Optimisation requirements

Introduction

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

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

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

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

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

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

Overview and summary of optimisation requirements

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

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

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

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

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

Processing

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

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

Provenance

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

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

Curation

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

In general, the principal concerns of curation are ensuring the accessibility and availability of research assets (especially, but not exclusively, data). High availability in particular requires effective replication procedures across multiple sites. It would be expedient to minimise the cost of synchronising replicas and to anticipate where user demand (for retrieval) is likely to be so as to minimise network congestion.

Cataloguing

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

Identification and citation

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

Optimisation methodology

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

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

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

If one of the goals of ENVRI+ is to provide an abstraction layer over a number of individual research infrastructures and a number of shared services that interact with the majority of those infrastructures to provide standardised solutions to common problems, then the embedding of knowledge at every level of the physical, virtual and social architecture that is a necessary result of this approach is essential to mediate and optimise the complex system of research interactions that are then possible.

Research Infrastructures

The following RIs contributed to developing optimisation requirements:

Euro-Argo: This RI is interested in: providing full contextual information for all of its datasets *(who, what, where, when, why, how);* local replication of datasets (to make processing more efficient); cloud replication of the Copernicus marine service in-situ data in order to make it more accessible to the marine research community.

IS-ENES2: This RI has an interest in: standardised interfaces for interacting with services; automated replication procedures for ensuring the availability of data across all continents; policies for the assignment of compute resources to user groups; funding for community computing resources.

SeaDataNet: The RI is interested in: minimising the footprint of marine data observation; helping automate the submission of datasets by data providers to the RI; balancing the trade-offs between centralisation and distribution for performance, control and visibility; tackling organisational bottlenecks.