How the ENVRI Reference Model helps to design research infrastructures

Alex Hardisty works at the School of Computer Science and Informatics at Cardiff University. His interests span applications of computing to problems in environmental informatics, including engineering of large-scale distributed computing systems (e-Infrastructures) for data management and processing, and socio-technical issues of new technology adoption. In the ENVRIplus project he leads task 5.2 further developing the ENVRI Reference Model.

Abraham Nieva de la Hidalga (BSc, MEd, MSc, PhD) is a full research associate at the School of Computer Science and Informatics, Cardiff University. He is currently collaborating in the ENVRIplus project. He worked as a programmer and system analyst for government and private institutions in Mexico. During those years he was involved in the development and maintenance of different administrative systems, with diverse development platforms and various architectural models. His current research interests are within the area of Software Engineering, particularly in the aspects of producing high quality systems and the characterisation of complex distributed systems.

The advances in automation, communication, sensing and computation enable experimental scientific processes to generate data and digital objects at unprecedentedly great speeds and volumes. Environmental issues are a significant part of the research agenda. Research infrastructures provide advanced capabilities for data acquisition, sharing, processing, and analysis; enabling advanced research and playing an ever-increasing role in the environmental sciences [2, 3].

The ENVRI project identified several recurring problems in the development of environmental research infrastructures: i) duplication of efforts to solve similar problems; ii) lack of standards to harmonise and accelerate development, and bring about interoperability; iii) large number of existing relevant models and data information systems within the domain; and iv) integration requires deep understanding of different RI systems. However, the complexity of RIs entails a long learning process. By adopting the ENVRI reference model (ENVRI RM) as an integrative approach, the environmental research community can secure interoperability between infrastructures, enable reuse, share resources, experiences and common language, reduce unnecessary duplication of effort, and speed up the understanding of RI specifics.

The ENVRI RM provides a set of modelling templates that address different aspects of an environmental research infrastructure and allow the sharing of experiences, solutions and common understanding.

The ENVRI RM is derived from an industrial standard widely used for defining complex open distributed systems: the Reference Model for Open Distributed Processing (RM-ODP) [4, 5]. The core idea behind RM-ODP is that different aspects of a system's design may be better described from different perspectives, and certain subtleties of design may only become evident when considering the correspondences between different viewpoints. RM-ODP (and by extension, the ENVRI RM) provides an opportunity to explore these perspectives within a single framework. The ENVRI RM divides the modelling of the systems of an environmental research infrastructure across five viewpoints, each one describing concrete aspects of the system definition (Figure 1).

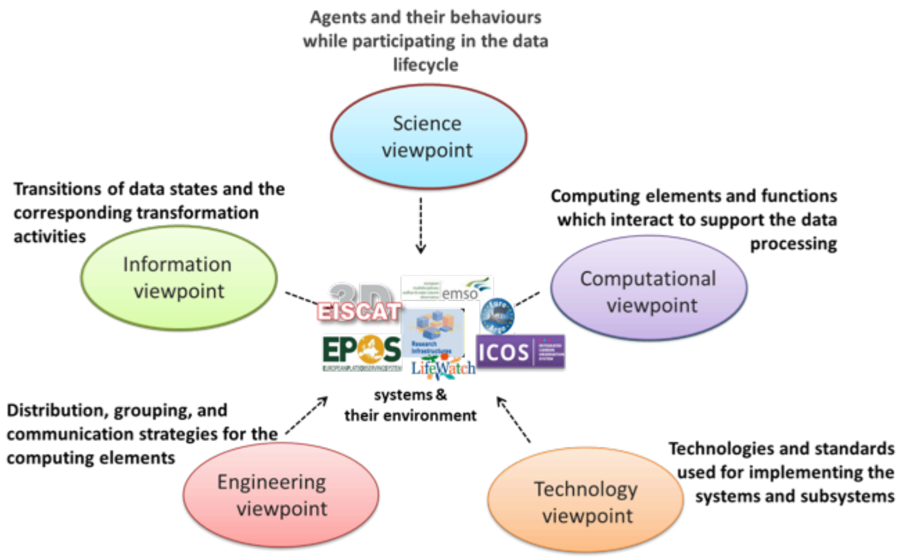


Figure 1 ENVRI RM Viewpoints. Science viewpoint: models specific actors participating in the processing of data as well as their behaviours and responsibilities. Information viewpoint: models the transitions between expected data states and maps them to typical transformation activities. Computational viewpoint: models computing elements and functions that interact to support data processing. Engineering viewpoint: models distribution, grouping, and communication strategies for the different computing elements in processing units. Technology viewpoint: models the integration of technologies and standards used for implementing the systems and subsystems

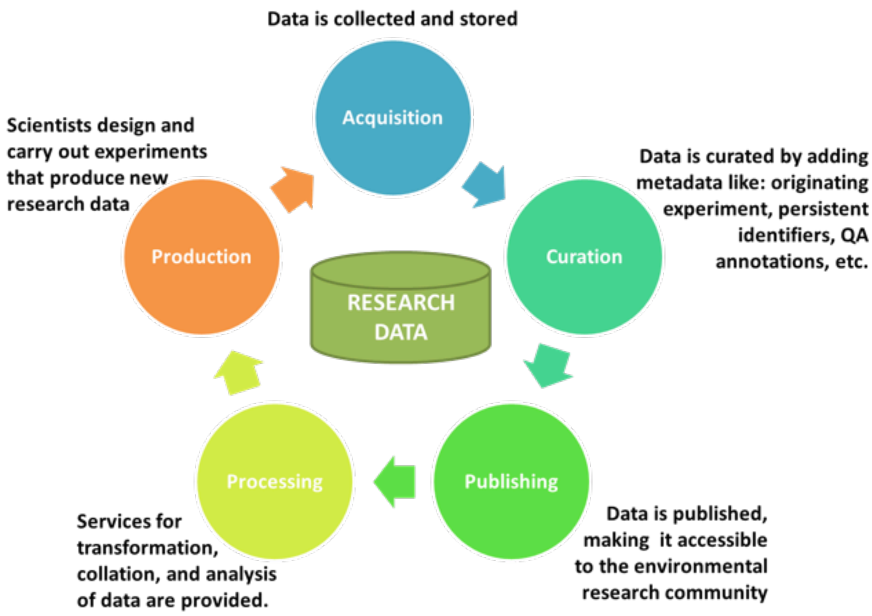
The reference model offers the mechanisms to map and keep the consistency of design across all five viewpoints, facilitating analysis, maintenance and evolution of these kinds of systems.

At the moment three of the viewpoints, Science, Information and Computational are fully developed [1], while the remaining two, Engineering and Technology are being developed in ENVRIplus.

The construction of the ENVRI RM has been derived from the original analysis of 6 research infrastructures (ICOS, EURO-Argo, EISCAT-3D, LifeWatch, EPOS, and EMSO). The analysis identified a basic data lifecycle with five main stages: production, acquisition, curation, publishing, and processing (Figure 2).

The ENVRI RM is built to support description of the systems and processes that are applied to the data, the communities involved, the systems and methods utilised, and the formats and standards applied. The recognition of the lifecycle revealed a set of functional requirements distributed across different stages. Most environmental science research infrastructures will implement some of these functions and delegate or outsource the implementation of other functions to some other infrastructures (e.g., foundational computing infrastructures like EGI, PRACE or national facilities).

Figure 2 Research data lifecycle. (1) Data is generated from an experiment or observation. (2) Data is collected. (3) Data is curated, adding metadata (such as: characterisation of source experiment, persistent identifier, methods and media used to acquire and store the data, QA). (4) Data is published allowing access to external parties. (5) Published data can be further processed (browsed, referred, downloaded, analysed). This will produce new data, marking the beginning of a new cycle.



The ENVRI RM is not prescriptive. It suggests different ways to align the requirements and functionalities according to their main purpose in the processing of data from acquisition to publication and use. The providers and sponsors of research infrastructures can model a system or parts of it from any viewpoint, preferably starting with the one that they are more confident with. And from there they can move on to model different viewpoints. Modelling research infrastructure is a team effort. Some of the participants in defining the RI may have more knowledge in some aspects of the systems, but no single person is expected to know all aspects in full detail.

Use of the ENVRI-RM enables: viewing the infrastructure as a set of systems and subsystems that are used by different communities to facilitate access to and processing of research data. The ENVRI RM fosters the use of a common language for describing these systems, facilitating understanding and communications across the research community. Increased use of the ENVRI RM will encourage the development of integration and coordination standards to interconnect research infrastructures and to make them more interoperable.

References

1. ENVRI Reference Model. http://envri.eu/rm

2. Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013, May). Addressing big data issues in scientific data infrastructure. In Collaboration Technologies and Systems (CTS), 2013 International Conference on (pp. 48-55). IEEE.

3. Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., ... & Porter, J. H. (2013). Big data and the future of ecology. Frontiers in Ecology and the Environment, 11(3), 156-162.

4. ISO/IEC, ITU-T (1995-1998). ISO/IEC 10746-1, 2, 3, 4| ITU-T Recommendation X. 901. X, 902, 1995-98.

5. Linington, P. F., Milosevic, Z., Tanaka, A., & Vallecillo, A. (2011). Building enterprise systems with ODP: an introduction to open distributed processing. CRC Press.