Example 3: Using the Reference Model in documentation (EMSO)

Descriptions of the Example

Researchers and architects of an ESFRI Environmental Research Infrastructure often encounter requests to describe their infrastructure, to introduce its particular architectural features, or to explain system requirements. The Reference Model offers a set of ready-to-use terminology with explicit definitions, which can be applied to various documentations. This example tells how the Reference Model has been used as a common language in writings to communicate with a community other than environmental science.

The Research Data Alliance (RDA) is established to accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability. This will be achieved through the development and adoption of infrastructure, policy, practice, standards, and other deliverables.

ENVRI has been actively supporting the RDA activities and made various contributions. In particular, ENVRI has been accepted as one of the use cases by the RDA Data Foundation and Terminology (DFT) working group, which has been set up to gather emerging requirements as well as to test research outcomes.

In preparing the use case, researchers and architects from two ENVRI-participating research infrastructures, EMSO and EPOS, used the terms and concepts defined in the Reference Model to describe architectural features of their research infrastructures. The resulting document from EMSO is presented below.

How to Use the Reference Model

The European research infrastructure EMSO is a European network of fixed-point, deep-seafloor and water column observatories deployed in key sites of the European Continental margin and Arctic. It aims to provide the technological and scientific framework for the investigation of the environmental processes related to the interaction between the geosphere, biosphere, and hydrosphere and for a sustainable management by long-term monitoring also with real-time data transmission. Since 2006, EMSO has been on the ESFRI (European Strategy Forum on Research Infrastructures) roadmap; it entered its construction phase in 2012. Within this framework, EMSO is contributing to large infrastructure integration projects such as ENVRI and COOPEUS. The EMSO infrastructure is geographically distributed in key sites of European waters, spanning from the Arctic, through the Atlantic and Mediterranean Sea to the Black Sea. It is presently consisting of thirteen sites that have been identified by the scientific community according to their importance respect to Marine Ecosystems, Climate Changes and Marine GeoHazards.

The data infrastructure for EMSO is being designed as a distributed system. Presently, EMSO data collected during experiments at each EMSO site are locally stored and organized in catalogues or relational databases run by the responsible regional EMSO nodes. The EMSO data architecture is currently adapted to the ENVRI Reference Model. As shown in Figure 1, according to the ENVRI-RM it includes the 5 ENVRI common subsystems. Concepts and terms defined in ENVRI-RM are used to illustrate the currently practiced common data management strategies for real time as well as archived data within the EMSO distributed data management system.

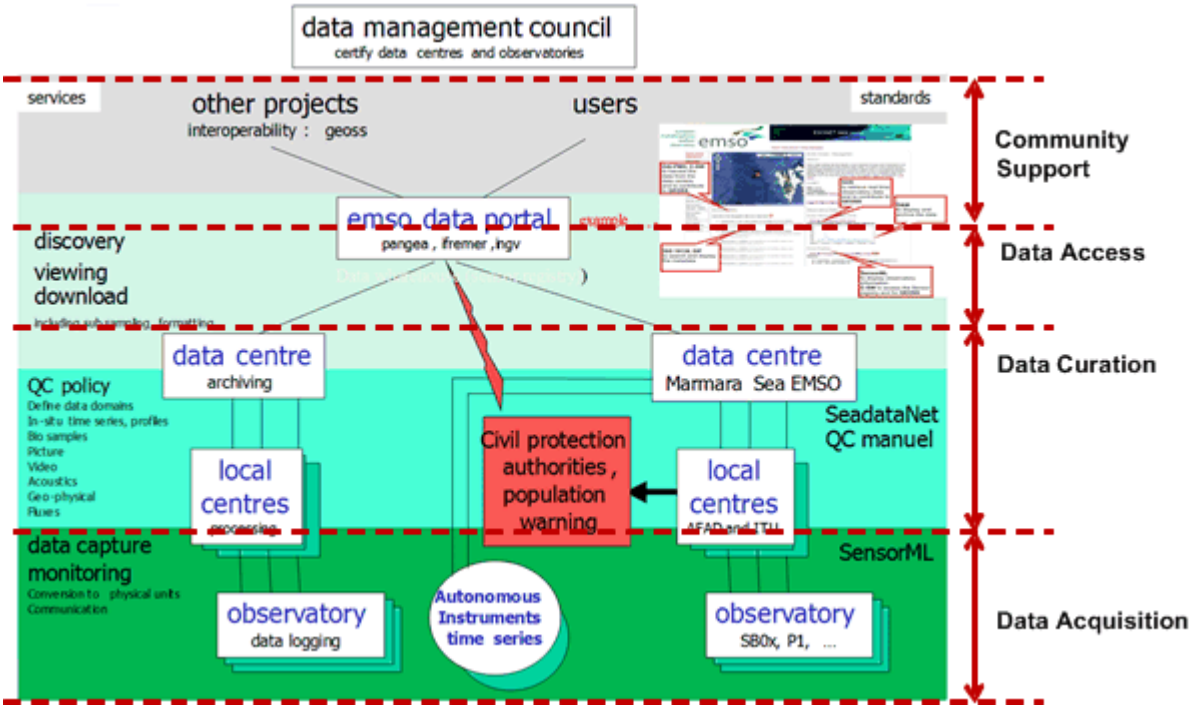

Figure 1: EMSO Distributed Data Management System

Data Acquisition

The EMSO **data acquisition sub-system** collects raw data from EMSO's marine observatories, which represent sensor arrays of varying geometry and various instruments or human observers, and brings the measures (data streams) into the system. **instrument configuration** and **design** is specified depending on the scientific demands and includes **sepcification of measurement**.

Depending on the deployment situation and nature of collected data, EMSO data is collected in real-time or delayed mode. Both **data collection** methods are performed by the regional nodes of EMSO that are responsible for the operation of marine observatories. Marine observatories have to deal with many technological challenges due to their extreme, deep sea deployment locations. Therefore data acquired by marine observatory sensor systems is most often temporarily staged within the instruments or the observatory's internal storage systems, and real-time transmission of data is only provided by observatories that are connected by submarine cables or permanent satellite connections. Whereas real time data are immediately available, the staged data becomes available for these systems only after visits during dedicated ship expeditions when the instruments are recovered or maintained. In addition, data are acquired through laboratory studies performed on material or samples collected at marine observatory sites such as multidisciplinary analyses of water samples, sediment cores, tow or trap catches.

Depending on the instrumentation and observatory design, on-site quality control and data filtering is applied, generally followed by a transformation process which converts the instrument specific data format into a transmission format required by EMSOs ***data curation*** and ***processing*** systems at the regional data centre nodes. The data collected by the *data acquisition sub-system* are transmitted to the ***data curation*** , to be maintained and archived there.

Data Curation

The EMSO **data curation sub-system** facilitates **data curation**, **quality** and **preservation** of scientific data. It is operated at the data centres responsible for archiving the data acquired by the EMSO regional nodes. Three major data centres are currently offering these services for EMSO data: UniHB (PANGAEA), INGV (MOIST) and IFREMER (EUROSITES).

data importer are provided by these institutions which either transfer the above mentioned data transmission format into an archival format or provide editorial tools and interfaces to ingest delayed mode data and laboratory analysis into their systems. Data which are intended to be transferred to the regional nodes data archives are quality checked, linked with an appropriate set of **metadata** according to international standards and persistently identified, depending on the archives internal standards and procedures. EMSO offers **catalogue service** and **data exporter** for each regional node. The node systems PANGAEA and MOIST services based on metadata standards such as ISO19115, GCMD-DIF and extended Dublin Core, for EUROSITES data, metadata is extracted from NetCDF files via a central EMSO service. **data exporter** are not yet fully implemented at all EMSO nodes, however it is planned to provide NetCDF export services for each node. The regional archives are responsible for cataloguing and long term preservation of these data that are provided for users via EMSO's *data access* and *discovery* subsystems.

Data Access and Discovery

The EMSO **data access sub-system** enables discovery and retrieval of data housed in data resources managed by a *data curation sub-system*. EMSO offers **discovery and access** via a common **metadata catalogue** and web portal which can be visited at http://dataportals.pangaea.de/emso. The portal is based on the brokerage system panFMP (http://www.panfmp.org) and uses Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) or simple file transfer via FTP/HTTP to harvest metadata from EMSOs distributed regional node data archives and their archival systems PANGAEA, MOIST and EUROSITES.

The EMSO data portal offers machine-human as well as machine-machine search facilities and discovery services based on the collected **metadata**. This includes a simple web-based user interface, a data search engine, which is offered at the EMSO data portal in a Google like style. In addition the data portal offers a common discovery service following the OpenSearch specification including the OpenSearch-Geo extension. A Open Geospatial Consortium (OGC) Catalogue Service for Web (CS-W) interface is currently under development.

A centralized data export service for these archived data is not implemented or planned, therefore, unless each EMSO data archive offers its own NetCDF data transformation service (see above) data requests are not yet processed by the EMSO data portal but are redirected to the hosting data archives which provide their own data access services for data retrieval.

Access to real time data is also offered via the EMSO data portal. EMSO has chosen to implement core standards of the OGC Sensor Web Enablement (SWE) suite of standards, such as Sensor Observation Service (SOS) and Observations and Measurements (O&M) to deliver real time data. These interfaces and formats are used to offer a common, web based SOS client which provides interactive visualizations of real time data.

Data Processing

Centralized **data processing sub-system**s that aggregate the data from various resources and provides computational capabilities and capacities for conducting data analysis and scientific experiments are not yet implemented for EMSO. Once more regional EMSO nodes and their data archives support NetCDF data export, it has been envisaged to introduce data visualization and plotting services at the EMSO data portal following the ESONET example. However presently, data processing services such as visualization, mining, as well as statistical services, are exclusively provided by each regional node and its responsible data centre.

Community Support

Centralized **community support sub-system** services to **profile management**, control and **tracking** **users**' activities and supports users to conduct their roles in communities are not yet implemented or planned for EMSO.

Summary

The EMSO example demonstrates how to use the common language defined by the Reference Model in documentation to communicate with the RDA community.

It has been recognised there is a common challenge when communicating with external organisations or communities -- "*your 'model' is not my 'model', your 'data' is not my 'data*"'. With a public accessible reference base, an external community who has little domain knowledge, such as the RDA, is able to understand the specific descriptions of EMSO by looking up the terminology in the Reference Model. In a way, using the Reference Model, the communication efficiency can be improved.

The ENVRI Reference Model provides a set of ready-to-use terminology, in principle:

- Terms in the Science Viewpoint can be used for describing requirements, use scenarios, and human activities;
- Terms in the Information Viewpoint for describing information objects handled in a system, their action types, constrains, states, and lifecycles; and
- Terms in the Computational Viewpoint for describing functionalities, computational components, interfaces and services.

A reader may have noticed there are some terms in the writing that are different from the ones linked back in the Reference Model. For example, " ***example3_setup*** *(… of each observatory)"* is linked to "***instrument configuration***". The intention is to show that in practice, to pursue the fitness, significance or beauty of the writing, an author may use different vocabularies to express the same concept. However, one can link them to the related concepts and definitions in the Reference Model to indicate the precise meanings. In this sense, using the Reference Model is different from using a dictionary – referring to the Reference Model places more emphasis on conceptual relativity.