits are recognised by the institutions ultimately providing the resources, as well as by the funding agencies that authorise funding for those institutions. Authenticating who is using an infrastructure, and where necessary providing authorisation for use of restricted resources, is also important in many cases, though this has to be balanced against the desire for open access to data for any interesting in pursuing research questions.

To be considered ‘interoperable’, these services should adhere to common standards wherever applicable and should present their functionality using standard interfaces that allow researchers to interact with them directly or as part of a more complex task workflow.

In the following sections, we consider in more detail some of the services that need to be developed and their requirements, focusing on the *provision* of research assets (curation, cataloguing and provenance), the *generation* of research outputs (processing, workflow and optimisation), and the *publication* of research outputs (identification, discovery and citation).

CURATION, CATALOGUING AND PROVENANCE

One of the most fundamental responsibilities conferred to research infrastructure is the reliable long term storage and preservation of accumulated research data, whether that be raw data extracted from instruments or observed in the field (or laboratory), or interpreted data obtained via analysis and processing. What distinguishes a modern e-research infrastructure from a simple data archive is how it makes that data accessible—an infrastructure should provide as clear a path as possible from the sources of data to scientists and their experiments, as well as provide a set of companion services that provide useful functions on data. This entails not merely storing the data, but also providing simple but effective mechanisms by which investigators can search data (and indeed other research assets) using various catalogues, and then retrieve those data to feed into computational processes.

Curation services therefore exist to support the curation of data at all points in the data lifecycle. They provide the means to ingest data gathered from the field or the laboratory into a curation framework.

The integration of data curation facilities with data acquisition networks results in continuous (or at least frequent) ingestion of new data over time that must be initially processed, quality checked and stored. Most environmental science research infrastructures have a ‘staggered’ curation sub-infrastructure, consisting (for example) of an initial ‘buffer’ of a certain capacity for newly-acquired data, followed by a regional data centre for ‘chunking’ data into discrete datasets (if necessary), quality processing and initial metadata annotation, followed by storage at a central facility for the medium-to-long term. Some experimental activities (e.g. volcanic monitoring) require access to new data almost immediately, necessitating the creation of dedicated pipelines for acquiring data as soon as it is available.

Once acquired, environmental data should be packaged into datasets based on geography, time, specific events or other criteria. These datasets should be annotated with characteristic information, including a unique identifier and a record of how the data was obtained. Datasets should be stored safely in some repository, and preferably replicated either to improve access (e.g. provide multiple redundant locations from which to retrieve data) or to ensure long term availability (e.g. store a copy of the data in an archive, with offline records produced on long-life media such as tape). Replica sites can be managed by the core research infrastructure, or can be farmed out to designated dedicated data infrastructure on the research infrastructure’s behalf. It is also important to institute policies for handling changes in storage and data formats over the long term, whether by preserving certain retrieval technologies or regularly migrating data to new models, so that resources remain accessible far into the future.

The provenance of curated artefacts is also important. Researchers need to know the source of data, the methods used to acquire them, the quality processes that they have been subjected to, and so forth. This information inspires trust in the infrastructure, ensures a degree of accountability, and promotes reproducibility of experiments. Provenance services should record the evolution of data by tracking each operation processed—such services have to be further developed, harmonized and integrated into existing and future research infrastructures. It is necessary to carefully consider how to integrate better provenance tracking within existing services and workflows however, as well as what tools are needed to work with the resulting provenance information. Provenance services need to trace the entire research data lifecycle from acquisition through curation through to processing. In a federated infrastructure, provenance recording also helps to correctly attribute the efforts of different participating institutions, which is important for their own reporting processes.

As already alluded to, data is not the only important asset maintained by research infrastructures. Code, documentation, instruments, tools, processes and other useful assets that contribute to the research process also need curation. The formulation of policies for preserving and ensuring access to those assets can be as fundamental a concern of research infrastructure development programmes as the curation of scientific datasets. Many of them have the same curation requirements as pure data—the need for persistent identifiers, rich metadata descriptions, and a high level of availability. The notion of 'research objects'—packages of information necessary for the reproducibility of research—is becoming especially important (Bechhofer et al. 2010).

Quality control and annotation can be considered essential curation services, but have differing requirements. Quality control focuses on identifying errors or gaps in data, and with flagging those errors and gaps, or even repairing them using extrapolation or default values where this does not unduly damage the integrity of the data (this is most common for continuous data streams rather than discrete measurements). Quality control is generally part of the standard workflow for fully ingesting newly-acquired data, and is often performed in near real time. Annotation focuses on enriching the metadata associated with a dataset and allowing researchers and other agents to make observations about the data and communicate those observations to other users of the data. The process of quality control may lead to annotation of the data, but annotation is general can be performed at any time while the data in question is under curation, at any point in its lifecycle after ingestion.

In a sense, interdisciplinary research begins with the drawing together of data from different sources. Interoperable data cataloguing allows for the discovery, access, retrieval and integration of data from multiple infrastructures, making it a key component in an interdisciplinary research environment. To support interoperability, it is imperative to ensure that datasets and other research assets are adequately prepared for use by various services. This entails associating substantive metadata, including provenance records, with every asset—for environmental science, rich metadata for geospatial semantic annotation is particularly important. In order to ensure that these metadata exist, scientists and technicians should be supported by a range of flexible services for automatic curation and semantic annotation so as to reduce the burden of producing metadata and thus increase the likelihood that good metadata practices are upheld. To realise this, curation services need to be developed that consider all of the different 'levels' of data (from raw to various degrees of interpretation, derivation and integration) and should comply with research-oriented standards such as OASIS[24] and INSPIRE[25] where relevant.

Meanwhile, a common data provenance service standard can provide data tracing services for data evolutions across different infrastructures. Standardized interfaces for querying, accessing and integrating provenance data can then be realized. In practice, linking all infrastructures to a single provenance service is not feasible, so instead the use of standards for provenance collection are applied in each infrastructure's dedicated provenance architecture so as to allow a distributed provenance network to emerge that can (in principle at least) be treated as a single unified service. Standardised interfaces for querying, accessing and integrating provenance data should be realised. Some degree of semantic linking is necessary to harmonise the key components and standards used for provenance and querying. W3C has embarked on the creation of provenance-oriented standards applicable to Web and other similar environments[26].

Being a provider of research assets is one of the primary roles of technical research infrastructure, but the key purpose of providing such assets is to allow researchers to use those assets to achieve some research output, such as by analysing datasets gathered from sensors. Oft times however the datasets provided by research infrastructure are difficult to process, either because of their large size, or because the processing necessary is challenging to configure and execute. Thus another important role adopted by many infrastructures is to provide facilities for computation close to the data itself (rather

[24] Organization for the Advancement of Structured Information Standards: https://www.oasis-open.org/.

[25] Infrastructure for Spatial Information in the European Community: http://inspire.ec.europa.eu/.

[26] PROV: http://www.w3.org/TR/prov-overview/.

than requiring all data to be transferred to a researcher's personal machine first) and access to pre-configured processing services (that can be trusted to produce accurate, high-precision results); this is our next concern.

PROCESSING AND OPTIMISATION

Environmental system-level science increasingly relies on large volumes of heterogeneous data as produced by various research infrastructures. Data processing services can make it significantly easier for scientists to aggregate data from multiple sources and to conduct a range of experiments and analyses upon those data, when those services are sufficiently well-designed and accessible. In principle, researchers can always retrieve data from data centres and perform any analysis they wish on that data using their own private facilities. In practice however, this poses a number of difficulties. For example, computation can be prohibitively expensive. Many analyses, especially deep analyses of large, co-dependent datasets, outstrip the capacity of desktop/laptop computers. High performance or high-throughput computation is not universally available, essentially locking out researchers whose sponsoring institutions have not had the foresight to invest in such facilities. Alternatively the facilities that are available may be overstretched, with time on them very limited. If certain processes are deemed valuable to the community at large, it should be made possible to acquire additional computational capacity within the auspices of a research infrastructure. Another problem is that data movement itself can be prohibitive, whether due to the size of datasets or limited bandwidth for network transfers. Doing comprehensive analysis and data mining on large datasets requires computational facilities and data to be brought together; traditionally, this entails bringing the data to the computer. With the (many) large datasets now being made available, the simple act of downloading all the available data needed to conduct a particular data-intensive process may by itself be hugely time consuming. Having computational facilities at the data centres, and scheduling processes there, removes the need to transport the data anywhere else—and research infrastructure initiatives can provide a framework for putting such facilities in place.

The environmental sciences are producers of ‘big data’—data that come in the form of a large number of varied datasets, many of which are themselves very large, or are generated very rapidly. In many cases these data are dispersed in small scattered datasets, which are updated frequently (with periods in minutes or even seconds). Parallelisation of computing tasks is often necessary to handle that update frequency, and to ensure that core data analyses and experiments can be performed on schedule, keeping up with the arrival rates of new data. In other cases the data arrive more slowly, with periods of hours or days, but to fully analyse and integrate all relevant datasets still requires extensive cross-correlation of data elements. This can also benefit from parallelisation.

In many cases researchers have to configure their own workflows. Writing code, preparing tools, and composing processes to realise a complete experimental pipeline requires time and considerable technical expertise. While many researchers are indeed very technically capable, this is still effort that is being diverted away from fundamental research and exploration of data. Some researchers will indeed be comfortable with, and wish to configure, their own workflows while exploring new methods. For many others however, support for common or fundamental tasks provided as a service by an infrastructure would greatly increase the efficiency of a research community—more so if there is support for the composition of tasks to create more complex workflows, along with the ability to share and reuse those workflows.

Moreover, there is a significant replication of common tasks. There are many standard processes that researchers in various scientific disciplines like to apply to certain kinds of data. If these processes are applied to a dataset within the scope of a specific research infrastructure, then the results of those processes can be shared with the community at large by the same infrastructure, avoiding a lot of unnecessary repetition of computation. It is sometimes difficult however to trust in the reliability of results produced by others, particularly if you are staking your own research on them. Processes conducted in private, isolated from oversight, may be subject to unknown flaws that cast doubt on the results then produced. Even should no indication of error be present, a scientist who wishes to use the results of some analysis on a dataset for their own research has to decide whether to put their trust in those prior results, or to repeat the analysis in their own environment. The citation of data, tools and methods in research addresses this concern to a degree, but research infrastructures can also address this concern in the trusted research environments that they offer to researchers. Computations performed within the auspices of an trusted research infrastructure can be annotated with metadata describing the provenance of the results, including pointers to data sources and to the specific methods and tools used to perform the computation, allowing investigators to make better judgements about the quality (and trustworthiness) of derived datasets.

Data processing services should make it easier for investigators to aggregate data from multiple sources and then perform systematic analysis on those data. Of increasing interest is how to support the entire lifecycle of computational experimentation by allowing researchers to take full advantage of the underlying e-infrastructure, being the computers and networks available for working with experimental data, available to them. Specific data processing services are often (but not always) domain-specific. However generic mechanisms and languages exist for enhancing the usability and integration of processing elements to support interdisciplinary system-level science. It is highly desirable that any provision of a data processing facility to deal with the requirements of research infrastructures reuses (to the greatest extent possible) tools already developed. In addition, the extensibility of processing services is of paramount importance—new algorithms, models and techniques need to be brought into any framework very easily to achieve significant impact. Workflow composition services focus on the engineering and technological aspects of managing entire lifecycles of computing tasks and application workflows for the efficient utilisation of underlying computational infrastructure. In particular, the service should enable scientists to enrich the data processing environment by easily injecting new algorithms to be also reused by others. There are a number of different workflow management systems designed for scientific computing, such as Pegasus[27], Taverna[28] and Kepler[29], though few have been integrated specifically into public research infrastructure. The use of 'big data' analysis tools such as Apache Hadoop[30] or Storm[31] can also

[27] Pegasus: http://pegasus.isi.edu/.
[28] Taverna: http://www.taverna.org.uk/.
[29] Kepler: https://kepler-project.org/
[30] Apache Hadoop: https://hadoop.apache.org/
[31] Apache Storm: http://storm.apache.org/

augment experimentation if effectively used—however the automation of experimental configuration remains difficult.

Flexible monitoring and diagnosis services for data processing allow researchers to verify that their experiments are operating as intended, and engenders trust in the system. By evaluating the characteristic experiments that researchers want to conduct, and developing common services, different possible avenues of optimisation can be identified. Much of this optimisation will be bespoke—custom solutions for specific problems. However there is also potential for generic optimisation, performed in advance or during runtime. Such generic performance optimisation focuses on mechanisms for making decisions about the deployment and orchestration of resources, services, data sources and potential execution infrastructures so as to increase the overall efficacy of the whole system, allowing agents to schedule the execution of environmental big data applications more efficiently. Service level agreements and modelling the infrastructure-level quality of service can augment this, allowing expert systems to make decisions on resources, services, data sources and potential execution infrastructures, and to then schedule the execution. Such services can extend existing optimisation mechanisms for resources, and provide an effective control model for applications at runtime. A semantic linking framework can support generic decision procedures at service, infrastructure and network levels, as well as provide effective mapping between application-level quality attributes onto infrastructure-level quality of service attributes of computing, storage and network. We consider some of the benefits of such a framework later in the chapter.

Given the generation of research outputs, it is very important that these outputs can be published in a manner that allows fellow researchers to verify, replicate and build upon them in order to further increase the body of available knowledge. It is also necessary that the raw assets used in the production of research outputs be made not only as accessible to researchers as possible, but also ‘publishable’, in the sense that interested parties can identify and refer back to them without needing particular knowledge of the research infrastructures which happen to provide them.

IDENTIFICATION, DISCOVERY AND CITATION

Research infrastructures support the activities of researchers by providing data, tools and services. However these assets are only useful if researchers are aware of their existence and find them sufficiently accessible. The capability to discover research assets is a significant problem in a global research context. The proliferation of research infrastructure presents a range of opportunities to the agile researcher, but these opportunities cannot be realised unless researchers are both made aware of the kind of assets available to them and are able to effectively seek them out on their own initiative. The role of discovery services in research infrastructure is to provide the tools needed by researchers to pull information about useful research assets on demand. To a lesser extent, notification services can also be used to push information to the researcher where it is deemed worthwhile to do so.

The ability to cite sources is fundamental to research. Statements can be verified, prior experiments can be replicated, and credit can be properly attributed. Increasingly it has become important to cite data, models (often in the form of code) and tools as well as prior research publications—this can be attributed to the massive increase in data volumes and the increasing complexity of data analysis, which has led to a state of affairs where, without the ability to retrieve the exact same datasets and analytical models and tools actually used in the research, there is little-to-no basis by which interested parties can actually validate the research of their peers and hold them accountable for their conclusions.

The discovery and citation of data and other assets relies on the ability to unambiguously identify objects. At the most fundamental level, this entails being able to describe the data to an extent that an agent familiar with the data can retrieve them on request. Given the vast quantities of datasets being handled, and the desire to automate basic curation functions, datasets are generally given their own unique name or identifier that can be used to recall the data on demand within a given context (such as a specific data centre or archive). Generally the 'names' of data have limited scope, only applicable within a single institutional context—when making data available publicly, it becomes important to try to ensure that it can be referred to using a genuinely unique identifier, so that conflicts with other similarly named data are avoided. Associating the object with another, more widely used namespace (such as used for URLs on the Web) can help with this. If the 'domain' of an identifier is unknown, then the identifier by itself may be insufficient for retrieving the data—the use of a globally unique identifier associated with a resolution service (that acts on behalf of a range of different data-carrying institutions) can assist with this. ePIC[32] is an example of an initiative that provides such identifiers as well as identifier resolution. Another system, used for scientific publications in particular, is the DOI system[33], which is also used by DataCite[34] to associate metadata with DOIs.

The principal role played by research infrastructures in the context of resource identification is simply to be the community-preferred place to find those resources. If the community knows that a given infrastructure maintains all the important research products in a given research sphere, then it can be used as the default portal for discovery of those products.

The ability to refer to data and the artefacts that allow for the manipulation of those data by citing their respective identifiers allows colleagues (and other agents) to retrieve research assets for themselves and provides a means to attribute those assets to the infrastructures and institutions responsible for making them available. Environmental research infrastructures integrate a large number of observational and experimental sites, administered by a variety of different institutions that are responsible for the operation, funding and maintenance of the different sites. It is often extremely important to these institutions that the research outputs produced using their resources is correctly attributed to them, as much for political and financial reasons as for scientific prestige. Thus any open access policy for data held by an infrastructure needs to acknowledge the source of the data and those responsible for making them available. It is also important that data providers are able to track the usage of their data, both to prove its importance and to refine their own understanding of how the data are used.

Optimisation of identification and citation models and technologies will be necessary because of the need to handle a truly vast number of different data objects—in a future where data, concepts, instruments and services are all citable, we can assume that there will be a corresponding explosion of persistent identifiers. The need to be able to efficiently and reliably resolve these identifiers and to

[32] European Persistent Identifier Consortium: http://www.pidconsortium.eu/.
[33] Digital Object Identifier: https://www.doi.org/.
[34] DataCite: https://www.datacite.org/.

direct investigators to the correct information artefacts is likely to become increasingly pressing. To identify resource use across federated infrastructures, or between interoperable infrastructures, it is desirable to implement common policy models for describing persistent identifiers for certain classes of data object, which can then be used to publish and cite data used in research. Several services for data identification (e.g. DataCite and ePIC) already exist, but there are still questions as to how best to apply them to the scientific process—for example, should different persistent identifiers be supplied for different versions of the same data set? Does a continuous data stream merit a direct identifier, or only the chunked output sets? Should raw data be given identifiers if in many cases they will rarely be accessed, or will be discarded after a few months, or should focus be given mainly to commonly-accessed derivative datasets? How should data generated during modelling or simulation be treated, given that it may (or may not) be more efficient to simply re-run the original process (which itself should be somehow citable)? At what point are researchers overwhelmed by a glut of persistent identifiers, and what is the role of data curators in what data should be permanently identified and what is designated 'limited access'?

In practice, any significant efforts to harmonise data citation requires collaboration with existing academic publishers, who wield considerable influence on current community behaviours and the effectiveness of citation mechanisms (which are currently focused on research paper citation, but increasingly involve generic citation mechanisms such as DOIs).

An interoperable data identification and citation service should aim to adhere common policy models for using persistent identifiers for publishing and citing data, and should use existing technologies where possible. It should furthermore be operated in close cooperation with existing initiatives such as RDA and the ICSU World Data System[35].

Resolving many of the issues associated with discovery, identification and citation requires common agreements among a range of different stakeholders, and many of the issues regarding the provision of research assets or the generation of research outputs also bear influence, because the internal composition of resources in a research infrastructure (whether technological or otherwise) determine

[35] International Council for Science World Data System: https://www.icsu-wds.org/.

what is often the 'simplest' or least invasive approach to (for example) assigning persistent identifiers to individual datasets, instruments, and other assets. As such, the adoption of standard architectures, taxonomies and other tools for describing research infrastructure can play a role in establishing a fundamental orthodoxy that makes many of these issues easier to resolve, and it is this that we now address.

4. BUILDING INTEROPERABLE COMPUTATIONAL INFRASTRUCTURES FOR RESEARCH

Interoperable tools which use standard APIs and can be used together in different configurations can make a huge impact on interdisciplinary research if made available to researchers. Interdisciplinary research after all relies on the integration of research processes founded in different research disciplines. If one accepts that the use of research infrastructure services is increasingly vital to expand the horizon of current innovation, then it is necessarily entailed that the integration of the experimental processes that are supported by those services can only happen if those services can be made to interact. Such interaction can be manually mediated by the efforts of technicians and (often) junior researchers, but this approach is both time-consuming and rarely generalizable to anything beyond the specific technologies being worked with. In essence, the time and intellectual capital of researchers is being increasingly diverted towards solving technical problems, rather than to genuine research.

It is unlikely that there will ever be a single unified research infrastructure for all aspects of science, yet the challenges humanity faces requires the ability to cross conventional scientific boundaries with a minimum of friction. Efforts are underway to consolidate within specific disciplines or areas of interest, to reduce the fragmentation of specific scientific communities. Nevertheless, we still need to accept that the needs of certain communities (as well as certain political realities) will always result in a degree of independence and technological drift. As such, we still need to be able to efficiently build bridges between different research infrastructures where the potential for interdisciplinary research exists. This requires a toolkit of interoperable data standards, protocols and service specifications that can be used to build the interoperability layers that must be inserted between technically-distinct infrastructures. Such interoperability layers are needed to streamline the interaction between data and services of different origins, automating where possible the establishment of pathways for interdisciplinary research, or at least simplifying the task of creating translation tools for combining specific services and data products.

The design, construction and maintenance of effective research infrastructure poses political, economic and technical challenges (Womersley 2010)—especially for primarily academic or research-oriented institutions—but these challenges are shared widely. In practice, considerable knowledge already exists regarding a range of issues typical to research infrastructure development. In that regard, it should be feasible to pool expertise already present in infrastructure projects in order to both share solutions and prevent the same mistakes from constantly recurring. To do so however, there need to be common forums for discussion and standard frames of reference (in terms of language and common understanding) by which to relate past experiences to new initiatives. In that regard, the foundation of interoperable architectures for research infrastructure must be a common model for research infrastructure that infrastructure developers and system architects can refer to. To improve cooperation and interoperability between infrastructure projects, attempts have been made to produce such a reference model for environmental science research infrastructures. The principal goals of such a model should be to capture high-level characteristics of operations common to environmental science research infrastructures, and to establish a lexicon for describing the parts and composition of such infrastructures to be used by research communities in future infrastructure development efforts.

Any archetypical model of environmental research infrastructure should not however exist in ignorance of the multitude of standards, protocols and policies already established for many of the operations of research infrastructure. Nor should it be ignorant of the current practices of existing infrastructure. In order to shape both the design and validation of a reference model, there should be a framework by which the concepts defined by the model can be related in terms of their semantics to relevant concepts articulated by different specifications, for example to link metadata concepts to the description of information flow in the lifecycle of a curated dataset. Such a semantic linking framework can then also be used to produce mappings between different controlled vocabularies (e.g. metadata standards, service descriptions, database schemas, etc.) that are needed to realise interoperability between different infrastructure services.

[Figure 2—Constructing interoperable research infrastructures.]

Given a well-defined reference model and semantic linking framework then, it only remains to consider how new services that fulfil the needs described in the previous section can be defined and deployed on suitable e-infrastructure. Figure 2 illustrates this approach, demonstrating how it cuts across the different key services needed by interoperable infrastructure described in section 3. The reference model and the semantic linking model inform the architecture design, which draws upon and informs the construction of all services. This architecture takes the requirements of the research infrastructure initiatives and the technologies provided by existing 'e-infrastructure' providers (being providers of storage, computational power and networking on demand), and guides the development of new services on top of those technologies and the adoption of those services by the research community.

The following sections argue for the rigorous modelling of infrastructure and consider some of the issues that must be faced in the course of such modelling.

REFERENCE MODELLING

Existing interoperability solutions mainly focus on specific levels of interaction: between infrastructures (Ngan et al. 2011), between middleware (Blair and Grace 2012), and between workflows (Zhao et al. 2006). Interoperation is typically achieved via iterative steps: building adapters or connectors between two infrastructures and then deriving new service layer models for standardization via community efforts. Such iterations can continuously promote the evolution of standards for infrastructures (and particularly those service layers), but will not completely solve all interoperability problems as long as the diversity between infrastructures remains great and there still exist missing links between standards (Riedel et al. 2009). Providing interoperability solutions only at a specific layer without a global view of the entire technology stack hampers the convergence of service layers. White et al. (2012) argued that an interoperability reference model is needed to complement the model of the application and infrastructure. This argument can be extended to the design of environmental science research infrastructures.

A reference model provides a framework for communicating complex concepts in precise terms, as well as a methodology for describing and rationalising the design and development of an instance of the modelled artefact. In recent years, the construction of a reference model for environmental research infrastructures has been seen as essential for developing the research field globally into one that can coherently address the inter-disciplinary challenges facing the Earth and society. This realisation has been partly driven by the rapid proliferation of new research infrastructures and the recurring problems that arise in their development and use.

Having a reference model to refer to during the development of a research infrastructure confers a number of benefits. For one, a reference model provides a common vocabulary for key concepts, helping a community to share and discuss ideas more efficiently and precisely. A reference model also helps a community converge on a single common vision by providing a means to clearly express it. A standard model can allow a proposed resource, service or technology to be evaluated in the context of the larger proposed infrastructure, making it easier to spot omitted functionality or violations of standard practice, and the exercise of fitting existing infrastructures into a standard model can make it easier to identify existing solutions to recurring problems. Finally a reference model helps to identify points at which interoperability has to occur, thus pinpointing the standards and protocols that might be applicable.

An example of a reference model developed for a specific research infrastructure is that of the LifeWatch Reference Model (Hernandez-Ernst et al. 2010), which provides guidelines for specification and implementation of the LifeWatch infrastructure. The LifeWatch Reference Model is built upon the ORCHESTRA Reference Model (Usländer et al. 2007), an architectural framework for distributed processing and geospatial computing, which is itself founded on the Reference Model for Open Distributed Processing (RM-ODP) (Linington et al. 2011). The approach taken for the

LifeWatch model was generalised for environmental research infrastructures in (Chen et al. 2013b, Zhao et al. 2015a)—this model decomposes 'research infrastructure' based on the five different viewpoints prescribed for distributed systems by RM-ODP (also known as ISO/IEC 10746). However the model cannot be deemed to be complete—not all viewpoints prescribed by ODP are addressed, and the validation of the model against real infrastructures is lacking. There is still need for a general reference model for environmental research infrastructure that encompasses the full scope of issues described earlier.

Nevertheless, the use of standards such as ODP that deconstruct complex systems by viewpoint seems to have merit. The fundamental idea that we can break down complexity by focusing on certain specific concerns in one context, with the presumption that any absent information will be present in another viewpoint, is appealing. It borrows from the idea of blueprints in construction and mechanical engineering. ODP in particular considers five viewpoints: *enterprise* (the interaction between agents in the system), *information* (the evolution and handling of information during execution of the system), *computation* (the decomposition and distribution of logical functionality in the system), *engineering* (the mapping of logical to physical resources as well as the data channels that exist between physical resources) and *technology* (the technologies and standards used by the system). In principle, there are other possible decompositions that could be used instead—however the fundamental idea (that of decomposition of complex systems by viewpoint) underpins one facet of semantic linking, as we describe below.

SEMANTIC LINKING

Interdisciplinary experimentation requires integration of data and methods from different scientific disciplines. These data and methods are increasingly being provided as part of dedicated research infrastructure. Different research communities have different working practices and use different technical standards to model data and processes, so technical incompatibilities often exist between datasets, tools and services deployed within different infrastructures. Composing an experimental workflow across research infrastructures often requires bespoke engineering to allow the different components to correctly interact with one another. Interdisciplinary data-intensive research therefore requires an understanding of all the workflow components (including data) involved in the activity—essentially the semantics (and pragmatics) of the different components needs to be understood before they can be made to interoperate. A formalised, standard vocabulary for shared concepts and processes can be used to define this understanding more precisely, and communicate it to others—such as provided by a reference model as described earlier. A generic, globally operational ontology that describes all aspects of research and computational infrastructures applicable in all contexts however is infeasible (not to mention cumbersome) to develop. Instead, the construction of interfaces for interoperability often depends on the ability to translate from one local controlled vocabulary to another, essentially ensuring that the inputs provided to various processes and services and the outputs extracted from those processes and services adhere to the expected formats, regardless of the actual provenance of those inputs and outputs. Providing a translation component between two different contexts can (and often has to be) done manually, especially if the tolerance for translation error is low. Nevertheless, if the vocabularies used in both the source and target contexts have been formally defined, it is at least possible to define a mapping between vocabularies that can then be used by a generic broker to manage the translation programmatically, rather than relying on custom brokers for every pairwise combination of foreign components.

The proliferation of semantic annotation of components, data and services e.g. in the form of Linked Open Data (Bizer et al. 2009) is founded on the principle that some of the burden of finding associations between disparate datasets and services should be taken off the shoulders of researchers (whose awareness of available datasets and tools may be unavoidably narrow), and taken on by the same discovery services that provide access to research assets in the first place by virtue of making it possible to automatically infer correspondences between the metadata attached to those assets. Semantic annotation provides the basis for semantic linking, the activity of providing translations between different concept-spaces so as to allow agents to reason about scientific (and other) data between as well as within specific semantic models.

Semantic linking is often investigated in the context of ontology matching, mapping or alignment (Ehrig 2007). 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.

Linking pairs of information models via semantic linking, allowing for incremental improvements to the interoperability of infrastructure components, may prove more pragmatic than waiting for a universal ontology to describe 'everything'. However a complete pairwise bridging of all information models used by all potentially interoperable research infrastructures is no more practical than the development of that one universal ontology, and moreover is unnecessary. Just as it is often necessary to use intermediate brokering for flexible service composition, it is necessary to use an intermediary concept model for semantic linking, reducing the number of required mappings between pairs of ontologies and their internal concepts (Martin et al. 2015).

The role of a semantic linking framework is simply to formalise the methodology for establishing semantic correspondences more efficiently, by allowing concepts in data models, specifications and other controlled vocabularies and languages to be linked via generic concepts defined by a core reference model, instead of to each other on a pairwise basis. By having a generic reference model for environmental science research infrastructure and associating as many existing standards for data, services and technologies to the concepts defined by the reference model as possible, it becomes easier to compare and indirectly link the standards themselves to one another, using the core reference model as a 'concept exchange'. This also serves the purpose of validating the reference model, as a complete model for environmental science research infrastructures can legitimately be expected to be able to describe most if not all of the concepts specified in any standard used by researchers in their experiments. Moreover, when the reference model takes the multi-viewpoint approach described in the previous section, there are additional benefits. Different standards and ontologies focus on different aspects of research and technology, so in principle there should be a natural fit between most models and a specific viewpoint (though models that straddle multiple viewpoints do exist). The internal correspondences between concepts in different viewpoints can therefore provide a means to find associations between models addressing different viewpoints via those correspondences, allowing the construction of a network of (indirect) concept relations. The coverage of a set of standards (e.g. as used collectively by a research infrastructure to describe all aspects of its operation) can be evaluated by how completely it maps to the set of concepts described across all viewpoints of the reference model.

The actual process of semantic linking between two concept ontologies (including between a dedicated vocabulary and a reference model) can be manual, automated or a mix of both. In any case it involves several iterations of the following steps:

1. **Pre-processing** of features by looking at a small set of excerpts from the overall ontology specification 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.
5. **Derivation** of the final linking between entities using different interpretation mechanisms, including the analysis of human experts.

Semantically linking information models from different environmental research infrastructures remains difficult however, even ignoring additional complications regarding (for example) multilingual research. The information resources (e.g. the datasets, documents and descriptions) from different infrastructures often do not share common vocabularies due to their individual idiosyncrasies coupled with the different contexts these information sources address. Moreover, the diversity of metadata standards used by different infrastructures (and in particular their potential evolution, extension or adjustment to address specific needs) make it costly to sustain and use any semantic linking model. To alleviate the cost, an effective linking model needs to focus on the interoperability gaps between research infrastructures, where semantic linking is most useful. It should also identify which viewpoints that specific concepts and concept models address, ensuring that the scope of a given concept, and thus what practical aspects of infrastructure design are influenced, are properly understood.

DEPLOYING NEW SERVICES ON E-INFRASTRUCTURE

Research infrastructure should either provide services that directly support key research activities, or interoperate with existing facilities already in use by research communities. In principle, these services should be:

- **Accessible**. The assets of a research infrastructure should be made as available as possible to a wide range of users, and any services should be made as simple as possible in order to allow users to integrate them quickly into their own working practices.

- **Accountable**. All resources should be properly annotated with appropriate metadata generated at all stages in an experimental workflow, so as to allow the provenance of research results to be traced from end product back to source.

- **Translatable**. A certain degree of semantic mapping may be required to bridge the operational gap between the different knowledge organising systems required by different scientific and technical domains, but tools and resource need to be formally documented in order to make this possible.

- **Adaptable**. Available resources change and user demands fluctuate; core research infrastructure services must be elastic and fault-tolerant, and provide programmatic interfaces for ad-hoc service composition.

- **Open yet secure***:* Although most research data is open, there is a need to protect more sensitive data, protect the privacy of researchers, attribute credit to individuals and organizations, embargo new research prior to publication and preserve authority and accountability constraints when transferring data between different technical and political domains.

Based on the demands collected from each domain—whether currently represented by a research infrastructure or not—requirements of research infrastructures and their individual current solutions should be characterized with consideration for underlying common technologies and engineering challenges. Common operations (covering general and overarching activities) are characterized in several iterative steps involving research communities, infrastructure developers and technology providers. A reference model developed for constructing research infrastructures can be applied in the design and implementation of cross-infrastructure common services as well. Such an approach is used to reduce risk; the risk of developing new services is ameliorated by pooling resources and drawing upon the expertise of a broader technical community. It is also used to maximise utilisation of e-infrastructures. There exist a number of initiatives that provide technological infrastructure, generally based on ‘Grid’ or ‘Cloud’ computing, which are intended to provide or host services for public research. Deploying new services on these e-infrastructures reduces the need to invest individually on new computational infrastructure and makes efficient use of prior public investments. Sharing responsibility and effort for the development of common services does not simply reduce risk, but also promotes the cross-pollination of ideas that leads to different infrastructure initiatives solving recurring problems in the same way, and using one another’s results to their mutual advantage, which maximises interoperability—the simplest way to ensure that the resources provided by different infrastructures interoperate is by using the same standards and technologies in the first place. Moreover, even for quite different datasets, processes and tools, if the means to interoperate with a standard service is developed internally by an infrastructure development, then it is simpler for semantic links to be developed with other artefacts in other infrastructures that have likewise have had an interoperation interface developed with the shared service.

Data-intensive approaches allow researchers to define assumptions, extract evidence and validate theories based on large quantities of observations, measurements, documents and other forms of data collected from a variety of possible sources. These approaches can only be effectively enabled however in the presence of a supporting ‘virtual research environment’, a kind of integrated desktop for common services acting as a virtual laboratory for researchers. Such virtual research environments should not only provide the necessary tools for searching, accessing and integrating data and software to realise the many different workflows that constitute scientists’ research activities, but should ideally also provide tools for enabling collaboration. Such environments must be underpinned by research infrastructure, essentially acting as the unified interface for all research activities.

Common services can be deployed in generic data infrastructures provided by publicly funded organisations such as EGI[36] and EUDAT, which can then operate them on behalf of specific research infrastructures. This approach aligns with current trends in the provision of computational infrastructure, especially grid-based (e.g. EGI), cloud-based (e.g. Helix Nebula[37] and EGI) and data-centric projects (e.g. EUDAT) (Jeffery et al. 2015), as well as the developments being proposed (and in some cases implemented) under the umbrella of community initiatives such as RDA. Similarly, infrastructures operated by commercial organisations can also be exploited.

Traditionally, research infrastructure has been built around data centres hosting data gathered by specific deployments of scientific instrumentation, or data gathered by field researchers. Most research infrastructure projects concentrate on the integration of multiple data centres behind a common service interface, or the standardisation of processes and data products in order to increase *internal* interoperability within a research community. However the development of truly interoperable research infrastructure requires support for complex application workflows that can be made available to any and all researchers under many different contexts (Mork et al. 2015). The process of brainstorming, planning and implementing data-intensive experiments has to be accomplished without having to acquire privileged access to limited resources. To realise this ideal of rapid innovation requires autonomous deployment and configuration of resources on demand, which is only feasible by enlisting scalable virtualised architecture such as that provided by the Cloud (e.g. via some intelligent workbench such as described in (Zhao et al. 2015b, 2015c)), and being able to

[36] European Grid Infrastructure: https://www.egi.eu/.
[37] Helix Nebula: http://www.helix-nebula.eu/.

optimise the movement and processing of data at a low level (and indeed the movement of code), e.g. by making use of programmable networks (Koulouzis et al. 2016).

Deployed services are ultimately validated only by how they are used by the research communities. Infrastructure development projects should define representative study cases by which to evaluate the utility of any prototype. Such study cases should be selected to preferably involve as broad a range of research interests as possible, across traditional disciplinary boundaries, and have a clear impact, such as the study of the mechanisms of carbon sequestration in the biosphere (Sedjo and Sohngen 2012). Any study case analysis should be articulated using a reference model, and any existing applicable standards (e.g. applying to the types of dataset typical used in the scenario) should be linked to relevant model concepts. Well-defined success criteria for this study case should be assigned and tracked alongside validation. The resources available or required on e-infrastructures should be identified, from network connectivity to data storage and processing capabilities. Data delivery to stakeholders should be optimized, and specific first test actions should be taken into account.

This process of fundamental modelling, creating a common conceptual vocabulary and understanding, semantic linking of applicable specifications and standards, and deployment of common operations to generic e-infrastructure is key to the construction of interoperable research infrastructure—with the ultimate goal of maximising the interoperability at infrastructure, service and application levels.

5. CONCLUSIONS

There is a duty of care that we have assumed for our world. To preserve our environment or to adapt it to our needs without unintended (and possibly disastrous) consequences, we need to be able to analyse and understand the hugely complex environmental systems that determine the state of our planet in so many different ways. These systems transcend the conventional boundaries of modern scientific disciplines, and so they demand extensive interdisciplinary collaboration by researchers of many different specialities using data drawn from a variety of sources. Such interdisciplinary collaboration requires interoperability of technology and information—the integration of disparate experimental methodologies necessitates an equivalent integration of data and processing across different operational contexts. To facilitate such interoperability, research infrastructures need to be constructed to be both extensible and flexible.

Extensible research infrastructure is needed to handle the integration of new experimental sites, new services and new data sources. It is tempting to define research infrastructure as beginning and ending with the integration of a specific set of facilities within a single organisational umbrella. In truth the range of research assets, experimental sites and data available to a research community changes continuously. Moreover, while there are reasons for individual institutions to retain some degree of identity in a network of research infrastructures (e.g. for proper attribution), the purpose of facilitating interoperation is to remove technical boundaries that might limit scientific experimentation. Thus the ability to integrate the research assets of neighbouring infrastructures in order to present an open, unified research infrastructure to users is essential.

The notion of flexible research infrastructure is simply a reinforcement of the notion of extensible infrastructure. Flexibility is required to take full advantage of available resources (including computational, storage and network resources), to provide redundancy (and therefore increased reliability), and to manage changes in infrastructure topology (whether that be due to redeployments of sensor networks, reconfigurations of services, or the setting up of temporary field laboratories for research, for example). One way by which flexibility of research infrastructure can be realised is via the adoption of Grid- or Cloud-specific e-infrastructure to provide selected services (such as for data storage or processing). Grid computing has been used by academic projects for over a decade to provide access to high performance computing and storage facilities. Cloud computing, which inherits many of the ideas of the Grid, provides elastic virtualised generic infrastructure for hosting a range of services with a minimum of prior planning and configuration. It can be used to provide agile on-demand experimental facilities for researchers without intensive prior negotiation, and as a technology it benefits from widespread industrial support and investment.

There is increasing interest on the part of environmental science research communities in exploiting high performance or high throughput computing (HPC and HTC respectively) as part of generic (i.e.

not installed for specific purposes) e-infrastructure. Building an open framework for data processing requires the integration of resources from many different infrastructures of different types. This includes not only the core research infrastructures, but also other generic 'data' or computing infrastructures that provide specific support services for data and processing. Access to such data and computing infrastructures are often moderated by umbrella organisations such as EUDAT, EGI and PRACE[38], which allow for unified brokerage of resources of various kinds from different physical sites. An example of the use of such e-infrastructure for data processing is to analyse and predict the spread of infectious diseases. Mosquito-borne infections resulting in diseases like West Nile Fever, Cikungunya, Dengue, Usutu and Sindbis have (re-)emerged in Europe during recent decades, the result of globalisation and climate change granting new opportunities for pathogens to colonise or re-establish themselves in new areas. Statistical correlation approaches such as species distribution modelling (SDM) are invaluable methods for predicting disease outbreaks. The Swedish LifeWatch portal[39] is used to provide high-quality biological data for mosquito species, while BioVeL[40] is used to access relevant environmental information and provide a series of ecological modelling algorithms. Finally, HTC resources provided by EGI are used to model a number of different climate scenarios for many different disease-carrying species.

To comprehend and plan complex, extensible and flexible research infrastructure requires a well-defined model of both the infrastructure and the context in which infrastructure exists (in terms of community, engineering, standards, and of course the research process itself). The various components and the different concerns of stakeholders can all be mapped out and represented by a multi-viewpoint model in a way that allows developers to identify common operations, recurring issues and gaps in their planning or implementation efforts. Such a reference model is also invaluable for disseminating the results to others and to generalise best practice and technology selection to be applicable to a number of different research infrastructures that might be defined using the same core model. If the model can be formally specified in a machine-readable format, then it can also be used to support semantic linking and other automated activities.

[38] Partnership for Advanced Computing in Europe: http://www.prace-ri.eu/.

[39] Analysportalen för biodiversitetsdata: https://www.analysisportal.se/.

[40] Biodiversity Virtual e-Laboratory: http://www.biovel.eu/.

A semantic linking framework provides a pragmatic means to support interoperability between data and services from different research infrastructures by guiding the construction of semantic mappings between different controlled vocabularies: metadata models, service specification standards, operational policies, etc. Such semantic mappings allow for analysis of the coverage of different models and specifications, but also facilitate the practical translation of data from one context to another. This potentially permits the construction of interoperability services between different operational environments, allowing (for instance) the construction of multi-infrastructure workflows by which researchers can conduct experiments using the resources and other assets made available by different e-infrastructures. In the absence of a unified research environment, such semantic linking may prove necessary for encouraging much-needed interdisciplinary research.

It is important for environmental science research infrastructures to embrace open data administration policies so as to provide additional support to researchers, as well as facilitate the contribution of research to governmental policy. Fundamentally, all of the advantages of semantic interoperability described so far can only be realised if as broad a church as possible of researchers from a range of institutions (e.g. universities, national research centres and industry) have access to the data and services provided by research infrastructure. Despite the challenges however, the prognosis for future environmental science research infrastructure is good. There is an increasing level of collaboration between different infrastructure development initiatives, and an increasing availability of dedicated computational infrastructure for generic data curation and processing, which can be adapted to the needs of research communities. An increasing recognition of the importance of data modelling, especially for cataloguing and tracing data provenance will lead to the introduction of better standards (and better adoption of those standards). In turn, given the support of semantic linking and other metadata management methodologies, this should allow the production of more unified service interfaces and greater interoperability, thus encouraging more system-level science to address the global environmental challenges that have motivated the construction of advanced environmental science research infrastructure in the first case.

ACKNOWLEDGEMENT

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreements No. 654182 (ENVRI^{PLUS} project), No. 676247 (VRE4EIC project) and No. 643963 (SWITCH project).

REFERENCES

Bechhofer, Sean, David De Roure, Matthew Gamble, Carole Goble, and Iain Buchan. "Research objects: Towards exchange and reuse of digital knowledge." *The Future of the Web for Collaborative Science* (2010).

Bizer, Christian, Tom Heath, and Tim Berners-Lee. "Linked data—the story so far." *Semantic Services, Interoperability and Web Applications: Emerging Concepts* (2009): 205-227.

Blair, Gordon, and Paul Grace. "Emergent middleware: Tackling the interoperability problem." *IEEE Internet Computing* 1 (2012): 78-82.

Chen, Yin, Alex Hardisty, Alun Preece et al. "Analysis of Common Requirements for Environmental Science Research Infrastructures." In *The International Symposium on Grids and Clouds (ISGC)*, vol. 2013. 2013a.

Chen, Yin, Paul Martin, Barbara Magagna et al. "A Common Reference Model for Environmental Science Research Infrastructures." In *EnviroInfo*, pp. 665-673. 2013b.

Ehrig, Marc. *Ontology Alignment: Bridging the Semantic Gap*. Springer-Verlag, 2007.

Foster, Ian, and Carl Kesselman. "Scaling system-level science: Scientific exploration and IT implications." *Computer* 11 (2006): 31-39.

Hernandez-Ernst, Vera, Axel Poigné, Jon Giddy et al. *LifeWatch deliverable 5.1.3: Technical construction plan (Reference Model).* LifeWatch, 2010.

Jeferry, Keith, George Kousiouris, Dimosthenis Kyriazis et al. "Challenges Emerging from Future Cloud Application Scenarios." *Procedia Computer Science* 68 (2015): 227-237.

Koulouzis, Spiros, Adam Belloum, Marian T. Bubak, Zhiming Zhao, Miroslav Živković, and Cees de Laat. "SDN-aware federation of distributed data." *Future Generation Computer Systems* 56 (2016): 64-76.

Linington, Peter F., Zoran Milosevic, Akira Tanaka, and Antonio Vallecillo. *Building enterprise systems with ODP: an introduction to open distributed processing*. CRC Press, 2011.

Llewellyn Smith, Chris, editor. *Knowledge, Networks and Nations: Global Scientific Collaboration in the 21st Century.* The Royal Society, 2011.

Martin, Paul, Paola Grosso, Barbara Magagna et al. "Open Information Linking for Environmental Research Infrastructures." In *e-Science (e-Science), 2015 IEEE 11th International Conference on*, pp. 513-520. IEEE, 2015.

Mork, Ryan, Paul Martin, and Zhiming Zhao. "Contemporary challenges for data-intensive scientific workflow management systems." In *Proceedings of the 10th Workshop on Workflows in Support of Large-Scale Science*, p. 4. ACM, 2015.

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.

Riedel, Morris, E. Laure, Th. Soddermann et al. "Interoperation of world-wide production e-Science infrastructures." *Concurrency and Computation: Practice and Experience* 21, no. 8 (2009): 961-990.

Sedjo, Roger, and Brent Sohngen. "Carbon sequestration in forests and soils." *Annu. Rev. Resour. Econ.* 4, no. 1 (2012): 127-144.

Thomas Usländer, editor. *Reference Model for the ORCHESTRA Architecture (RM-OA) V2 (Rev 2.1).* Orchestra Project, 2007.

Wagner, Caroline S. *The new invisible college: Science for development*. Brookings Institution Press, 2009.

White, Laura, Norman Wilde, Thomas Reichherzer et al. "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.

Womersley, John, editor. *Cost control and management issues of global research infrastructures.* European Commission, Directorate-General for Research, 2010.

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 Second IEEE International Conference on*, pp. 11-11. IEEE, 2006.

Zhao, Zhiming, Paul Martin, Paola Grosso et al. "Reference Model Guided System Design and Implementation for Interoperable Environmental Research Infrastructures." In *e-Science (e-Science), 2015 IEEE 11th International Conference on*, pp. 551-556. IEEE, 2015a.

Zhao, Zhiming, Paul Martin, Junchao Wang et al. "Developing and Operating Time Critical Applications in Clouds: The State of the Art and the SWITCH Approach." *Procedia Computer Science* 68 (2015b): 17-28.

Zhao, Zhiming, Arie Taal, Andrew Jones et al. "A Software Workbench for Interactive, Time Critical and Highly self-adaptive cloud applications (SWITCH)." In *Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on*, pp. 1181-1184. IEEE, 2015c.