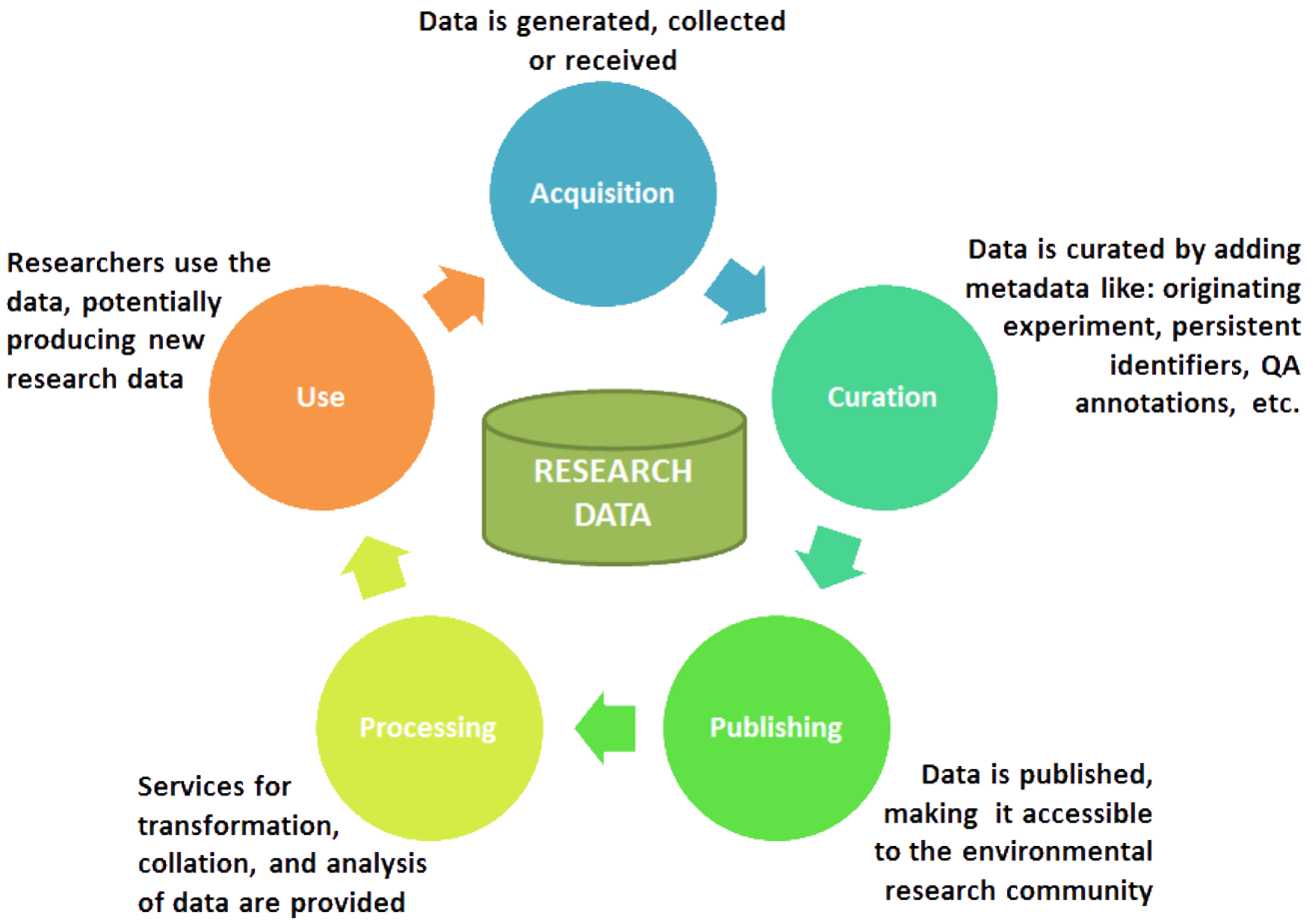
Model Overview

The Research Data Lifecycle within Environmental Research Infrastructures

The ENVRI and ENVRIplus project investigated a collection of more than 20 representative environmental research infrastructures (RIs) from different areas. By examining these research infrastructures and their characteristics, a common data lifecycle was identified. The data lifecycle is structured in five phases: *Data Acquisition*, *Data Curation*, *Data Publishing*, *Data Processing* and *Data Use* The fundamental reason of the division of the data lifecycle is based on the observation that all applications, services and software tools are designed and implemented around five major activities: acquiring data, storing and preserving data, making the data publicly available, providing services for further data processing, and using the data to derive different data products. This data lifecycle is fairly general and all research infrastructures investigated exhibit behaviour that aligns with its phases. Consequently, the ENVRI-RM is structured in line with the five phases of the data life-cycle.

This lifecycle begins with the acquisition of data from a network of integrated data collecting entities (seismographs, weather stations, robotic buoys, human observers, or simulators) which is then registered and curated within a number of data stores belonging to an infrastructure or one of its delegate infrastructures. This data is then made accessible to parties external to the infrastructure, as well as to services within the infrastructure. This results in a natural partitioning into data acquisition, curation and publishing. In addition, RIs may provide services for processing data, the results of this processing can then produce new data to be stored within the infrastructure. Finally, the broader research community outside of the RI can design experiments and analyses on the published data and produce new data, which in turn can be passed to the same RI or to other RI for curation, publishing and processing, restarting the lifecycle.

The activities of each research infrastructure can align with this lifecycle. However, research infrastructures will tend to optimise and concentrate more on some phases. For instance, some research infrastructures concentrate mostly on the acquisition of data, while others focus their expertise on curation or publishing. ENVRI RM assumes that the research infrastructures can complement and integrate with each other to support the entire data lifecycle. Integration is achieved through providing a set of capabilities via interfaces invoked within systems (or subsystems) which can be used within the infrastructures but also across boundaries. In the ENVRI RM, an interface is an abstraction of the behaviour of an object that consists of a subset of the interactions expected of that object together with the constraints imposed on their occurrence.



The Research Data Lifecycle

Data Acquisition

*In the **data acquisition phase** the research infrastructure collects raw data from registered sources to be stored and made accessible within the infrastructure.*

The data acquisition phase supports collecting raw data from sensor arrays and other instruments, as well as from human observers, and brings those data into the data management part (ie., ICT sub-systems) of the research infrastructure. Within the ENVRI-RM, the acquisition phase is considered to begin upon point of data entry into the RI systems. The acquisition phase as modeled in the ENVRI RM starts from the design of the experiment. Acquisition is typically distributed across networks of observatories and stations. The data acquired is generally assumed to be non-reproducible, being associated with a specific (possibly continuous) event in time and place. As such, the assignment of provenance (particularly data source and timestamp) is essential. Real-time data streams may be temporarily stored, sampled, filtered and processed (e.g., based on applied quality control criteria) before being ready for curation. Control software is often deployed to manage and schedule the execution and monitoring of data flows. Data collected during the acquisition phase ultimately enters the data curation phase for preservation, usually within a specific time period.

Data Curation

*In the **data curation phase** the research infrastructure stores, manages and ensures access to all persistent data-sets produced within the infrastructure.*

The data curation phase facilitates quality control and preservation of scientific data. The data curation functionalities are typically implemented across one or more dedicated data centres. Data handled at this phase include raw data products, metadata and processed data. Where possible, processed data should be reproducible by executing the same process on the same source data-sets, supported by provenance data. Operations such as data quality verification, identification, annotation, cataloguing, replication and archival are often provided. Access to curated data from outside the infrastructure is brokered through independent data access mechanisms. There is usually an emphasis on non-functional requirements for data curation satisfying availability, reliability, utility, throughput, responsiveness, security and scalability criteria.

Data Publishing

*In the **data publishing phase** the research infrastructure enables discovery and retrieval of scientific data to internal and external parties.*

The data publishing phase enables discovery and retrieval of data housed in data resources managed as part of data curation. Data publishing often provide mechanisms for presenting or delivering data products. Query and search tools allow users or upstream services to discover data based on metadata or semantic linkages. Data handled during publishing need not be homogeneous. When supporting heterogeneous data, different types of data (often pulled from a variety of distributed data resources) can be converted into uniform representations with uniform semantics resolved by a data discovery service. Services for harvesting, compressing and packaging data and metadata, as well as encoding services for secure transfer can be provided. Data publishing is controlled using rights management, authentication, and authorisation policies.

Data Processing

*In the **data processing phase** the research infrastructure provides a toolbox of services for performing a variety of data processing tasks. The scope of data processing is very wide.*

The data processing phase enables the aggregation of data from various sources, as well as conduct of experiments and analyses upon that data. During this phase data tends to be manipulated, leading to both/either derived and/or recombined data. To support data processing, a research infrastructure is likely to offer service operations for statistical analysis and data mining, as well as facilities for carrying out scientific experiments, modelling and simulation, and visualisation. Performance requirements for processing scientific data during this phase tend to be concerned with scalability, which can be addressed at the level of engineering and technical solutions to be considered (e.g., by making use of Cloud computing services). The data products generated during processing may themselves be curated and preserver within the RI.

Data Use

*In the **data use phase** the research infrastructure supports users of an infrastructure in gaining access to data and facilitating the preservation of derived data products.*

The data use phase provides functionalities that manage and track users' activities while supporting the users to conduct their research activities which may result in the creation of new data products. Data 'handled' and produced at this phase are typically user-generated data and communications. The data use phase requires supporting activities such as interactive visualisation, standardised authentication, authorisation and accounting protocols, and the use of virtual organisations. This is the most advanced form of data processing, at this phase the research infrastructure implements an interface with the wider world in which it exists.

Lifecycle Support Inter- and Intra- Research Infrastructure Relationships

Each research infrastructure supports the data lifecycle to a different degree. According to the scope of a particular research infrastructure, some core activities align strongly with some of the phases while other phases are not so comprehensively supported. In this case, the integration of the research infrastructures and their external supporting systems and services help in the overall fulfilment of the research data lifecycle. For these cases, the major integration points are those at the transition between phases of the data lifecycle. These integration points are important to build the internal subsystems of the research infrastructure, as well as to integrate the research infrastructure with other research infrastructures.

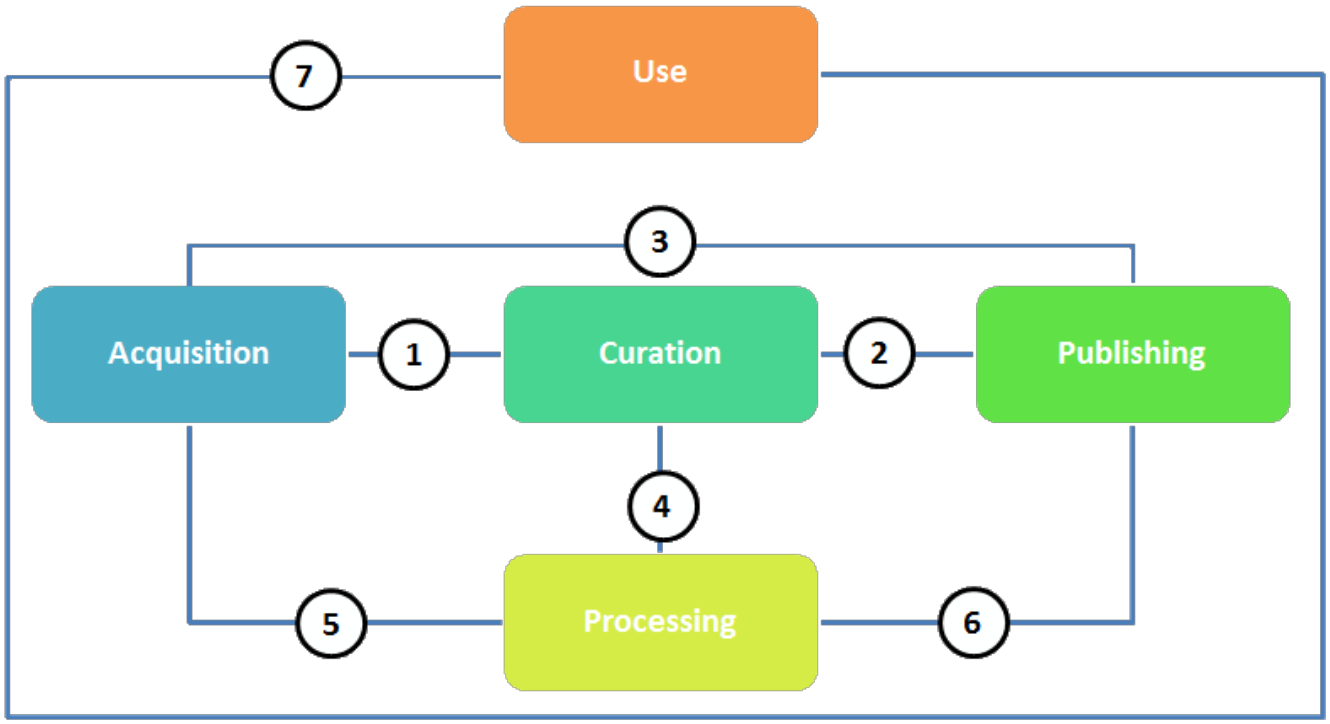


Illustration of the major integration (reference) points between different phases of the data lifecycle.

The integration points described as follows refer to the components supporting a phase of the data lifecycle. However, the components being integrated can be within the same research infrastructure or in different research infrastructures.

1. **Acquisition/Curation** by which components specialized in data acquisition are integrated with components which manage data curation.
2. **Curation/Publishing** by which components specialized in data curation are integrated with components which support data publishing.
3. **Acquisition/Publishing** by which components specialized in data acquisition are integrated components which support data publishing.
4. **Curation/Processing** by which components specialized in data curation are integrated with components which support data processing.
5. **Acquisition/Processing** by which components specialized in data acquisition are integrated with components which support data processing.
6. **Processing/Publishing** by which the components specialized in data processing are integrated with components which support data publishing.
7. **Use/All** by which entities outside the research infrastructure may be allowed to provide, access, or use data at different phases of the data lifecycle.

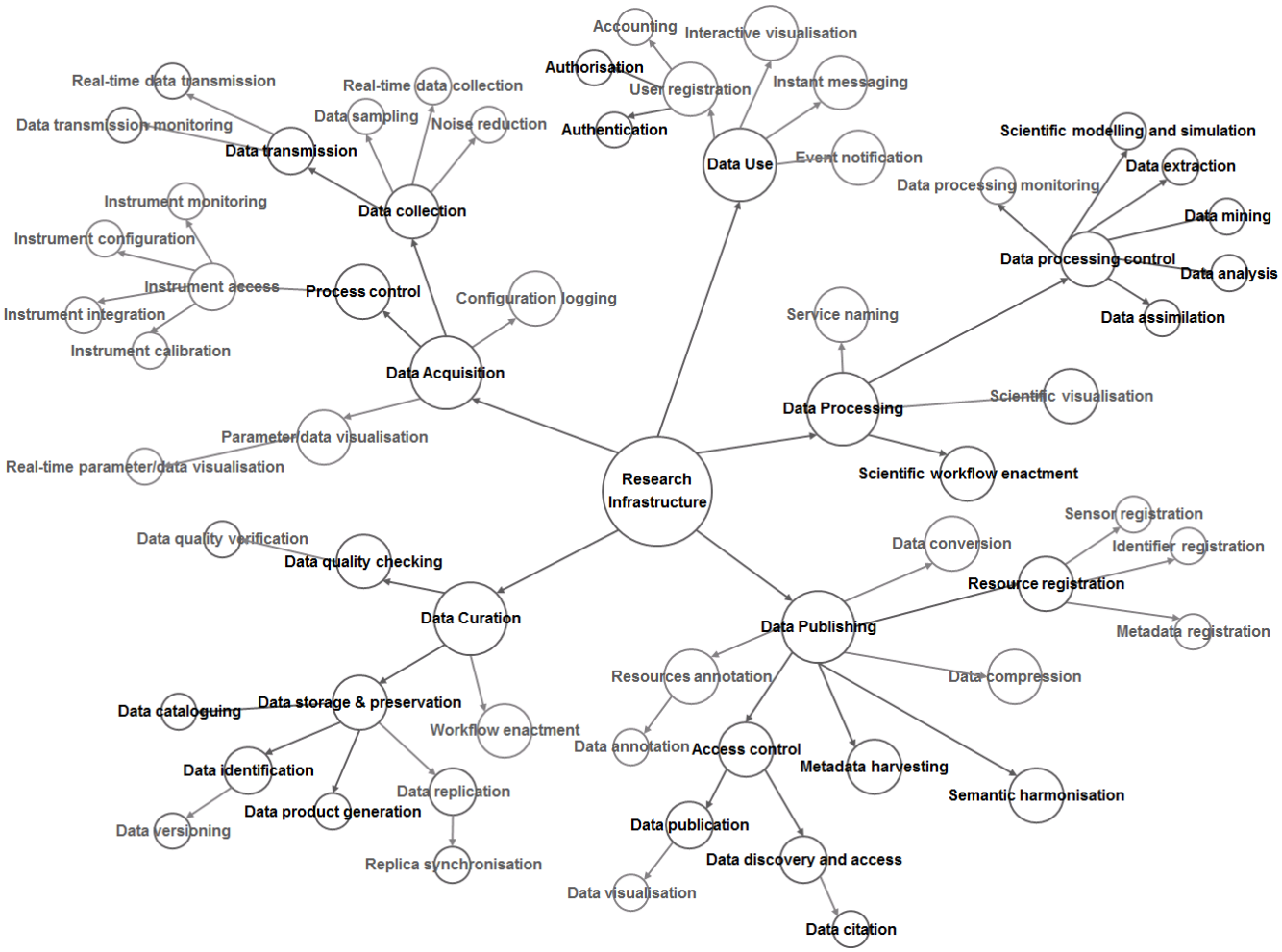
No notion of direction is implied in the definition of these points of reference. Relations with direction only appear when interfaces are superimposed on reference points, and then they can be unidirectional in either or both directions, or bidirectional - according to the nature of the interface(s).

Depending on the distribution of resources in an implemented infrastructure, some of these integration points may not be present in the infrastructure. They take particular importance however when considering scenarios where a research infrastructure delegates or outsources functionalities to other infrastructures. For example, EPOS and LifeWatch both delegate data acquisition and some data curation activities to national-level and/or domain-specific infrastructures, but provide data processing services over the data held by those infrastructures. Thus reference points 4 and 5 become of great importance to the construction of those projects.

Common Functions within a Common Lifecycle

Analysis of requirements of environmental research infrastructures during the ENVRI and ENVRIplus projects has resulted in the identification of a set of common functionalities. These functionalities can be classified according to the five phases of the data lifecycle. The requirements encompass a range of concerns, from the fundamental (*e.g.* data collection and storage, data discovery and access and data security) to more specific challenges (*e.g.,* data versioning, instrument monitoring and interactive visualisation).

In order to better manage the range of requirements, and in order to ensure rapid verification of compliance with the ENVRI-RM, a *minimal model* has been identified which describes the fundamental functionality necessary to describe an environmental research infrastructure. The minimal model is a practical tool to produce a partial specification of a research infrastructure which nonetheless reflects the final shape of the complete infrastructure without the need for significant refactoring. Further refinement of the models using the ENVRI-RM allow producing more refined models of designated priority areas, according to the purpose for which the models are created.



Radial depiction of ENVRI-RM requirements. The black labels correspond to the minimal model requirements.

The definitions of the minimal set of functions are given as follows (a full list of common functions is provided in Appendix A Common Requirements of Environmental Research Infrastructures):

(A) Data Acquisition

Process Control: Functionality that receives input status, applies a set of logic statements or control algorithms, and generates a set of analogue / digital outputs to change the logic states of devices.

Data Collection: Functionality that obtains digital values from a sensor instrument, associating consistent timestamps and necessary metadata.

Data Transmission: Functionality that transfers data over a communication channel using specified network protocols.

(B) Data Curation

Data Quality Checking: Functionality that detects and corrects (or removes) corrupt, inconsistent or inaccurate records from datasets.

Data Identification: Functionality that assigns (global) permanent unique identifiers to data products.

Data Cataloguing: Functionality that associates a data object with one or more metadata objects which contain data descriptions.

Data Product Generation: Functionality that processes data against requirement specifications and standardised formats and descriptions.

Data Storage & Preservation: Functionality that deposits (over the long-term) data and metadata or other supplementary data and methods according to specified policies, and then to make them accessible on request.

(C) Data Publishing

Access Control: Functionality that approves or disapproves of access requests based on specified access policies.

Metadata Harvesting: Functionality that (regularly) collects metadata in agreed formats from different sources.

Resource Registration: Functionality that creates an entry in a resource registry and inserts a resource object or a reference to a resource object with specified representation and semantics.

Data Publication: Functionality that provides clean, well-annotated, anonymity-preserving datasets in a suitable format, and by following specified data-publication and sharing policies to make the datasets publically accessible or to those who agree to certain conditions of use, and to individuals who meet certain professional criteria.

Data Citation: Functionality that assigns an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications and/or from other data collections.

Semantic Harmonisation: Functionality that unifies similar data (knowledge) models based on the consensus of collaborative domain experts to achieve better data (knowledge) reuse and semantic interoperability.

Data Discovery and Access: Functionality that retrieves requested data from a data resource by using suitable search technology.

(D). Data Processing

Data Assimilation: Functionality that combines observational data with output from a numerical model to produce an optimal estimate of the evolving state of the system.

Data Analysis: Functionality that inspects, cleans, and transforms data, providing data models which highlight useful information, suggest conclusions, and support decision making.

Data Mining: Functionality that supports the discovery of patterns in large datasets.

Data Extraction: Functionality that retrieves data out of (unstructured) data sources, including web pages, emails, documents, PDFs, scanned text, mainframe reports, and spool files.

Scientific Modelling and Simulation: Functionality that supports the generation of abstract, conceptual, graphical or mathematical models, and to run an instances of those models.

Scientific Workflow Enactment: Functionality provided as a specialisation of Workflow Enactment supporting the composition and execution of computational or data manipulation steps in a scientific application. Important processing results should be recorded for provenance purposes.

Data Processing Control: Functionality that initiates calculations and manages the outputs to be returned to the client.

(E) Data use

Authentication: Functionality that verifies the credentials of a user.

Authorisation: Functionality that specifies access rights to resources.