Example 1: Using the Reference Model to Guide Research Activities (EISCAT 3D - EGI)

Descriptions of the Example

This example explains the usage of the Reference Model in a pilot project that investigates the big data strategies for the EISCAT 3D research infrastructure. The Reference Model serves as a knowledge base to guide various research activities.

EISCAT, the *E*uropean *I*ncoherent *Scat*ter Scientific Association, was established to conduct research on the lower, middle and upper atmosphere and ionosphere using the incoherent scatter radar technique. This technique is the most powerful ground-based tool for these research applications. A next generation incoherent scatter radar system, EISCAT 3D, is being designed. The multi-static radars to be used will be a tool to carry out plasma physics experiments in the natural environment, a novel atmospheric monitoring instrument for climate and space weather studies, and an essential element in multi-instrument campaigns to study the polar ionosphere and magnetosphere. It will be a world-leading international research infrastructure, using the incoherent scatter technique to study how the Earth's atmosphere is coupled to space.

The design of the EISCAT 3D opens up opportunities for physicists to explore many new research fields. On the other hand, it also introduces significant challenges in handling large-scale experimental data that will be massively generated at great speeds and volumes. During its first operation stage in 2018, EISCAT 3D will produce 5PB data per year, and the total data volume will rise up to 40PB per year in its full operations stage in 2023. This challenge is typically referred to as a big data problem and requires solutions beyond the capabilities of conventional database technologies.

EISCAT is currently considering the use of e-Science technologies to deliver strategies for handling its big data products. Advanced e-Science infrastructure projects such as EGI, PRACE, and their enabling technologies are making large-scale computational capacities more accessible to researchers of all scientific disciplines. Emerging infrastructures, such as cloud systems proposed by the Helix Nebula project and by the EGI Federated Cloud Task Force, or the data infrastructure to be provided by EUDAT will extend possibilities even further.

As a potential of e-science partner for EISCAT, we present EGI. EGI was established in 2010 as a Europe-wide federation of national computing and storage resources. The EGI collaboration is coordinated by EGI.eu, a not-for-profit foundation created to manage the infrastructure on behalf of its participants: National Grid Initiatives and European Intergovernmental Research Organisations. Resources in EGI are provided by about 350 resource centres from the NGIs who are distributed across 55 countries in Europe, the Asia-Pacific region, Canada and Latin America. These providers operate more than 370,000 logical CPUs, 248 PB disk and 176 PB of disk capacity (June 2013 statistics) to drive research and innovation in Europe and beyond.

Since February 2013, a pilot project has been set up within ENVRI, which establishes a partnership between EISCAT, EGI and EUDAT, aiming to identify and allocate solutions that directly benefit EISCAT 3D, which can also be reused in other ESFRI projects involved in ENVRI. ENVRI WP3 has been involved in this investigation, and uses the Reference Model to guide various research activities, including;

- Analysis of the EISCAT 3D data infrastructure; Capturing requirements from the EISCAT 3D scientific community concerning applications that work with and process data.
- Analysis of EGI and EUDAT services; Identifying the gaps between the generic service infrastructures of these providers and the domain-specific requirements of EISCAT 3D.
- Provide recommendations to EISCAT 3D for the setup of up a big data strategy and a big data infrastructure for its community. Setup demonstrators/proof-of-concept systems if resources permit.

Having fulfilled these tasks, the Reference Model is proving to be useful as a knowledge base that can be referred when conducting various system analysis and design activities.

How to Use the Reference Model

In the following, we describe how the Reference Model is used to conduct several system analysis tasks.

Analysis of the EISCAT 3D Data Infrastructure

The initial challenge for the pilot project is to understand the EISCAT 3D data infrastructure. The existing design documents of EISCAT 3D has been focused on the incoherent scatter radar technologies. As shown in Figure 1, its data infrastructure is embedded within the overall design of the observatory system that is difficult for a computer scientist/technologist having little physics knowledge background to understand.

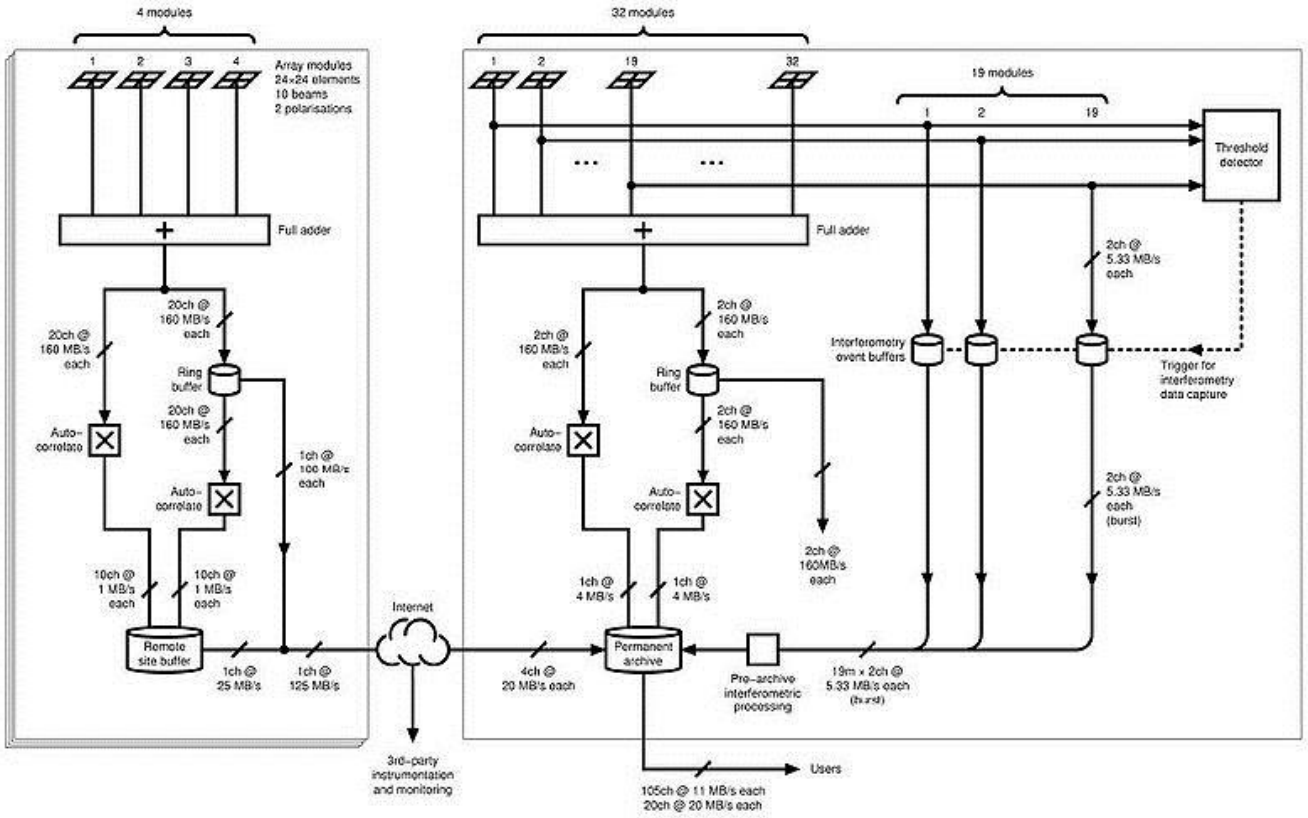

Figure 1: The original design of EISCAT 3D data infrastructure is embedded within the overall observatory system design

We use the 💡 **ENVRI_Common_Subsystem** framework to decompose the computational elements, clarifying the boundary between the radar network and data infrastructure, which results in Figure 2. This diagram now, instead of Figure 1, is frequently used in presentations and discussions of the EISCAT 3D data infrastructure.

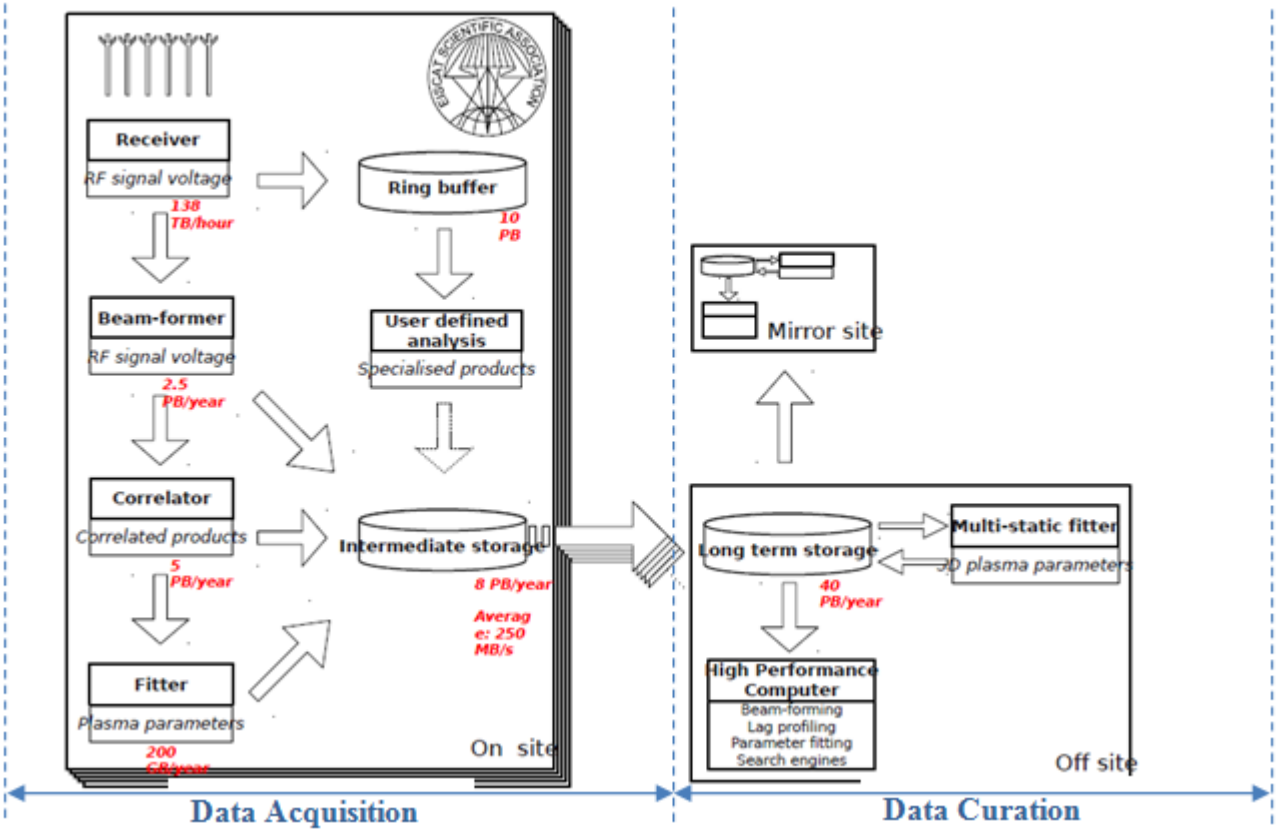

Figure 2: Using the 5 ENVRI Common Subsystem to interpret the EISCAT 3D data infrastructure makes it easy to communicate with computer scientists/technologists

Figure 2 illustrates that the EISCAT 3D functional components can be placed into 2 ENVRI common subsystems, **subsys_acq** and **subsys_cu** . Briefly, at the **subsys_acq**, the raw signal voltage data will be generated by the antenna *Receivers* at the speed of 125 TB/hr, and be temporarily stored in a *Ring buffer*. A second stream of RF signal voltages will be passed to a *Beam-former* to generate the beam-formed data (1MHz). Continually, the beam-formed data will be processed by a *Correlator* to generate correlation analysis data based on standard methods. Then, the correlation data will be delivered to a *Fitter* to produce the fitted data (1GB/year). In order to support different user requirements, EISCAT 3D will allow users to access and process the raw voltage data in the *Ring buffer* and to generate the specialised products based on self-defined analysis algorithms. Both raw data and their products will be stored in *Intermediate storage* (11PB/year), from where they will be delivered to the central site within the curation subsystem.

In **the curation subsystem** , *Long-Term Storage* will preserve the raw voltage data and their products. A *High Performance Computer* will be used for data searching and processing (e.g., beam forming, lag profiling or other correlation, and parameter fitting). Searching facilities will enable user to search over all data products and to identify significant data signatures. A *Multi-static fitter* will be installed to process the stored raw voltage data to generate the 3D plasma parameters that will then be stored back in *Long-Term Storage*. A complete copy of *Long-Term Storage* data will be established at mirror sites; related data processing and searching tools will be provided.

While it is made clear that the design specification covers 2 of 5 common subsystems described in the ENVRI Reference Model, we understand functionalities of the other 3 subsystems are currently missing. The reason of this is likely due to resource limitations. However, the absent 3 subsystems are crucial for a big data system such as EISCAT 3D. Without providing services to support data discovery, access, processing and user community, the value of EISCAT 3D big data cannot be unlocked, and expensively generated and archived scientific data will be useless.

Using the Reference Model as the analysis tool, we identified the missing pieces of the design specification, which gives the direction for future investigation.

Analysis of EGI Enabling Services and Construction of an Integrated Infrastructure

We need to understand the functionalities of EGI services and how to integrate them to support the EISCAT 3D requirements.

A set of generic services are enabled by the EGI e-Infrastructure, including:

- AMGA Metadata catalogue
- LFC File catalogue
- Storage elements
- File Transfer Service
- Portal for application development & hosting (e.g. SCI-BUS)
- Access control

Showing in Table 1, by examining the functionalities of the EGI services and mapping them to the ENVRI Reference Model computational model objects, we understand these services fall into 2 ENVRI common subsystems: Curation and Community Support.

Table 1: Mapping EGI Services to the Reference Model Elements (from **Computational Viewpoint** perspective)

EGI Services	**ENVRI- RM Computational Objects**	**ENVRI Common Subsystem**
AMGA Metadata catalogue	**catalogue service**	**CV Data Curation**
LFC File catalogue	**catalogue service**	**CV Data Curation**
Storage elements	**data store controller**	**CV Data Curation**
File Transfer Service	**data transfer service**	**CV Data Curation**
Portal for application development & hosting	**virtual laboratory**	**CV Data Use**
Access control	**security service**	**CV Data Use**

Above analysis gives clues to a solution for integrating the EGI technologies into the EISCAT 3D data infrastructure. Depicted in Figure 3, a secondary **CV Data Curation** (seen as the mirror site of the EISCAT 3D central archive in Figure 1) can be established using the EGI infrastructure and its services. Data from EISCAT 3D central archive (or the acquisition subsystem) can be staged into the EGI storages, and be managed using LFC File Catalogue and AMGA Metadata Catalogue. At the front end, an EISCAT science gateway can be established, seen as part of a **CV Data Use**, to provide access control (e.g., authentication, authorisation, and single sign-on) and application portals (e.g., to which processing and data- mining applications from EISCAT 3D can be plugged in).

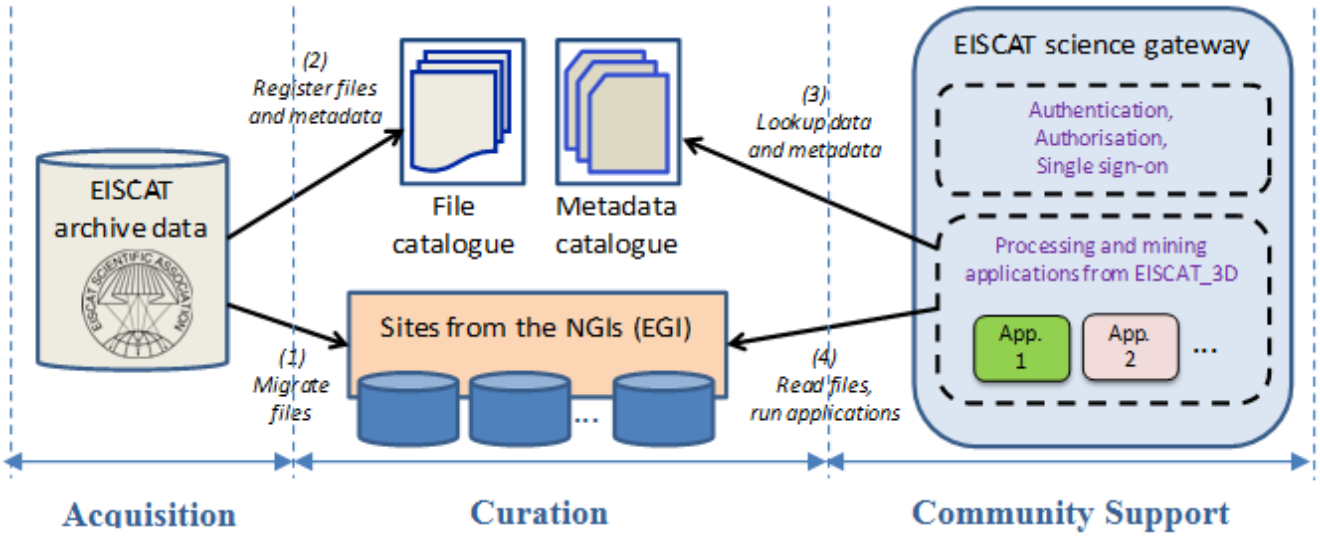

Figure 3: An integrated infrastructure of EGI and EISCAT 3D

Using the Reference Model, functional elements of both EISCAT 3D and EGI can be placed into a uniform framework, which provides a way of thinking about the construction of the integrated infrastructure.

Evaluation of the Feasibilities of the EGI Infrastructures and Services in Supporting EISCAT 3D Requirements

Using the common framework enabled by the Reference Model, we can analyse and compare the EGI and EUDAT generic service infrastructure and the requirements from a domain-specific data infrastructure such as EISCAT 3D, and we understand that there are significant gaps in-between, including but not limited to:

- Staging services to ship scientific data from observatory networks into the EGI generic service infrastructure (and to get the data off) are missing. Such a staging service should be able to transmit both big chunk of data (up to petabyte) and continuing updates/real-time data streams during operations. Such a service should satisfy performance requirements, including:
  - Robust. Environmental scientific research needs high quality data. In particularly, during important natural events, losing observation data is unaffordable. Fault-tolerance is desirable, which requests the transmission service can be self-recover from the interruption point without restarting the whole transmission process.
  - Fast, e.g., in the case of EISCAT 3D, the 10PB ring-buffer can only hold data for about 3 days, and the big observation data need to be transferred to the archive storage fast enough to avoid being overwritten.
  - Cheap, e.g., the observatory networks are remote from the EGI computing farm. Using high-capacity pipes are possible but expensive. Software solutions such as, intelligent network protocols, optimisation, data compression, are desirable.
- Cost effective large storage facilities and long-term archiving mechanisms are urgently needed. Environmental data, in particular for climate research, need to be preserved over the long-term to be useful. Being Grid-oriented, EGI is not designed for data archiving purposes. Although large storage capabilities are potentially available through NGI participants, EGI does not guarantee long-term persistent data preservation. Curation services such as advanced data identification, cataloguing and replication are absent from the EGI service list.
- The EGI infrastructure needs to adapt in order to handle emerging big- data phenomena. The challenge is how to integrate what is new with what already exists. Services such as job schedulers need to be redesigned to take into account the trade-off of moving big data; intelligent data partitioning services should be investigated as a way to improve the performance of big data processing.
- Advanced searching and data discovery facilities are urgently needed. It is often said that data volume, velocity, and variety define big data, but the unique characteristic of big data is the manner in which the value is discovered [38]. Unlike conventional analysis approaches where the simple summing of a known value reveals a result, big data analytics and the science behind them filter low value or low-density data to reveal high value or high-density data [38]. Novel approaches are needed to discover meaningful insights through deep, complex search, e. g., using machine learning, statistical modelling, graph algorithms. Without facilities to unlock the value of big data, expensively generated and archived scientific data will be useless.
- Community support services are insufficient. The big data phenomena will eventually lead to a new data-centric way of conceptualising, organising and carrying out research activities that could lead to an introduction of new approach to conducting science. A new generation of data scientists is emerging with new requirements. Service facilities should be planned to support their needs. These together should enable the EISCAT 3D community to design new applications that are capable to work with big data, and can implement these on cutting-edge European Distributed Computing Infrastructures.
- Currently, EUDAT has taken up the role to implement a collaborative data infrastructure, however only a few services are available, storage facilities are insufficient, and policies for usage are unclear. Among our current investigations, we are investigating the possibility of integrating EUDAT services into EGI infrastructure, seen as a layer on top of the EGI federated computing facility. The analysis of the EUDAT services is included in Example 2: Using the Reference Model as an Analysis Tool (EUDAT) of the Reference Model.

Summary

In this example, we have shown that the Reference Model could be used to conduct various system analysis tasks. Using the Reference Model we have:

- Clarified the boundary of EISCAT 3D data infrastructure and identified missing functionalities in the design;
- Provided a solution to integrate the EGI services into EISCAT 3D data infrastructure;
- Identified gaps between the EGI generic service infrastructure with the requirements from a domain specific research infrastructure, EISCAT 3D.

We have shown that the Reference Model offered a research infrastructure:

- A knowledge base containing useful information could be referred in various system analysis and design activities;
- A uniform platform into which computational elements of different infrastructures could be fitted, enabling comparison and analysis;
- A way of thinking of constructions of plausible system architectures.