ENVRI | Common Operations of Environmental Research Infrastructures

Towards a generic Big Data e-Infrastructure

Report of the H2020 Proposal Meeting, 10-11 Feb, Cardiff

Yin Chen

Cardiff University
ChenY58@cardiff. ac.uk

Our Story

The ENVRI Project

- Interoperability between ESFRI ENV RIs
- Common solutions to common problems

Started
11/2011

Now
02/2014

Ends
10/2014

Common Requirements

Common Services/Components

ENVRI Reference Model

- Evaluation/adoption of ENVRI RM
- Experiments with e-Science technology
- Evaluation of EGI/EUDAT Services

The Pilot Study

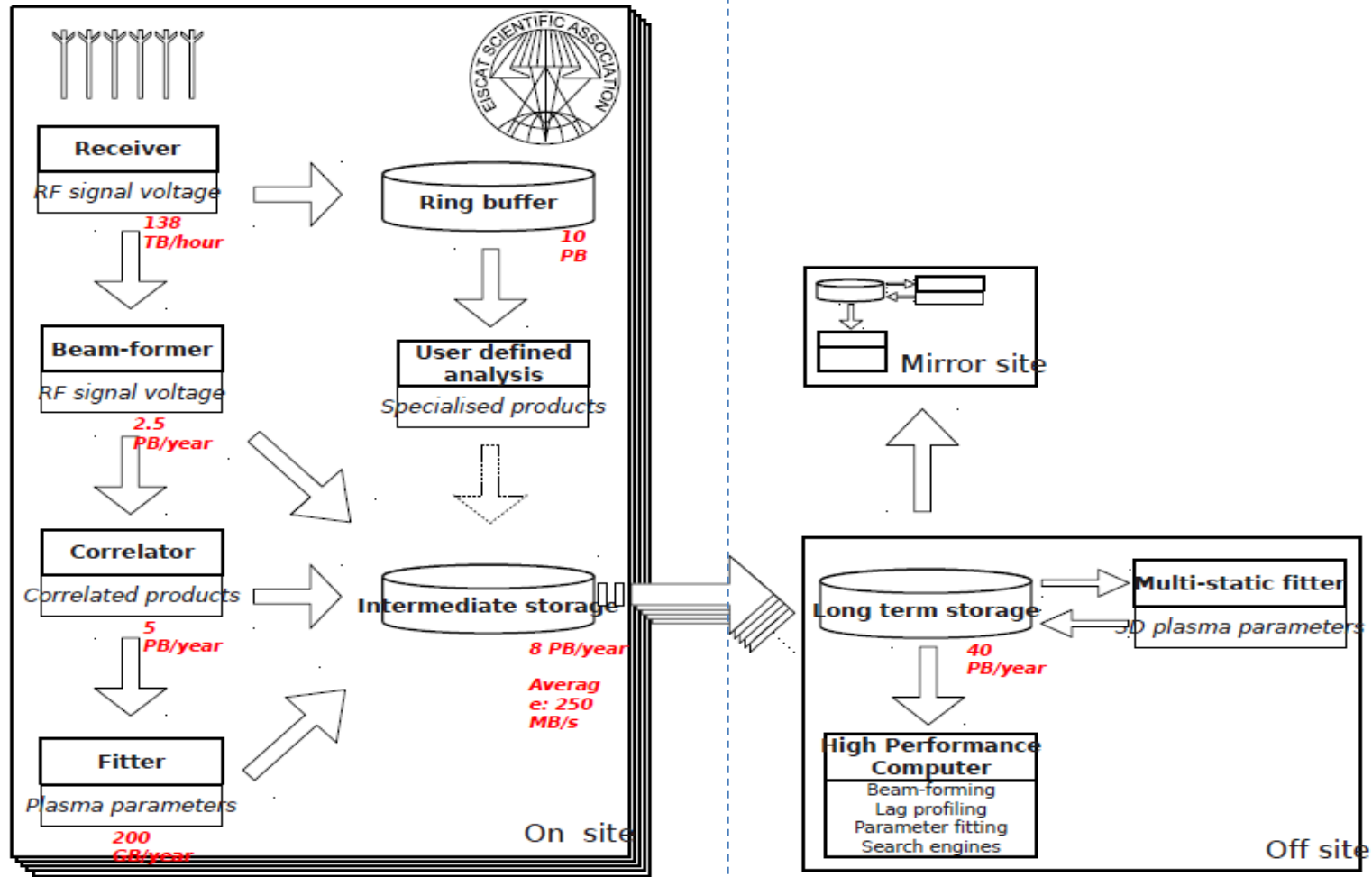- Community Standard
- Common language
- Unified framework

Started
02/2013

Ends
09/2014

Horizon 2020 Proposal

- Common services for the Big Data
- Support EISCAT Science Community
- Training of Data Scientists

Preparation
02/2013

Submission
09/2014

Starts
03/2015

Our Problems

3+1Vs Challenges in the case of EISCAT 3D

- **Volume.**
  - **8.5Gb/s*3sites=8PB/month=100PB/year**
  - **Operate for 30 years, data products to be stored for > 10 years**
- **Velocity.**
  - **Each antenna : 120MB/s**
  - **160 * antenna group (100 antennas): 2 Gbit/s/group**
  - **5* Ringbuffer: each 125 TB/h**
- **Variety.**
  - **Measurements: different versions, formats, replicas, external sources ...**
  - **System information: configuration, monitoring, logs/provenance ...**
  - **Users' metadata/data: experiments, analysis, sharing, communications ...**
- **Value.**
  - **How to discover meaningful insights from low-value-density data**
  - **Needs new approaches to the deep, complex analysis e.g., machine learning, statistical modelling, graph algorithms etc.**
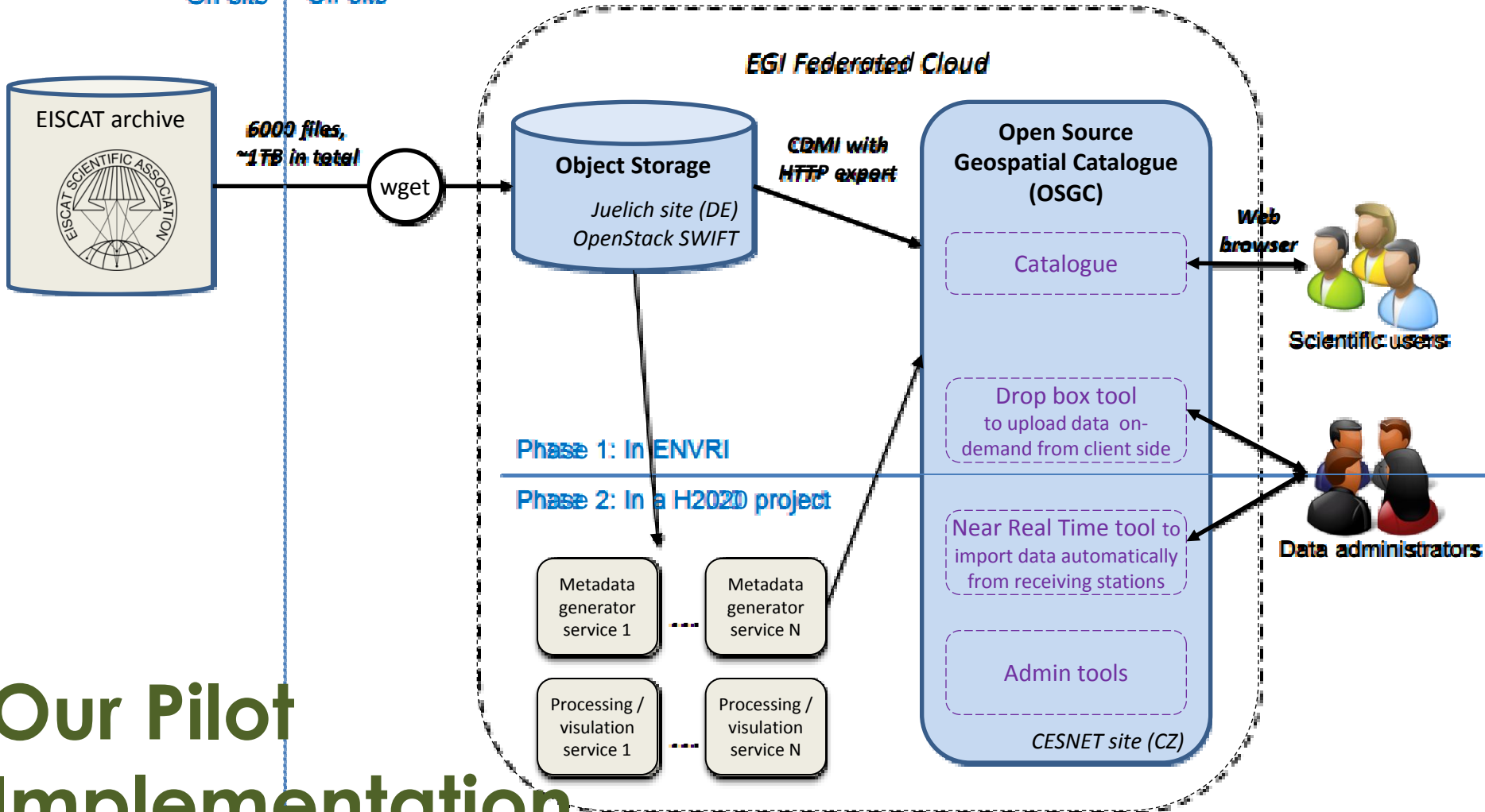 - **Go beyond traditional approaches to the space physics**

Our Pilot Investigations

- **Understanding of the EISCAT 3D data system**
- **Understanding of the EISCAT user perspectives**
- **Experiment with e-Science approaches**
  - **EGI + opensearch**
  - **EUDAT**
  - **DIRAC**
- **Testing with EISCAT 3D 1 polar year's data**

Our Future Directions

- **Staging services to ship scientific data from observatory networks into the EGI generic service infrastructure and get data off**
- **Cost effective large storage facilities + long-term archiving mechanism**
- **Comprehensive curation services**
- **Advance searching facilities**
- **Facilities to support new data scientists**

Our Achievements in the Pilot Study

- **Understanding of the problems**

- **Experiment with e-Science approaches (EGI, EUDAT, Opensearch)**

- **Identifying future requirements**

- **Establishment of a consortium with shared visions**
  - *13 organizations, 7 nations*
  - *Providers of data, storage, service + user community*

- **Raising of awareness via publications & presentations**
  - *2013 EISCAT Symposium, Aug 2013*
  - *Data Science Symposium 2013,*
  - *Radiovetenskap och Kommunikation 2013*
  - *EGI Technical Forum 2013, Sep 2013,*
  - *EGI Inspired Issue 13, Oct 2013*
  - *EGI Towards H2020 workshop, Dec. 2013*

- **Wiki, mailinglist, doc archive**

Our Aims in a H2020 Project

- **Objective1: A generic e-Infrastructure for Big Data**
  - **Identify common requirements, challenging issues, state-of-the-art design experiences, e.g., LOFAR, LHC, SKA, etc.**
  - **Proof of concepts of data infrastructure-enabling software**
- **Objective2: Support to EISCAT Community**
  - **Real-time data access**
  - **Community driven co-design**
  - **Virtual research environments**
  - **Support of long tail of scientist**
  - **Global data sharing and integration**
- **Objective3: Training of Data Scientists**
  - **A new data-centric way of organising research activities**
  - **New approaches to solve problems**
  - **Lead to significant scientific breakthroughs**

Our Visions and Missions

- **Provide a EU-level generic e-Infrastructure for big data**
- **Supports of ESFRI RIs, e.g, EISCAT 3D, others**
- **Integrate existing EU e-Science efforts (EGI, DIRAC, EUDAT, ESA)**
- **Focus on Big Volume Data**
- **Provide world-leading Big data technology**
- **Cost effectively :**
  - **Minimise the costs, maximise the benefits**
- **Support Science moves to the 4th paradigm**
  - **Data centric-> new approaches ->scientific breakthroughs**
- **Support the needs of data scientists**
  - **Community driven co-design**
- **Support the needs of long-tail scientists**

Organisations Involved

- **Cardiff University, UK**
- **CNRS, France**
- **CSC-IT Center for Sicnece, Finland**
- **DIRAC (via its partners)**
- **EGI.eu, The Netherlands**
- **EISCAT, Sweden**
- **EUDAT (via its partners)**
- **European Space Agency, Italy**
- **STFC Rutherford Appleton Lab, UK**
- **Umeå Universitete, Sweden**
- **University of Amsterdam, The Netherlands**
- **University of Barcelona, Spain**
- **University of Edinburgh, UK**