# Towards a Big Data Strategy for EISCAT\_3D

Yin Chen, Ingemar Häggström, Gergely Sipos, Małgorzata Krakowian, Nuno Ferreira, Alex Hardisty, Ville Savolainen

**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 5PB data per year, and the total data volume will rise up to 40PB 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 the European Grid Infrastructure (EGI) [4] and PRACE [6], 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 HelixNebula project [9] and by the EGI Federated Cloud Task Force [10], or the data infrastructure to be provided by EUDAT [5] will extend possibilities even further. To identify services and solutions from state-of-the-art e-science infrastructure providers that can address the data pre-processing, post-processing, publishing needs of EISCAT\_3D, the EISCAT\_3D, EGI and EUDAT collaborations established a ‘study case’ partnership [7] in February 2013. The partnership has been setup in the ENVRI project [8], where these three collaborations are already involved. The study case will identify and allocate solutions that directly benefit EISCAT\_3D and that expected to be reusable in other ESFRI projects of ENVRI as well[[1]](#footnote-1).

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.

EISCAT\_3DEISCAT\_3D

In the context of the study case we propose the set 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 advanced 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. These were not possible in the previous system, become the EISCAT radars were not operated 24/7, in addition, due to high cost of storages at that time, the raw voltage data were not able to store.
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 and those that 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 partnership 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

The architecture of EISCAT\_3D data management system is depicted in Figure 1. The system comprises 5 distributed data acquisition subsystems located at individual radar sites and a central data curation subsystem located at the central site. Mirrored copies of the central data archive (and software tools) will be established, for example, at EGI and other EISCAT associated organisations. Each data acquisition site will be connected to the central site through a Wide Area network (WAN) using a 500MB/s ethernet. A 10GB/s line is planned to build between the central site and a selected mirror site to assist the transmission of the high volumes of data.

At each data acquisition subsystem, (1) the RF signal voltages are generated by the antenna *Receivers* at the speed of 125 TB/hr, and are temporarily stored in a *Ring buffer*. The planned storage capability for such a *Ring buffer* is 10PB (which is sufficient to hold up the raw data for 3 days before being rewritten). A second stream of (1) the RF signal voltages is passed to a *Beam-fomer* to generate (2) the narrow band raw voltage beam-formed data (1MHz). The data volume of (2) is about 10PB/year. Continually, (2) the beam-formed data are processed by a *Correlator* to generate (3) the correlation analyses data based on standard methods. The volume of (3) is expected to be 1PB/year. Then, (3) the correlation data are delivered to a *Fitter* to produce (4) the fitted data (1GB/year). In order to support different user requirements, users are allowed to access and process (1) the raw voltage data in the *Ring buffer* to generate (5) the specialised products based on self-defined analysis algorithms. All 5 types of data, the raw data and their products, are stored in a *Intermediate storage* (11PB/year), from where to be delivered to the central site.

At the centre site, a *Long-Term Storage* will preserves the raw voltage data and their products collected and generated at the data acquisition subsystem. 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 (1) raw voltage data to generate (6) the 3D plasma parameters, and (6) will be stored back to the *Long-Term Storage*.

At the mirror sites, a complete copy of the *Long-Term Storage* data and related data processing and searching tools are planned to establish.

**Architecture of the proof of concept system**

The proof of concept system would take the first step towards the specification and setup of the ‘Off site’ component of the EISCAT\_3D architecture (bottom-right in Figure 1). This component will act as a Virtual Research Environment (VRE) that integrates resources, services from EGI and EUDAT with archive data and applications from EISCAT. The VRE should integrate:

From e-infrastructures (EGI and EUDAT):

1. Storage elements from the National Grid Infrastructures of EGI to provide capacity for EISCAT\_3D files. Presumably this can come from those institutes at national level that both have storage capacity and whose Governments support EISCAT.
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.
3. User requirements for searching. Initially, we consider the following searching scenarios:
* Plasma parameters: Look for specific numbers or combinations of them. For example, find high electron temperatures above a given value at a specific height region. This maybe restricted with values of other parameters, like electron density. This tends to be simple cases.
* Correlated data. Look for patterns in the profiles of correlations. Any feature that one want to search for is convolved with the ambiguity function of the specific experiment. Also look for spectral features (a feature in the recorded frequency spectrum)
* Voltage data: Look for patterns in the time sequences of data. Any feature is modulated by the transmitter waveform and intersection between transmitter and receiver antenna beam patterns.

Typically, a researcher would define what plasma state to find a given feature. So firstly, he/she searches over the plasma parameters to find the times of possible 'hits'. Then he/she takes those times to classify the data using the correlated and Voltage data. In this case, a searching engine maybe firstly reduce the data to look through by a rather large number maybe 1000 times. In principle the higher level of data acts as metadata to the lower level.

The challenging issue is that the data is very complex for regular users and the setup of the different radar data change from run to run. In the precious EISCAT system, enough information are available to enable cross-experiment searches, but there is no mechanism prepared for it. In addition the radars are presently run on campaign basis, so 'big' searches have not been needed. EISCAT\_3D will run 24/7, which brings new opportunities to find new types of events and need a searching engine.

Another challenge is the performance issue. As a pilot study a Svalbard radar has been set up and run for 1 year continuously in 2007/8 during International Polar Year. Experiments have been carried on for on-site searching on the lower levels of data for specific events (need refs), which took a week for processing. In which case, the events to look for have been rather spectacular and easy recognisable (NIEALs).

The OpenRadar format is used for the correlated and Voltage data, describing various types of metadata, mainly technical, e.g., antenna pointing directions and transmitter behaviours (<http://www.openradar.org/cgi-bin/openradar.cgi/Documents/OpenRadarRFSignalObject>)

In the existing EISCAT system, Madrigal, <http://www.eiscat.se/madrigal/>, is used for the plasma parameters, and the lower levels of data are stored in a MySQL database, with mainly experiment name, radar site and UT hour in the table and pointers to the data.



Figure 2. Architecture of the EGI - EISCAT\_3D proof of concept system

**Next steps**

Within this partnership the members will:

1. Collect data, application and other types of requirements from the EISCAT community for the proof of concept system. Expected output of this work are:
	1. Details about the EISCAT archive. (Such as information about the structure, accessibility, used file formats, expected metadata structure, possible need for replication, expected user base, expected usage pattern, etc.)
	2. Details about the applications that EISCAT researchers would like to use to process, visualise, and in any other way manage and interpret the archive data.
2. Identify and evaluate technologies from EGI, EUDAT (and possibly from other e-infrastructures) that could satisfy the requirements and that can work together within a single architecture. Solutions that are capable of migrating and registering large number of files and metadata on EGI resources should be the first area of focus. Other areas of interest are portal framework, application repository, data visualisation and mining tools.
3. Update and document the proof of concept architecture draft that can be seen on Figure 2.
4. Setup a prototype system based on technologies and resources that are available for the architecture at that time. Perform usability, scalability and performance tests on the prototype system and its applications.
5. Draw lessons learnt and make recommendations for EISCAT towards the setup of the ‘Off site’ component of the EISCAT\_3D system, and for ENVRI concerning the setup of ESFRI infrastructures in the environmental sciences.

**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>
9. HelixNebula the Science Cloud: <http://helix-nebula.eu/>
10. EGI Federated Cloud: <http://go.egi.eu/cloud>
1. The ESFRI projects of ENVRI are EISCAT\_3D, EMSO, EPOS, Euro-Argo, ICOS, LifeWatch. [↑](#footnote-ref-1)