# Towards a Big Data Strategy for EISCAT-3D

Yin Chen, Ingemar Häggström, Gergely Sipos, Małgorzata Krakowian

**Motivations**

The European Incoherent Scatter Scientific Association, EISCAT [2], operates the world’s largest system of incoherent scatter radar installations and other radio diagnostics. It is a unique resource to observe the high-latitude atmosphere and ionosphere which are important for studies of the relationship between Solar and Terrestrial conditions as well as the coupling of the different altitude regions in the Earth’s atmosphere. The design of the next generation incoherent scatter radar system, EISCAT-3D [3], 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 which will be massively generated at great speeds and volumes. During its first operation stage in 2018, EISCAT-3D will produce 1PB data per year, and the total data volume will rise up to 10PB per year in its full operations stage in 2023. This refers to so-called big data problem, whose size is beyond the capabilities of the current database and software technology [1]. To unlock the value from these huge volumes of data, new forms of processing and platforms of tools are needed.

Advanced e-Science infrastructures such as, EGI [4], EUDAT [5], and PRACE [6], and their enabling technologies are making large-scale computational capacities more accessible to researchers of all scientific disciplines. The European Grid Infrastructure (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. EUDAT is a European project aiming to take the first steps towards building a Collaborative Data Infrastructure for European scientific data products. It will offer services for data storage and replication, data staging to computational resources (and vice versa) and services for data cataloguing and discovery. PRACE is the pan-European supercomputing infrastructure that forms the top-tier of HPC provision across Europe, with the aim of enabling high impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness. Mention should also be made of the Helix Nebula public/private partnership initiative between CERN, EMBL and ESA together with a consortium of major ICT industry partners, with the aim of providing sustainable large-scale cloud computing capability for science.

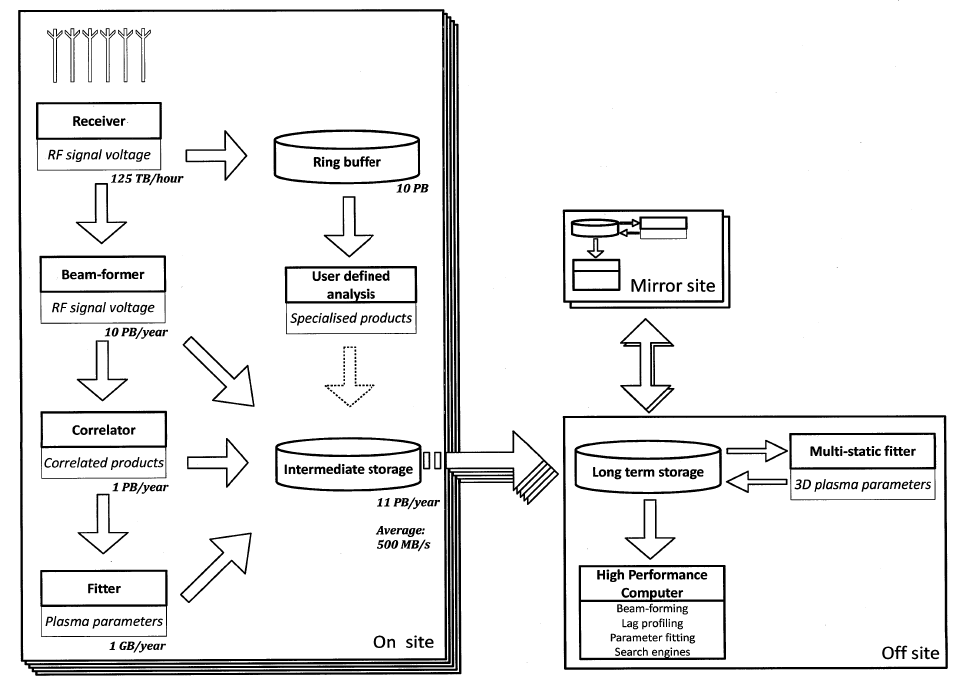
EISCAT-3D and EGI.eu established a ‘study case’ collaboration [7] in February 2013 inside the ENVRI project [8]. The study case aims to identify existing services and solutions from EGI that can address the data pre-processing, post-processing, publishing needs of EISCAT-3D, and which could be reused in other ESFRI projects of ENVRI as well. In the context of this study case we propose setting up a proof of concept system that would help EISCAT take the first steps towards a strategy that can help handling big data in EISCAT-3D. The proof of concept system would be able to:

1. Stage EISCAT-3D lower-level data (voltage data) into a large-scale, distributed e-science storage system, such as storage elements of EGI or EUDAT. The data to be used in the proof of concept is the ~60TB EISCAT archive that has been collected between 1981-2013.
2. Provide advance discovery facilities for scientists to search through all levels of the data and identify specific signatures, for example: plasma features, meteors, space debris, astronomical features, etc.
3. Provide data processing and mining facilities (applications) such as: auto-correlation and spatial/temporal integration, to allow individual scientists to analyse data as their will.

**Towards an EISCAT-3D big data strategy**

With the proof of concept system we are interested in identifying those technological solutions that already exist or still need to be developed to manage data of EISCAT-3D. The proof of concept will help us set up the prototype version of certain parts of this system, and will help us advance the searching technology of large-scale data. Particularly it will provide novel solutions for handling big data in EISCAT-3D, for example, to study the space physicists' query behaviours and identify frequently used query patterns, to investigate data partitioning strategies to improve the searching performance, to provide various searching facilities to deal with the large-scale dataset, e.g., similarity searching, Top-K searching and so on. The collaboration shall start with urgent and important requirements collected from the EISCAT community.

The new data processing and searching strategy will offer more flexible way for EISCAT users to analyse and discover interesting data patterns which are not yet available. Space physicists will be able to make better use of the observation data and exploit the growing wealth of them. This will eventually lead to a new data-centric way of conceptualising, organising and carrying out research activities which could lead to an introduction of new approaches to solve problems that were previously considered extremely hard or, in some cases, impossible to solve and also lead to serendipitous discoveries and significant breakthrough [1].



**Figure 1**: Architecture of EISCAT-3D

**Architecture of the proof of concept system**

The proof of concept system should be a Virtual Research Environment (VRE) that integrates resources, services, data and applications from EGI and EISCAT (See also Figure 1). The VRE should integrate:

From EGI:

1. Storage elements from the NGIs to provide capacity for EISCAT-3D files.
2. A file catalogue (e.g. LFC or DIRAC) and a metadata catalogue (e.g. AMGA or DIRAC) to register and to make files searchable through science domain specific metadata.
3. An application registry where data processing and mining software from EISCAT can be stored and made accessible for those who wish to process the EISCAT files.
4. Computing sites and a workload management system that can execute applications from the application registry on storage resources of the NGIs.
5. A web based science gateway system that provides a graphical environment for researchers to interact with the EISCAT data and with the processing applications.
6. A security system that is integrated with the science gateway and provides single sign-on, authentication and authorisation (access control) for users and service operators.

From EISCAT:

1. Data processing and mining applications that can understand and work with the EISCAT files and metadata.
2. Files of the EISCAT archive.



Figure 2. Architecture of the EGI - EISCAT-3D proof of concept system

**Next steps**

Before the proof of concept system can be set up the partners have to:

1. Provide clarifications about the EISCAT archive. Information about the structure, accessibility, used file formats, expected metadata structure, possible need for replication, expected user base, expected usage pattern, etc. (Action for EISCAT)
2. Identify existing software solutions that could be used in the EISCAT VRE system. Solutions that are capable of migrating and registering large number of files and metadata on EGI resources should be the first area of focus. (Action for EGI)
3. Provide details on the data processing and data mining applications that users need to interpret and process the EISCAT archive. Study the scalability of these applications with respect to data scalability and compute scalability (Action for EISCAT and EGI)

The partners will organise an online workshop to facilitate the completion of these actions. The workshop agenda will include presentations by experts from EISCAT and from EGI about data, about applications and about software solutions that should be integrated into the EISCAT VRE. Discussion time will be allocated after each presentation to ensure that the partners can collect enough information to complete the above actions.

**References**

[1] C. Thanos, S. Manegold and M. Kersten, “Big Data”, *ERCIM Special Theme: Big Data*, No. 89, Apr. 2012.

[2] EISCAT: <http://www.eiscat.com>

[3] EISCAT-3D: <https://www.eiscat3d.se/node>

[4] EGI: <http://www.egi.eu>

[5] EUDAT: <http://www.eudat.eu>

[6] PRACE: <http://www.prace-project.eu>

[7] EGI - EISCAT-3D – EURO-ARGO study case in ENVRI: <https://wiki.egi.eu/wiki/EGI_ENVRI>

[8] ENVRI project: <http://envri.eu>