Example 2: Using the Reference Model as an Analysis Tool (EUDAT)

Description of the Example

This study case provide an example for ESFRI Environmental Research Infrastructures project managers and architects to use the ENVRI Reference Model as an analysis tool to review an emerging technology, the EUDAT data infrastructure and its service components. Such an analysis can help them better understand the newly developed technologies and decide on how to make use of the generic services provided in their own research infrastructures.

The EU-funded EUDAT project is developing a pan-European data infrastructure supporting multiple research communities. Such a generic data infrastructure is seen as a layer in the overall European scientific e-infrastructure to complement the computing layer (EGI, DEISA, PRACE) and the networking layer (GEANT).

The design activities of EUDAT are driven by use-case-based community requirements EUDAT reviews the approaches and requirements of different communities, such as linguistics (CLARIN), solid earth sciences (EPOS), climate sciences (ENES), environmental sciences (LIFEWATCH), and biological and medical sciences (VPH), identifying common services, and provides computational solutions. Initially, 4 services are provided within EUDAT data infrastructure:

- **Safe replication**: which enables communities to replicate datasets -- using the integrated Rule-Oriented Data System (iRODS) as a replication middleware -- within data centre sites, with persistent identifiers automatically assigned to the digital objects in order to keep track of all the replicas;
- **Data staging**: which enables easy movement of large amounts of data between EUDAT storage resources and workspace areas on high-performance computing (HPC) systems to be further processed.
- **Metadata Catalogue**: which allows researchers to easily access metadata of data (or their collections) stored in EUDAT nodes. EUDAT will also harvest external metadata (which contains pointers to actual data) from stable metadata providers to create a comprehensive joint catalogue that will help researchers to find interesting data objects and collections.
- **Simple Storage**: which allows registered users to upload “long tail” data objects (large in number but small in size), and share such objects with other researchers.

We use the concepts developed in the ENVRI Reference Model to analyse the EUDAT data infrastructure and its service components. Only cursory analysis is provided, since the main purpose of the study case is to illustrate the usage of the ENVRI Reference model.

How to Use the Reference Model

Analysis of EUDAT common services and components

The ENVRI Reference Model models an archetypical environmental research infrastructure (RI). As a service infrastructure, EUDAT itself is therefore not an implementation of the Reference Model, but is rather a source of implementations for instances of objects required by any RI implementing the Model.

Table 1: Mapping EUDAT Services to the Reference Model Elements

EUDAT Services	Computational Viewpoint	ENVRI Common Subsystem
Safe replication	**data transfer service**	**CV Data Curation**
Staging	**data importer**	**CV Data Curation**
Metadata Catalogue	**catalogue service**	**CV Data Curation**
Simple Store	**data store controller**	**CV Data Curation**

From the ***computational*** perspective, EUDAT offers services that can be used to instantiate various objects in the Reference Model. For example EUDAT’s Safe Replication facilities can implement various required services within the **subsys_cur** of an environmental RI:

- **CV Data Acquisition**: EUDAT does not offer facilities for **subsys_acq**, relying on data already gathered by client RIs.
- **CV Data Curation**: EUDAT can provide instances of any of the computational objects used for data curation (including **data store controller**, **data transfer service** and **catalogue service**), either in place of or complementary to instances provided by an environmental RI – the extent to which EUDAT assumes the curation role for an infrastructure will vary from case-to-case.
- **CV Data Access**: Data access to EUDAT curated data is brokered by EUDAT, whilst the RI would broker RI-curated data. In practice the RI **broker** would sit in front of the EUDAT broker, forwarding data requests that involve data delegated to EUDAT.
- **CV Data Processing**: EUDAT do not offer data processing (beyond *metadata annotation*) as a core service; *workflow enactment* is being investigated as a future service however, which would allow a later version of the EUDAT platform to implement elements of a **subsys_pro**.
- Whilst certain aspects of EUDAT such as the Simple Store for researchers might be directly accessible as an independent **gateway service** , in general EUDAT sits behind a client RI, its services hidden behind the RI’s native services from the perspective of the RI’s user community. It would be likely however that the ‘virtual laboratories’ by which community groups interact with an RI would be in some way augmented by EUDAT services; in particular, implementations of the **security service** would integrate the EUDAT AAI service to allow seamless integration of EUDAT-held datasets with locally-held RI datasets.

The most immediately apparent conclusion that can be drawn from cursory analysis of EUDAT services in the context of the Reference Model is that EUDAT can potentially implement the entire **CV Data Curation** of an environmental RI; however in practice, one would expect that an RI would retain a certain amount of data locally (particularly raw data that is expensive to transfer off-site), necessitating a more nuanced division of labour between the RI and EUDAT. In particular, EUDAT provides replication services, allowing the co-existence of RI and EUDAT data stores holding the

same data, and EUDAT provides metadata (including global persistent identifier) services, allowing EUDAT to provide any **catalogue service** (probably complementary to catalogue services maintained by an environmental RI itself). The delegation of services will be a product of negotiation between the environmental RI and the EUDAT project (some degree of automation may be possible, but likely sufficient for only smaller projects).

Summary

The principal potential benefit of using the Reference Model in general is the ability to precisely identify components required by an environmental RI and then identify how (if at all) the RI implements those components. In the EUDAT context, EUDAT provides a number of services that implement certain components (primarily in **CV Data Curation**); it should therefore be possible to identify the equivalent services in a modelled RI and determine whether or not there is a benefit to delegating those services to EUDAT. This decision may be based on cost (particularly related to economies of scale) and development time (in cases where the RI has not yet implemented the service, but may be able to use the EUDAT service instead).