Digital RI for Open Science:

The National Practice in CSTCloud

Jianhui LI (lijh@cnic.cn)

Computer Network Information Center, CAS

CBAS deputy Director General

CODATA VP

2023.6.29

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

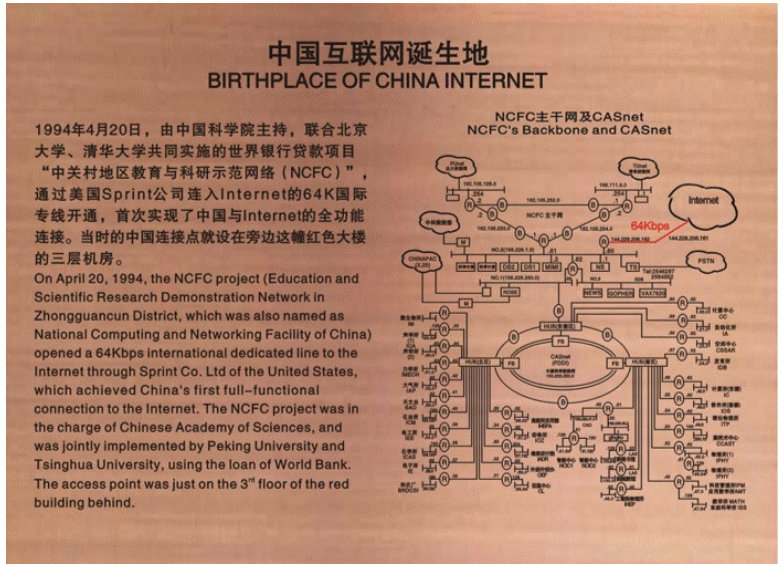
CNIC Introduction

- Computer Network Information Center (CNIC) is a special research institute of Chinese Academy of Sciences (CAS) to power scientific research and innovation, advance scientific research management, and promote public science outreach by advanced technologies. CNIC is the cradle of China's digital transformation and open science with data, network, computing and public outreach initiatives.

Hello Science | Hello World

Data Center
Located In Huairou District
Beijing, China

Main Campus
Located In Haidian District
Beijing, China



Requirements for E-infrastructure in CAS

Data-intensive & big data driven Science

Major Research Infrastructures

National scientific data centers

Field Stations all around China

Major international cooperation projects

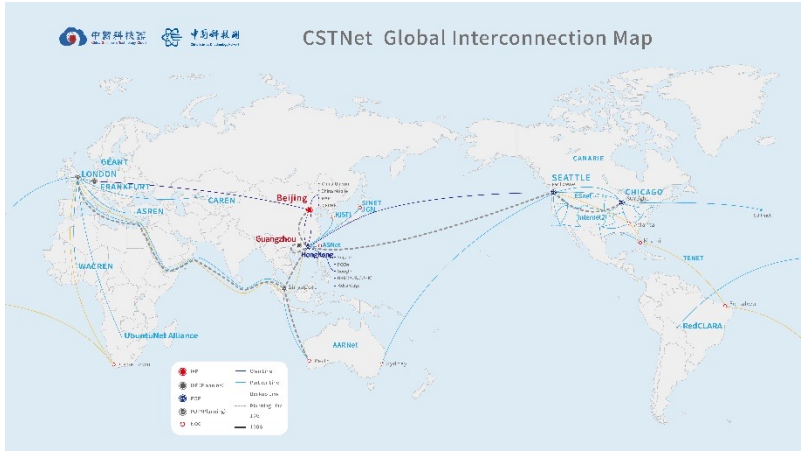
- Decades investment in cyber-infrastructure development and e-science in CAS, totally over 2 billion RMB

E-Infrastructures in CAS

NETWORK

中国科技网 China Science & Technology Network

CSTNet has 12 Branches across China, and provides services to approx. 1MM. end users from 300 institutions. Her domestic bandwidth is 131G and international bandwidth is 113G.Below shows the CSTNet Global Interconnection.

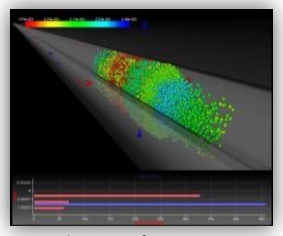


COMPUTING

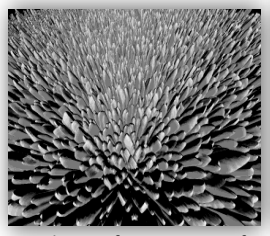
Chinese National Grid

CNGrid has connected 19 supercomputing centers, the capability of clustering computing has reached 200PF, supported for thousands of national science and technology program and key engineering. 40+ independent high performance computing software.

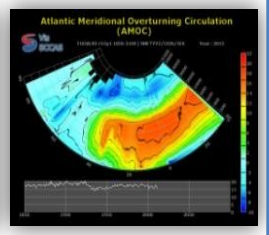
CAS Supercomputing Grid environment creates a three-layer architecture environment and has integrated more than 15PF computing resources from 35 CAS institutions



Visualization of ADS transmutation system



Simulation of optimization of titanium alloy structure



Simulation of the Atlantic ocean circulation

OPEN RESEARCH DATA

20 National Scientific Data Centers are launched by the Ministry of Science and Technology,P.R.C as the pilot to improve research data sharing following the national rules, “Measures for Managing Scientific Data”. Among all of which, 11 are from the Chinese Academy of Sciences.

No.	National Scientific Data Center
1	National high energy physics science data center
2	National genome science data center
3	National microbial science data center
4	National space science data center
5	National astronomical sciences data center
6	National earth c
7	National polar s
8	National Qingha
9	National ecolog
10	National materi data center

No.	National Scientific Data Center
11	National glacial frozen deserts science data center
12	National metrological science data center
13	National earth system science data center
14	National population health science data center
15	National basic sciences public science data center
16	National agricultural science data center
17	National forestry and grassland science data center
18	National meteorological science data center
19	National earthquake science data center
20	National Marine science data center

China National Policy

- Scientific Data Management Regulations issued by the State Council in March 17, 2018
  - intended to clarify the responsibilities of officials and scientists who regulate and use the information.
  - improve the management, security, accuracy and openness of scientific data,
- Scientists have to submit data to related National Scientific Data Center for archiving and open services
- Scientific data sets will also better identify their origins and researchers, allowing clearer citations and stronger protection of intellectual property

Established 20 National Scientific Data Centers

MOST launched 20 National Scientific Data Centers in 2019

Responsibilities of National Scientific Data Centers

-Integration of scientific data

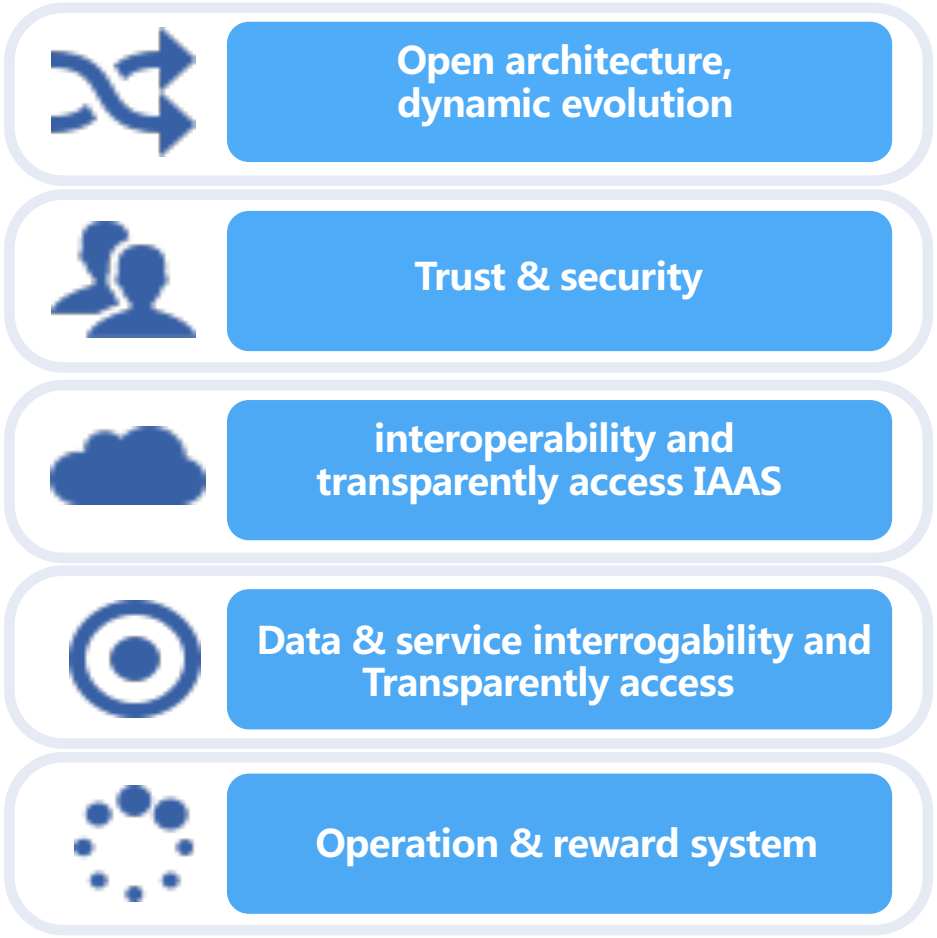
-Classification, processing and mining of scientific data

- Open and share scientific data

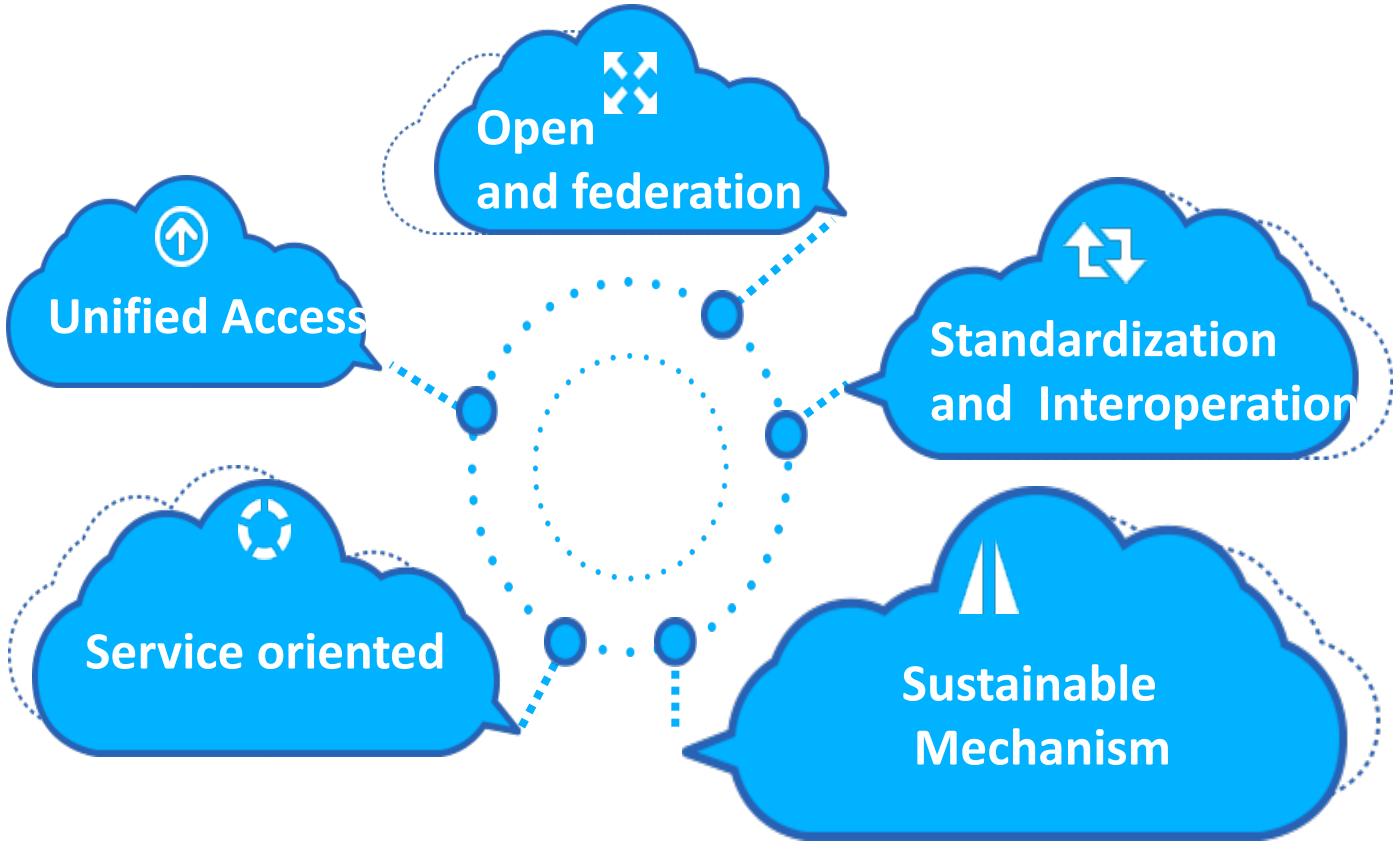
-Strengthen exchanges and cooperation in scientific data nationally and internationally

Integrate distributed e-infrastructure for Open Science

- Open science infrastructure features: **Federated, Accessible, Internationally Interconnected, Interoperable** (UNESCO, 2021).



challenges



principles

CSTCloud: A national research e-infrastructure

- CSTCloud supports multidisciplinary open scientific researches with integrated cloud services for the discovery, usage and delivery of S&T resources.
- Continually funded by CAS Informatization Program(2017-2020,2020-2025)

2016.12
- 2016,12 The State Council issued the National Informatization Plan for the 13th Five-Year Plan, "CSTCloud" was included in the priority action

2017.01
- 2017.01 The Chinese Academy of Sciences issued the "13th Five-Year Plan for Informatization Development of the Chinese Academy of Sciences, establishing the task of "CSTCloud"

2017.12
- 2017,12 Launching Ceremony of "CSTCloud"

2018.04
2018.04 CSTCloud portal launched

2018.11
- 2018.11 CSTCloud Phase I Construction Task, passed the review (focus on resource performance construction)

2019.04
- 2019.04 CSTCloud Phase II Construction Plan Passed the review (focus on convergence capability and service)

2019.07
- 2019.07 Resource Aggregation Workshop

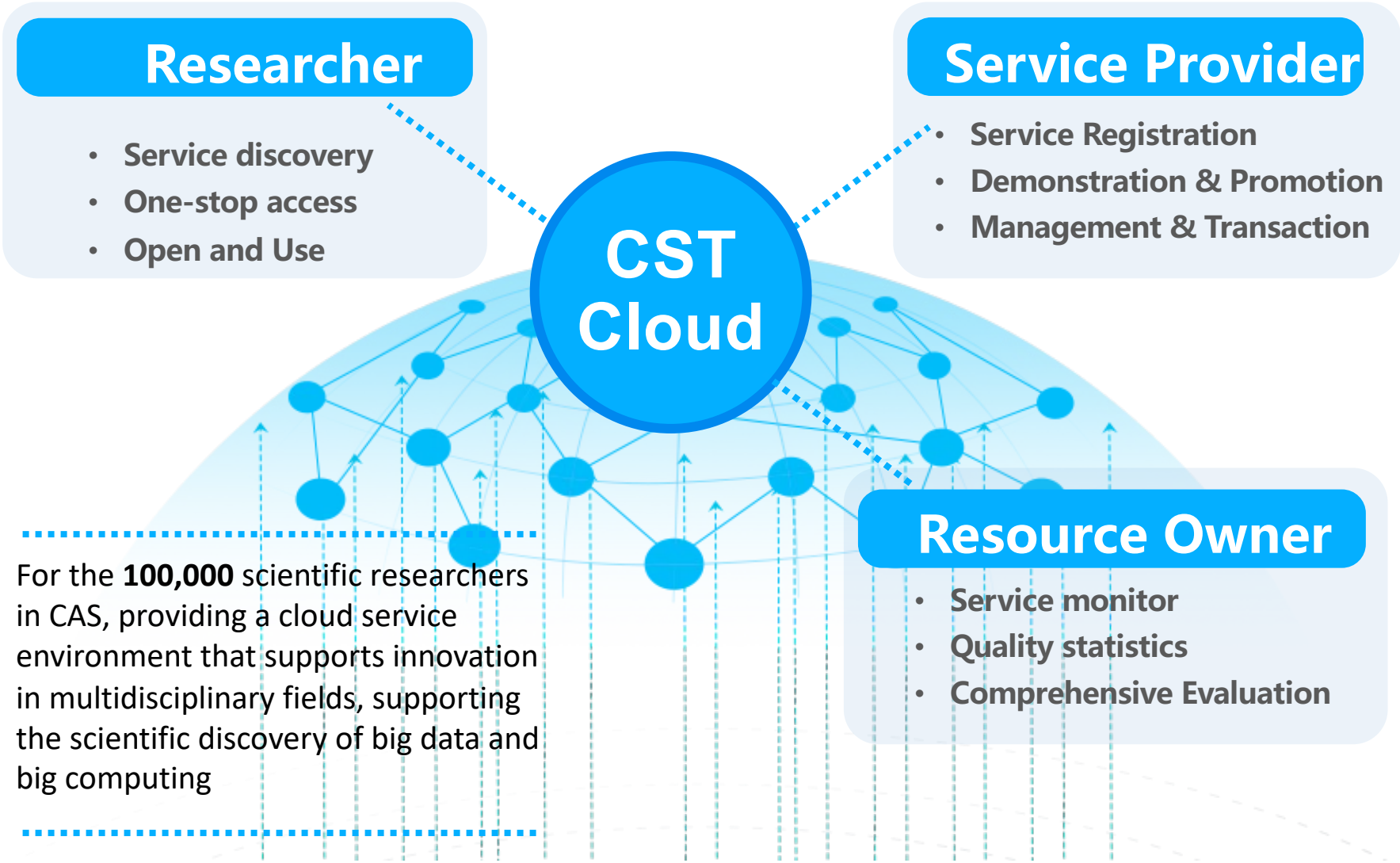
2019.12
2019.12 CSTCloud 2.0 launched

- 2019.06 MoU Signed between CNIC & EGI.eu, mobilizing and co-design of GSOC Initiative

2020.6

2020.11
- 2020.11 GOSC workshop held only jointly by CNIC, EGI and CODATA.

CSTCloud: Targeted users and tailored services



For the **100,000** scientific researchers in CAS, providing a cloud service environment that supports innovation in multidisciplinary fields, supporting the scientific discovery of big data and big computing

CSTCloud Service Catalogue and System

AI Operation and Maintenance System

National Large Research Infrastructure
National Science Data Center
National Field Scientific Observation and Research Station

Space, Astronomy, High Energy Physics, Microbiology ... Thematic Cloud

Information Technology Innovation and Research Demonstration Application

SAAS

- New ARP
- Website Group
- Science Communication
- CAS Mail System
- Cloud Conferencing
- Cloud Box
- Conference Service Platform
- Overleaf
- Open Source Community

PAAS

- CSTCloud Passport
- CSTCloud AAI
- Jupyter Hub
- Cloud Database
- CAS Science Data Center
- National & CAS Grid
- National Basic Data Center
- eduroam CN Root

IAAS

- Domestic and International Network
- HPC & AI Heterogeneous System
- Federated Cloud and Storage

Security Monitoring and Protection System

CSTCloud: Online Services

Resources & Services

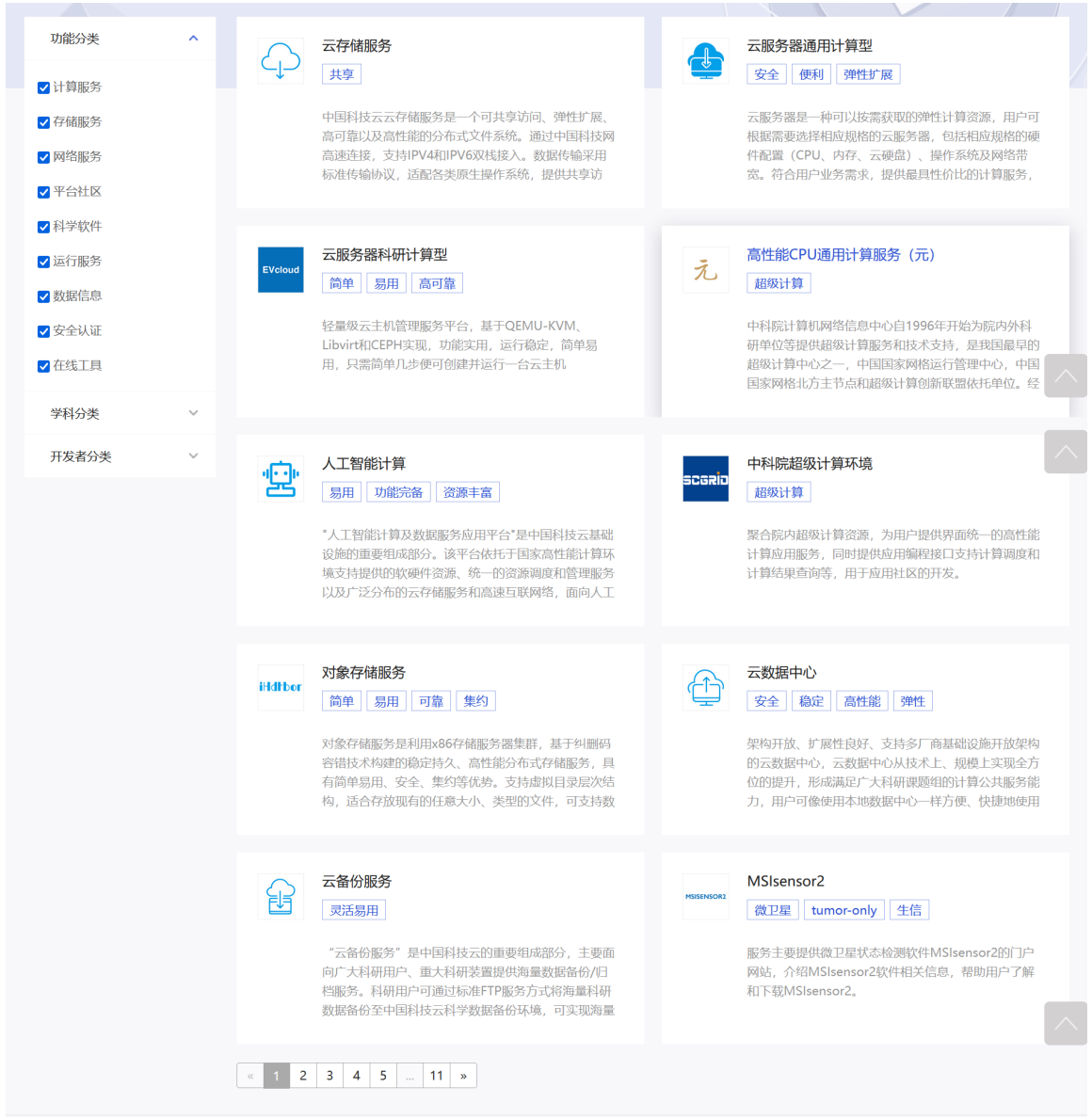
Computing Resources： **315**Pflops

Data Resources： **PB-class**

Storage Resources： **150PB**

Software Resources： **1000+**

Network Resources： **100G**



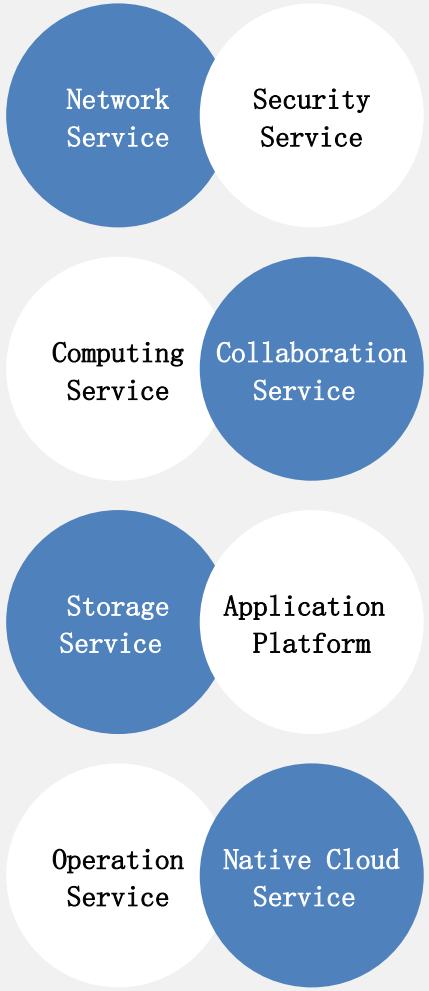
Service Catalogue

01
- Internet Access eduroam
- Data Private and VPN
- Field Data Transmission Network
- International Data Transmission Network

02
- Science Big Data Cloud Service
- Advanced Security Cloud Service
- HPC Federated Cloud Service
- Federated Cloud Service
- CN Grid and CAS Grid

03
- Science Data Storage and Archiving
- gStore
- Cloud Box
- Cloud Database

04
- Intelligent Network Operation System
- HPC Operation System
- Big Data log analysis System
- Integrated Information Monitoring Platform



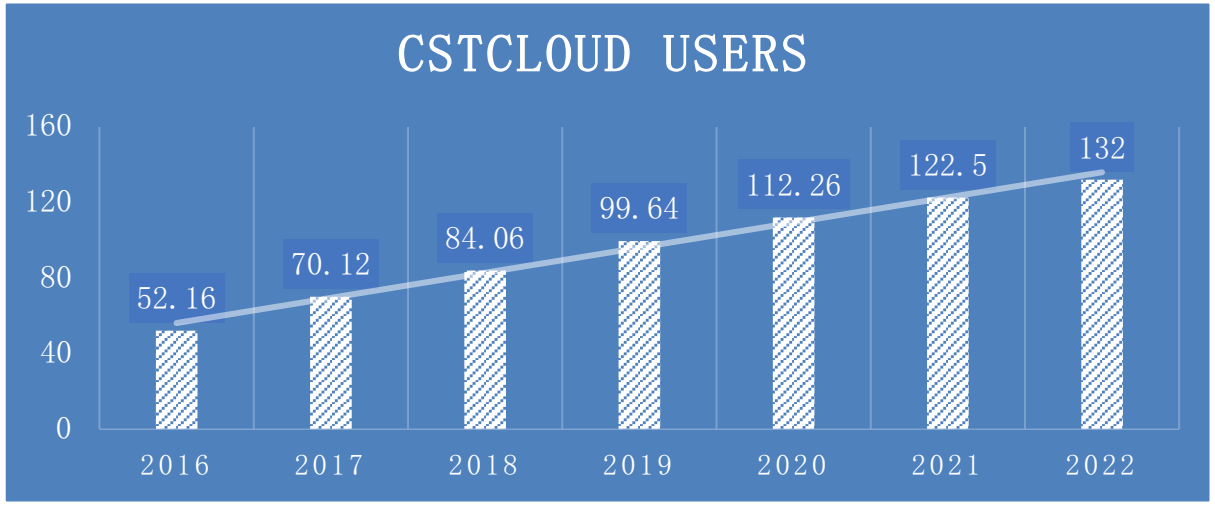
05
- CSTCloud Passport
- CSTCloud AAI

06
- CAS Mail System
- CAS Cloud Conferencing
- Conference Service Platform
- Overleaf
- Indico

07
- Open Source Community
- Large Research Infrastructure Sharing Platform
- Instrument and Equipment Sharing Platform

08
- JupyterHub
- Code Run
- Draw.io
- FAAS Service
- Grass
- ...

CSTCloud Service Application and Analysis



- CSTNet
  - About 300 institutes , 1 million users
  - 100G domestic and 100G international interconnection
  - The best quality in China Network Carriers
- HPC
  - 200PFlops Storage 100GB
  - >600 users
  - >1.7 million jobs

- Cloud Computing and storage
  - >1000 cloud hosts
  - 100PB Storage
  - 12 data centers federated
- CAS Passport/Check-in
  - >1200 applications and 1.32 million registered users
- eduroam
  - 400 institutes and universities

CSTCloud: Open Data services

CAS Data Cloud

CAS Data Cloud integrates resources and services from the perspective of three major types of services: infrastructure, data resources, and application platform. Up to now, the cloud platform has integrated nearly **960.14 TB** of shareable scientific data, with **132,500** users, **178,863,200** cumulative online visits, and **22,867,200** cumulative data downloads.

Data set： 18369 **Online Resources：** 3.11 PB

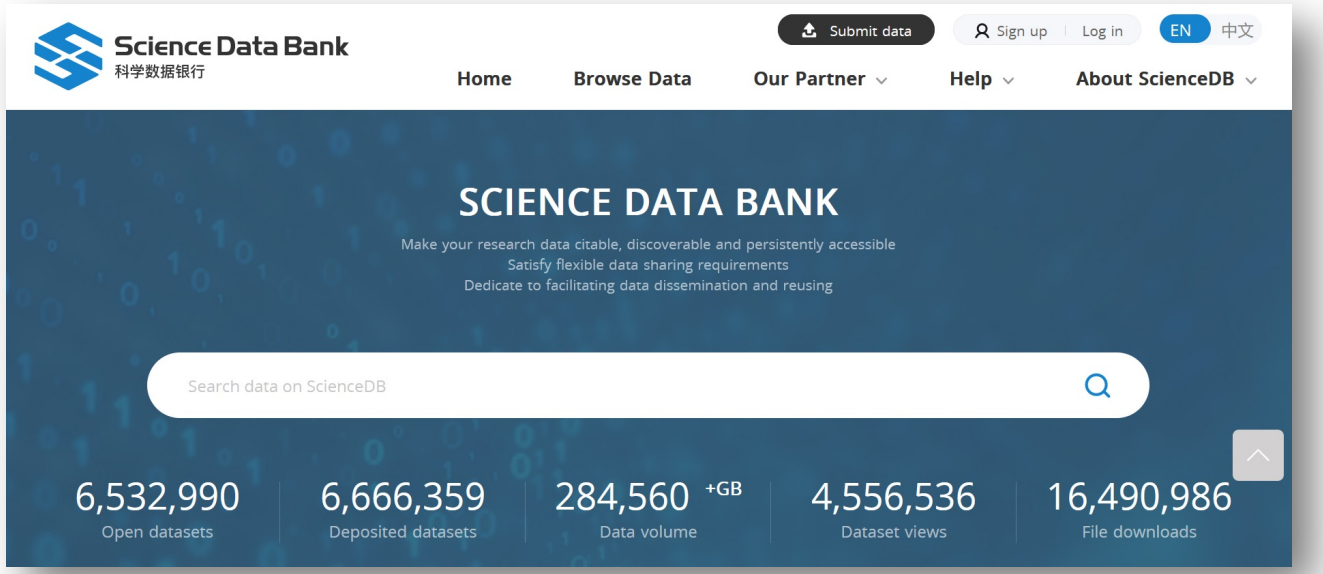
Data Center： 35 **Total number of visits：** 116 million

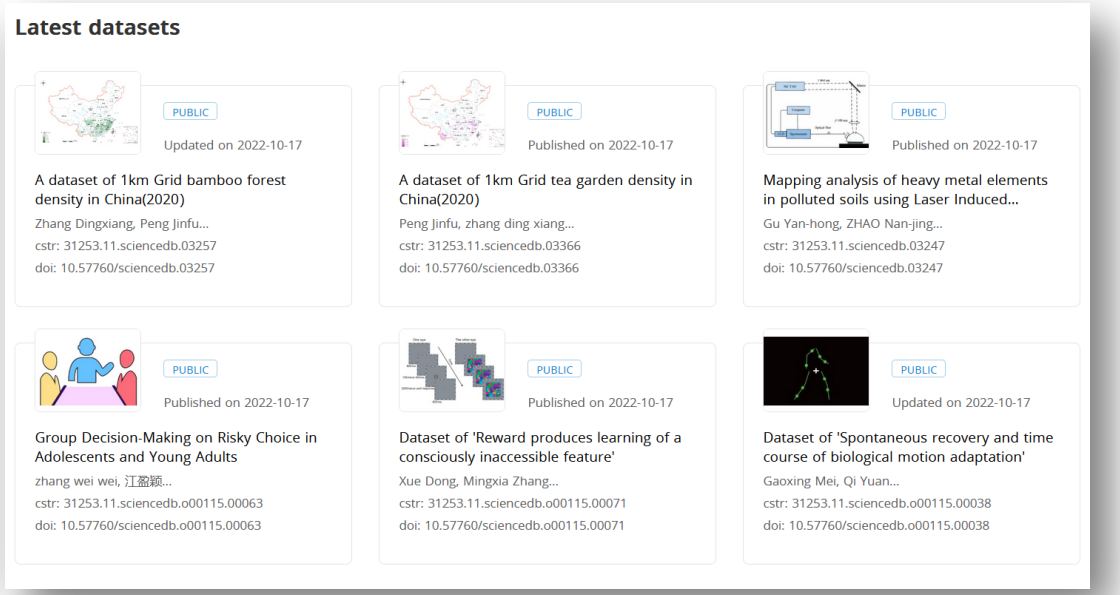
RESOURCE CATEGORY：

CSTCloud: Research Data repository

ScienceDB http://www.scidb.cn/en

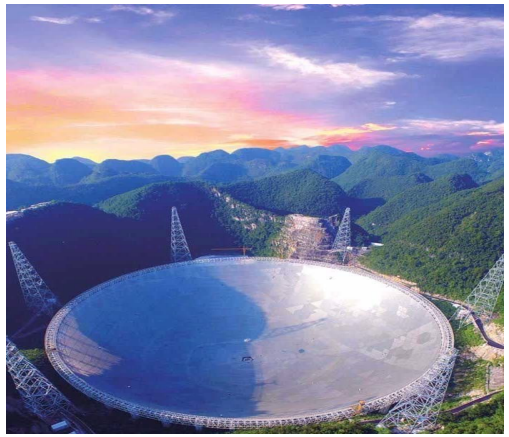
Science Data Bank is a public and non-profit generalist data repository which is developed and maintained by the CNIC, implementing the FAIR principles of data sharing. The repository accepts submissions from the entire scientific community across disciplinary boundaries and file formats, and publishes all kinds of research data with no charge of fee.





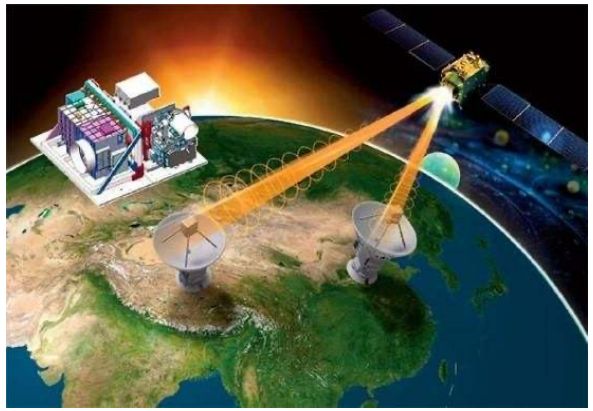
CSTCloud for big science research

FAST

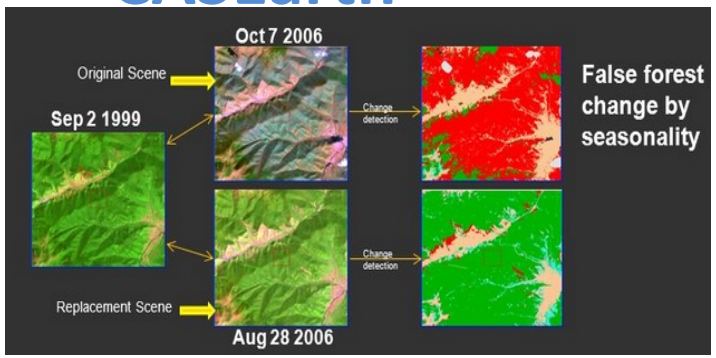


Wukong

Mozi



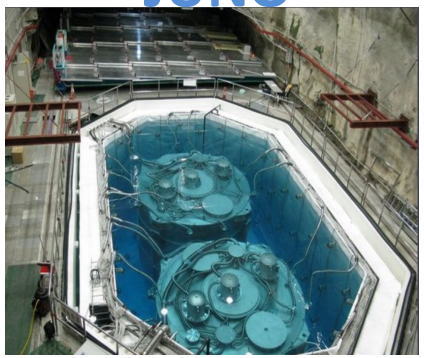
CASEarth



iter

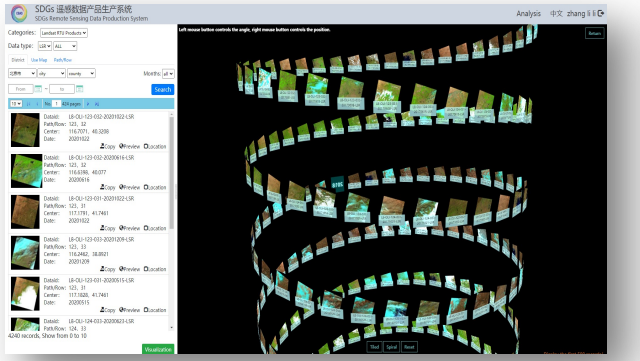
the way to new energy

JUNO

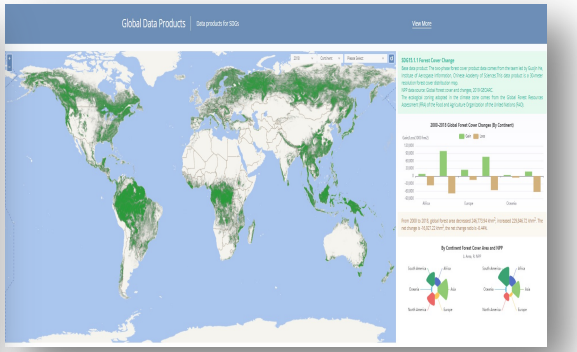


CSTCloud for SDGs

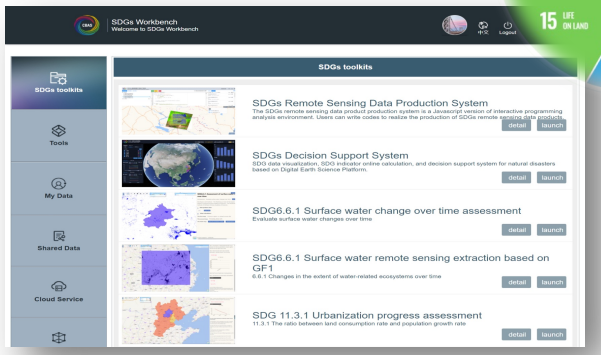
- **The *SDG Big Data Platform*** provides rich data resources and cloud services for SDGs Decision Making and science discovery, that
  - With **10PB data** covering multiple subjects, such as geography, remote sensing, ground monitoring, and social statistics;
  - **Characterize and Profile Scientific Workflows** featuring lifecycle Big Earth Data management, including massive data storage, curation, computation, analysis and visualization;
  - Integrate **over one hundred algorithms and tools** for advanced data analysis and management;
  - Provide **one-stop cloud services** through the **bilingual portals to** support the UN Sustainable Development Goals.



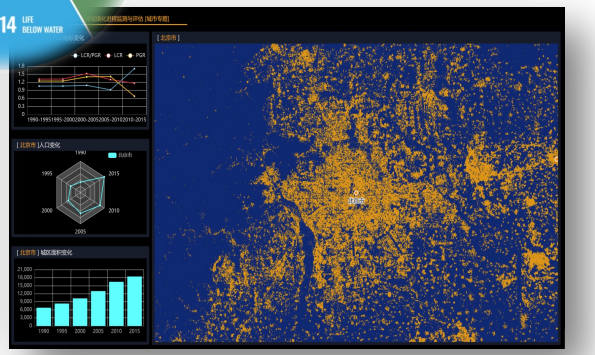
Databank Remote Sensing Data Engine



SDGs Data Products



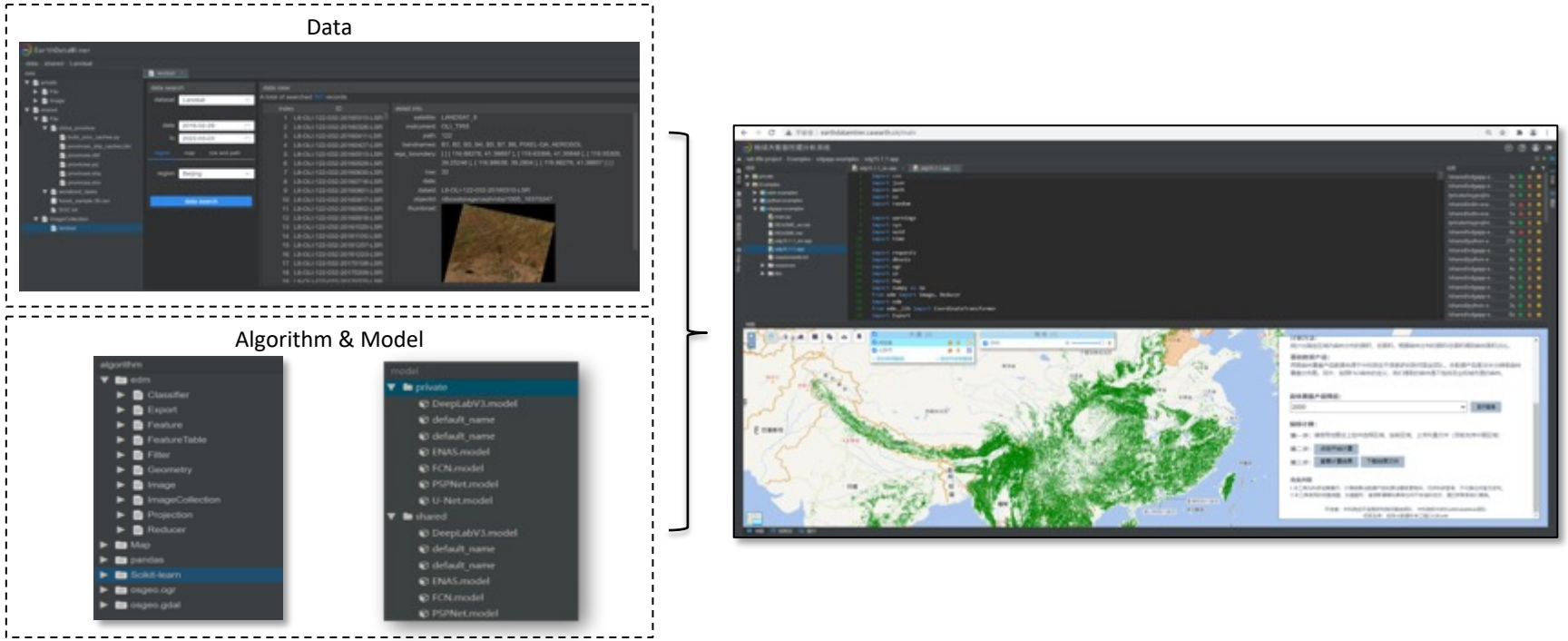
SDGs workbench



SDGs Decision Making

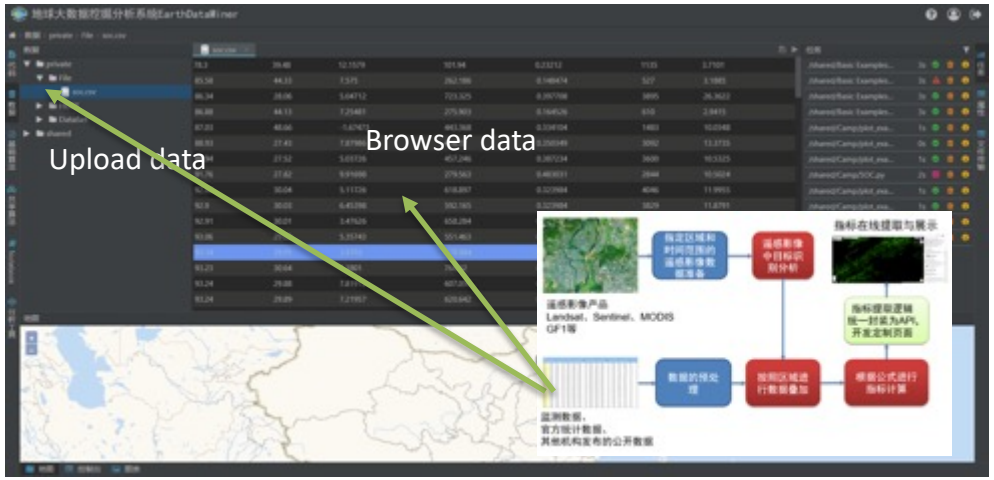
Data Analysis As Service: EarthDataMiner

- Online interactive data analysis environment
  - Integrated with DataBank: upload models, select data, and process data products through instruction operations
  - Algorithm & Model library: more than 150 algorithm developed and provide cloud service: FAAS(Function As A Service)
  - Web IDE: supporting users to write data analysis code (Python) online

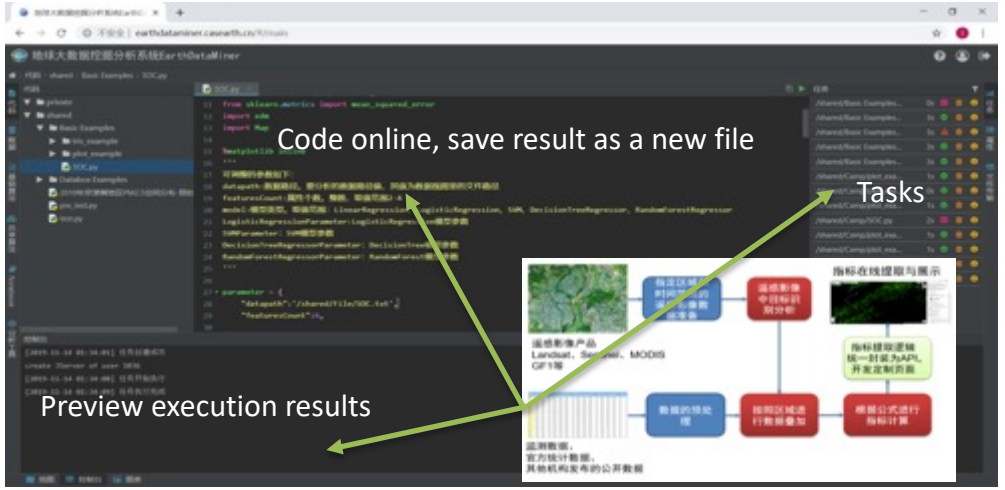


SDGs indicator on-demand computing

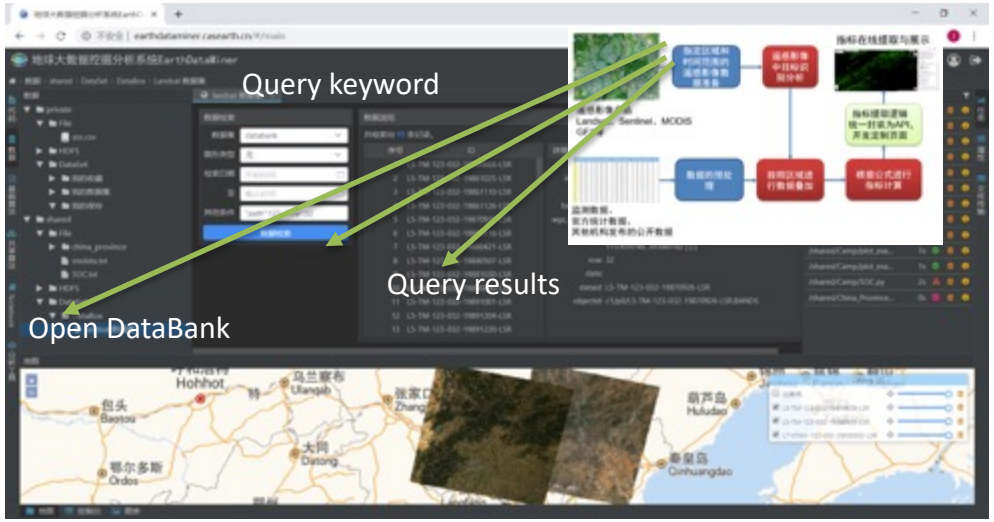
1. Upload data to the EarthDataMiner system



2. Write python code to preprocess data



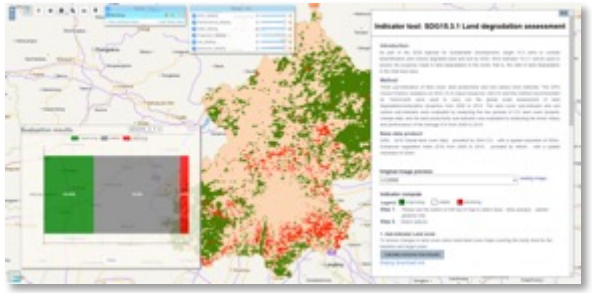
3. Search radar data from Databank develop by CASEarth



4. Compute the SDGs Indicator and Visualizaiotn

SDGs indicators on demand analysis Tools

SDG15.3.1 Land degradation index assessment



SDG6.6.1 Surface water change over time indicator

SDG13.2.2 Annual average CO2 concentration evaluation

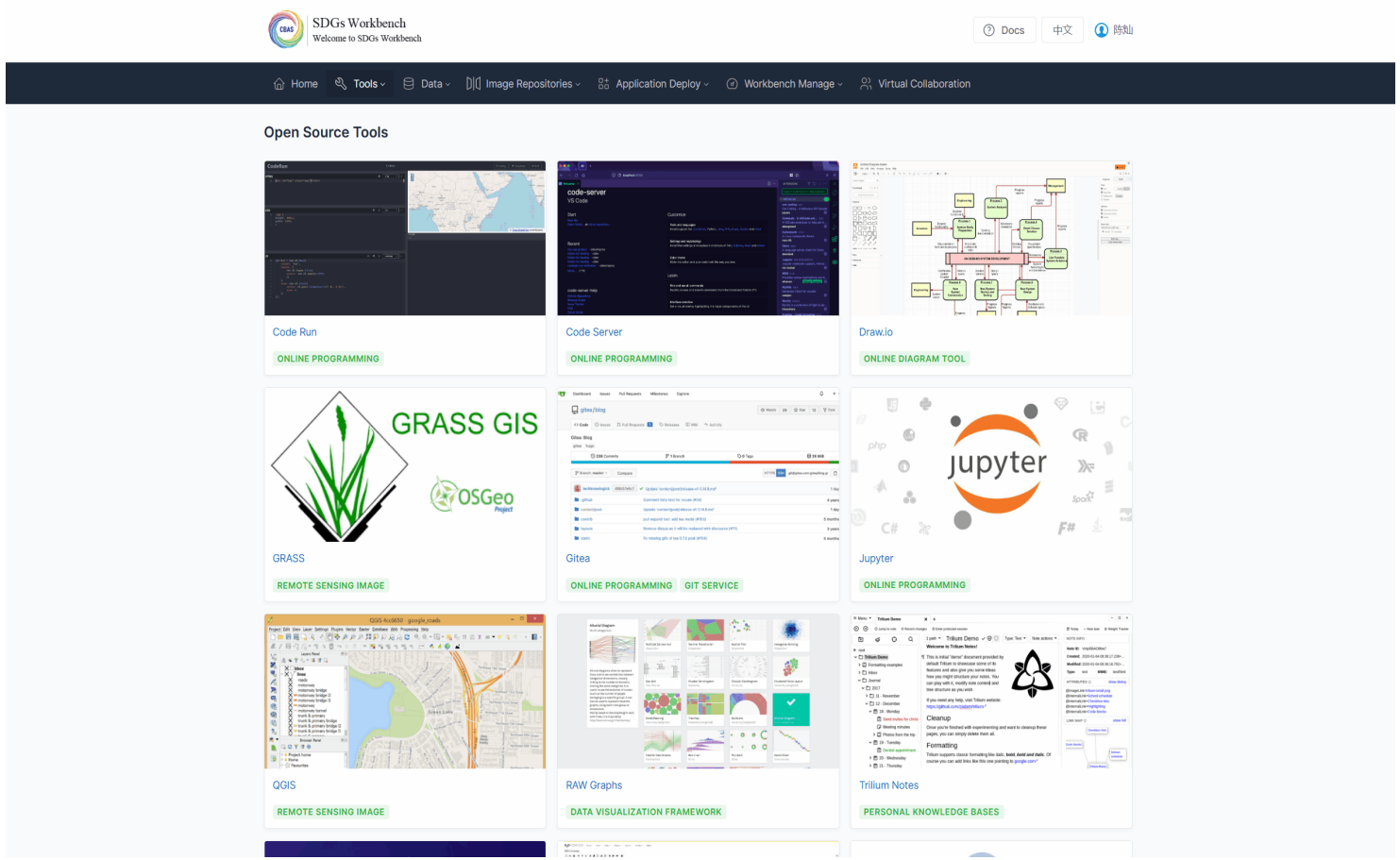
SDG13.1.1 Natural disaster impact evaluation

SDG15.1.1 Forest coverage detection

SDG11.3.1 Global urbanization index monitoring

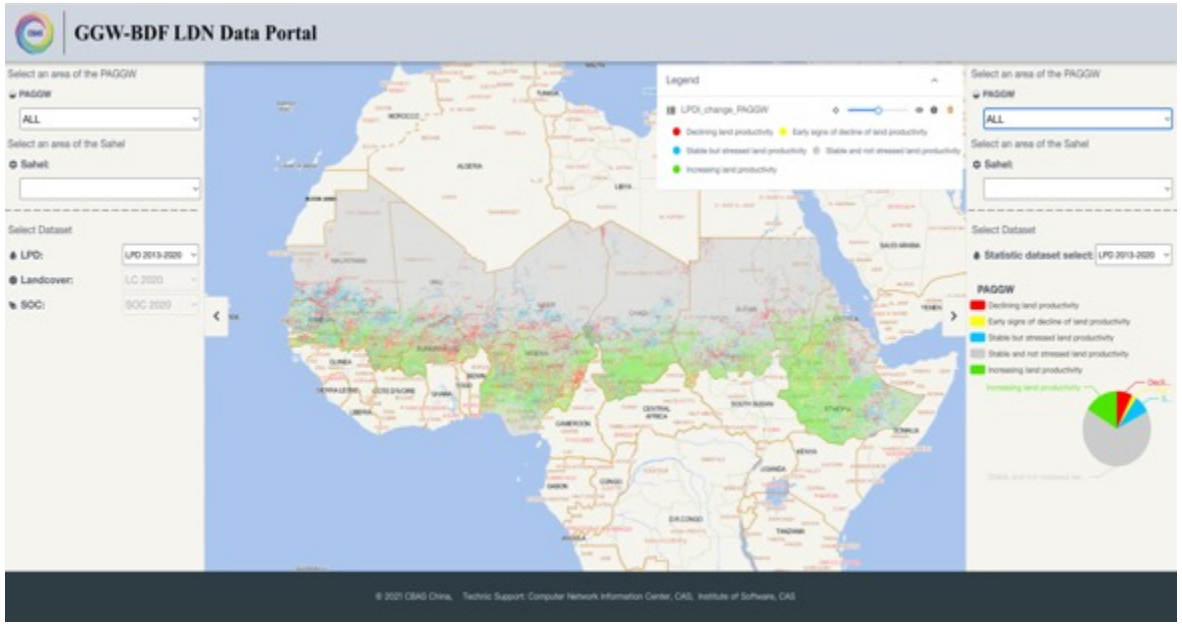
Converged Virtual Research Environment : SDG Workbench

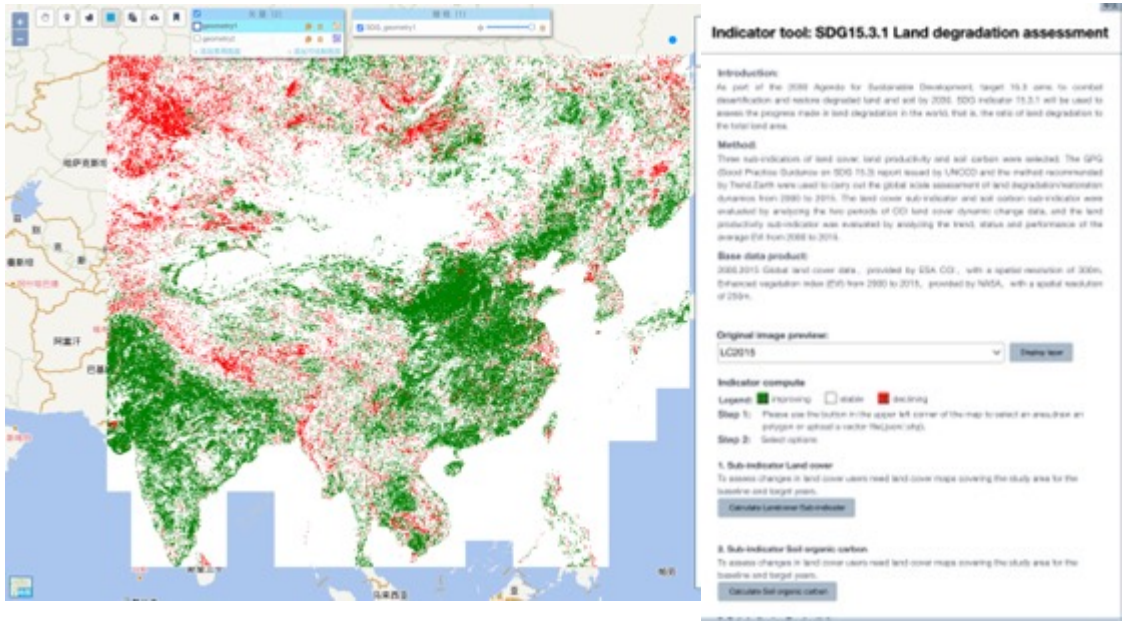
- Data be accessed by applications/machine Transparently
- SDGs Tools
  - ✓ Integrated Tools
  - ✓ Data Analysis Tools
  - ✓ Data Products Tools
- Open-source Tools
  - ✓ Developing
  - ✓ Machine Learning
  - ✓ Data Visualization
- Creating / Using / Releasing Spark Cluster On-demand
- Cloud-Native DevOps CI/CD
- Virtual Collaboration
  - Setup a virtual team or virtual organization



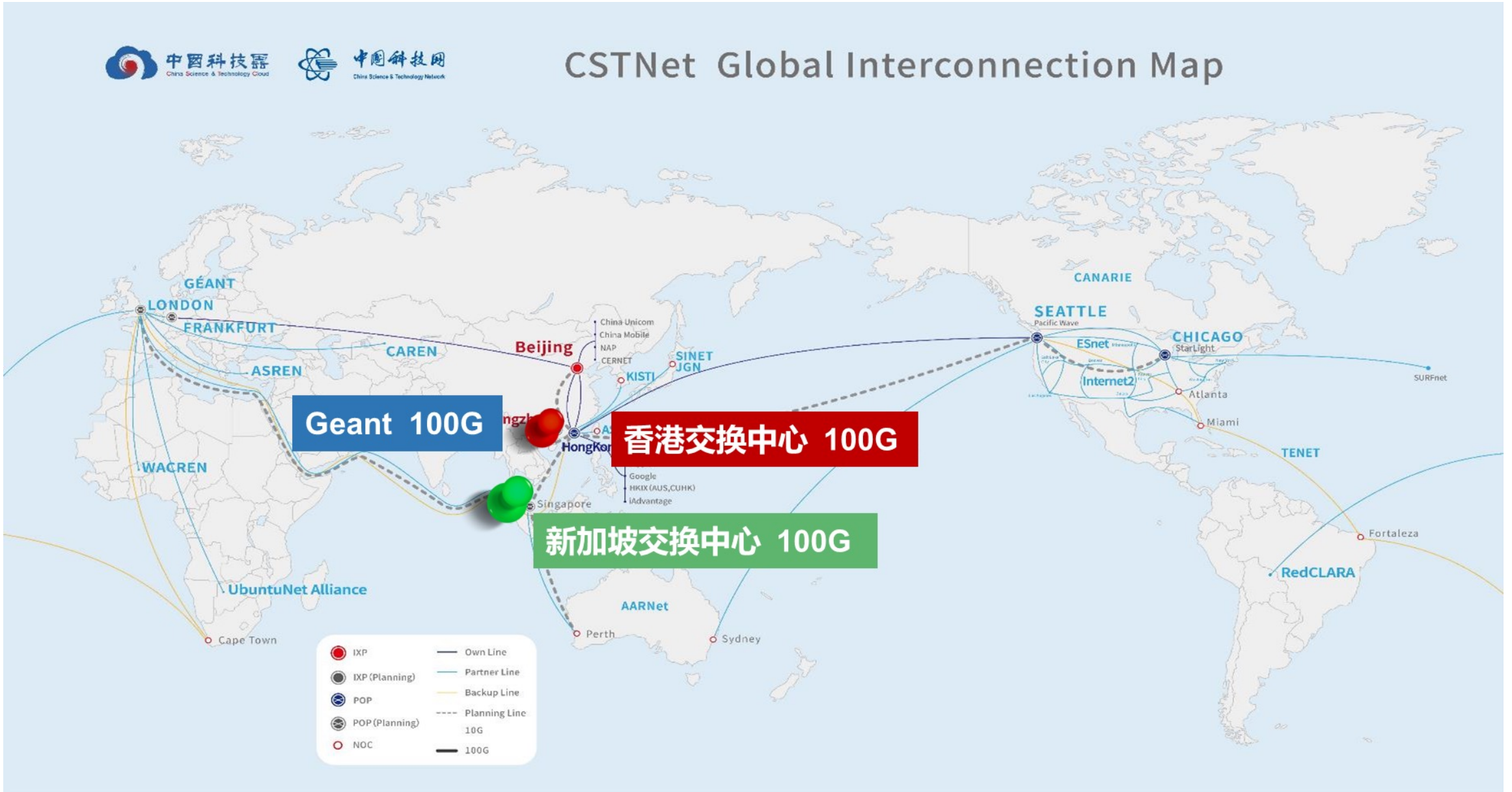
Example: GGW-BDF & SDG15.3.1 LDN

- provide high spatial resolution datasets on land cover, land productivity and soil carbon change to better track the progress of LDN.



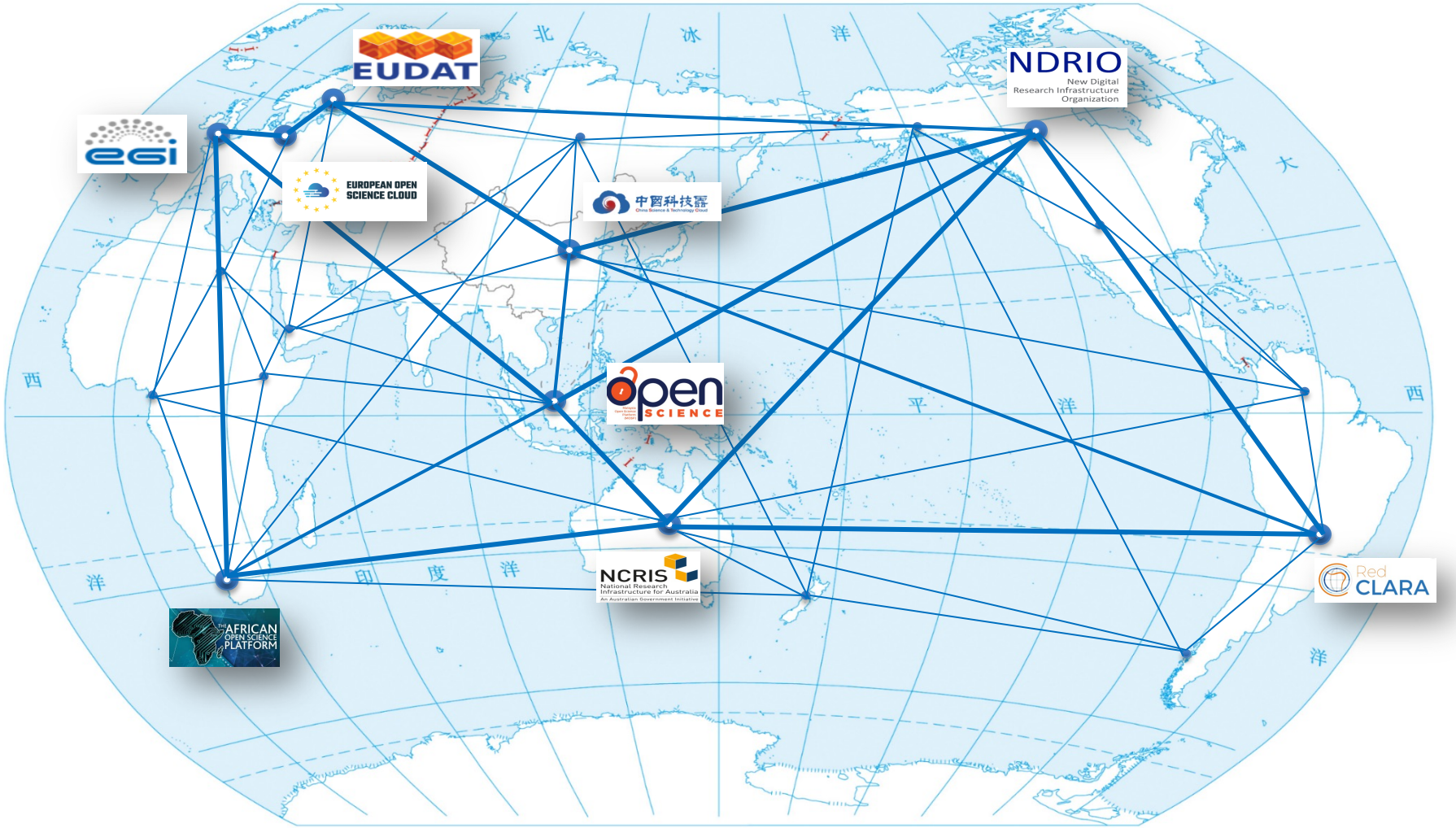


CSTNet Global interconnection



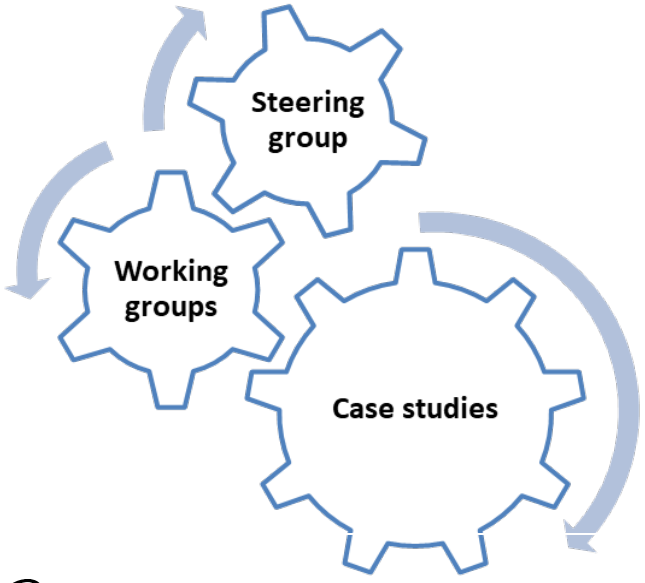
Connecting CSTCloud with the World

- **A cross-continental federated e-infrastructure and virtual research environment for global cooperation and open science.**

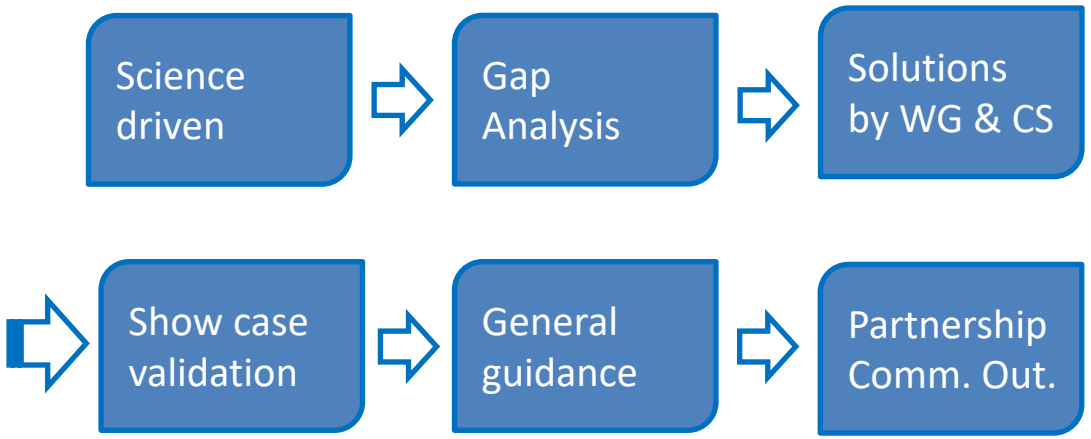


The Global Open Science Cloud (GOSC) Initiative

- **The Global Open Science Cloud** is a robust network to bridge trusted research e-infrastructures among which different research resources are connected and all stakeholders are linked for innovative science discovery in the dynamically evolving global open science environment.
- Important early project as part of ISC CODATA Decadal Program with seed funding from CAS.



3 Groups guiding the GOSC co-design



6 steps towards GOSC co-development

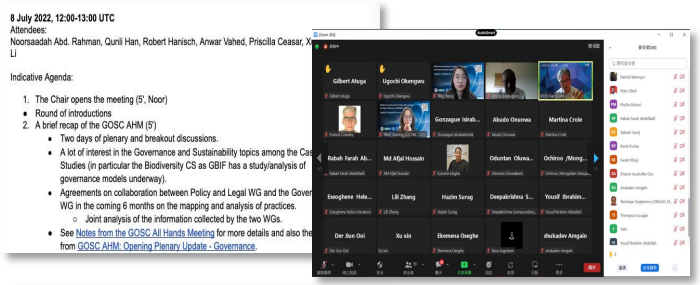
CODATA
International Science Council
UNESCO

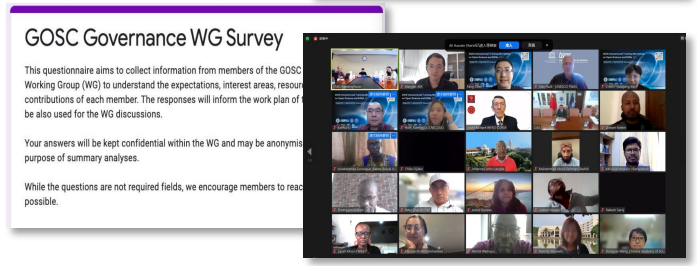
GOSC: Working Groups

Strategy, governance and sustainability

Qunli Han

Noorsaadah Abdul Rahman





Technical Infrastructure

Jianhui Li

Mark Dietrich

Happy Sithole

GOSC 2nd AHM: Updates from Working Groups and Case Studies

Technical Infrastructure WG

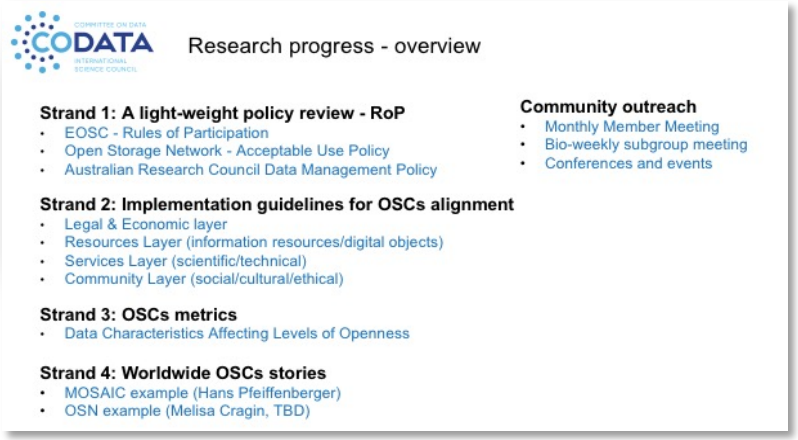
What are your main achievements since early June?

- Continued work in Sub Working Groups to develop technical reference frameworks for key functions:
  - AAI (Authentication, Authorisation and Identification)
  - Data Access/Storage Authorisation
  - Service and Data Discovery
  - Remote Compute Execution/Authorisation
- Interactions with sister WGs to identify common or overlapping challenges
- Developing an Overall Technical Framework for GOSC – to guide creation of testbeds

Policy and legal

Lili Zhang

Sarah Jones



Data Interoperability

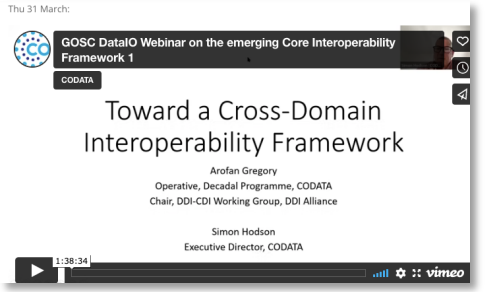
Lianglin Hu

Christine Kirkpatrick

Milan Ojsteršek

Pascal Heus

Natasha Simons



GOSC Data Interoperability WG

Case Study template

Case Study:	ICSM ANZ Metadata Working Group (MDWG) https://www.icsm.gov.au/what-we-do/metadata-working-group
Participant name, affiliation	Irina Bastrakova (ICSM / ANZ), Melanie Barlow (ARDC), Lesley Wyborn (ANU), Rowan Brownlee (ARDC)
Participant email address	Irina.Bastrakova@ga.gov.au Melanie.Barlow@ardc.edu.au Lesley.Wyborn@ardc.edu.au Rowan.Brownlee@ardc.edu.au
Date	Wednesday 4 May 2022

GOSC: Case Studies

Incoherent scatter radar data

Ingemar Häggström

Xin'an Yue

Biodiversity and Ecology

Joe Mille

Zhishu Xiao

Diffraction Data

John Helliwell

Genji Kurisu

SDG-13 climate change and natural disasters

Gensuo Jia

Monthip Sriratana

Bapon Fakhruddin

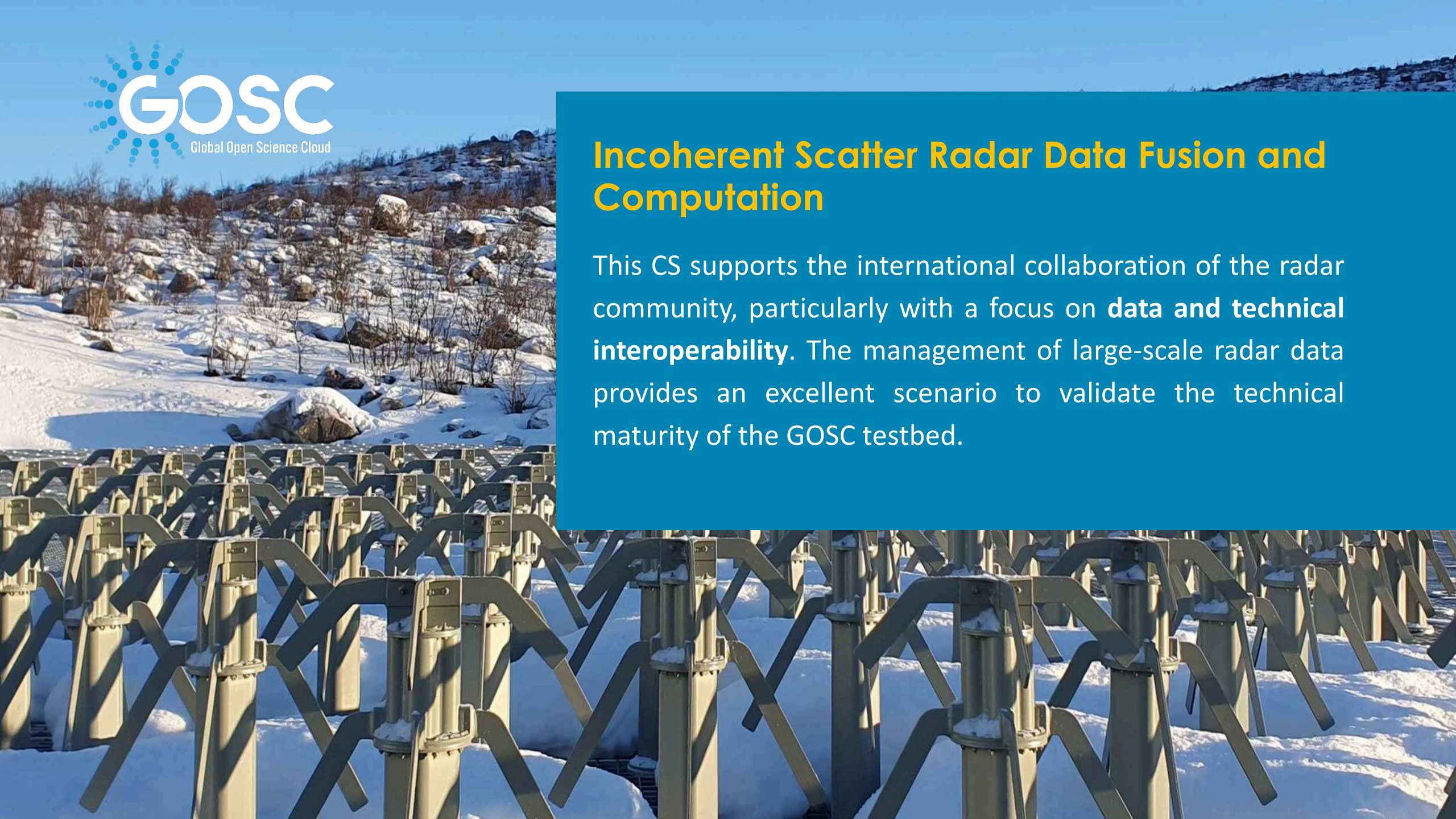
Sensitive data in population health

Jildau Bouwman

Lei Liu

Lauren Maxwell

Francisca Oladipo



GOSC
Global Open Science Cloud

Incoherent Scatter Radar Data Fusion and Computation

This CS supports the international collaboration of the radar community, particularly with a focus on **data and technical interoperability**. The management of large-scale radar data provides an excellent scenario to validate the technical maturity of the GOSC testbed.

Integrating Machine-derived Camera-trap Data into the Global Shared Science Networks

Through a collaborative platform for global camera-trap data sharing and analysis service, this CS will contribute to global biodiversity research, especially focusing on **distributed big data management, intelligent analysis, and cloud computing** for high-quality integration and optimization for camera-trap data management.

GOSC
Global Open Science Cloud

Open Reproducible Raw Diffraction Data for Access in Pandemics

This CS would lead to **single, definitive, protein models** derived from their raw diffraction data sets, which is important to the crystallographic community and the broader research community for drug discovery, especially for the pandemic crises that COVID-19 poses for society.

GOSC
Global Open Science Cloud

Sensitive Data Federation Analysis Model in Population Health

This CS seeks to demonstrate better ways of sensitive data sharing. Commonly agreed FAIR implementation profiles will be created based on the **FAIR data points** established in various 'GOSC' regions. The feasibility of distributed analytics over datasets held in various regions will then be explored and demonstrated**. Community-accepted standards** will be used throughout to facilitate the implementation of this CS.

GOSC
Global Open Science Cloud

From Raw Biodiversity Data to Operational Indicators through the Essential Biodiversity Variables (EBV-OSC)

The objective of this CS is to operationalize EBV indicators by targeting the highest levels of FAIRness (Findable, Accessible, Interoperability, Reusable) for both **data and source code implementation**, so that data and tools can be widely shared and reused.

GOSC
Global Open Science Cloud

SDG-13 Climate Change and Natural Disasters

CASEarth for Sustainable Development Goals (CASEarth4SDGs) is a platform system of data sharing and online computing for monitoring, measuring, and evaluating SDG indicators. Supported by this system, this CS mainly focuses on climate change and natural disasters in the SDG-13 field, addressing **technical, semantic, and policy interoperability to support decision-making.**

GOSC: International Programme Office (IPO)

- Expanding GOSC cooperation network and maintaining membership
- Facilitating conferences, training workshops and other events to increase social visibility
- Coordinating community outreach, day-to-day operation and future development activities of GOSC

GOSC IPO CNIC Team

Lili ZHANG
CNIC

Xueting Li
CNIC

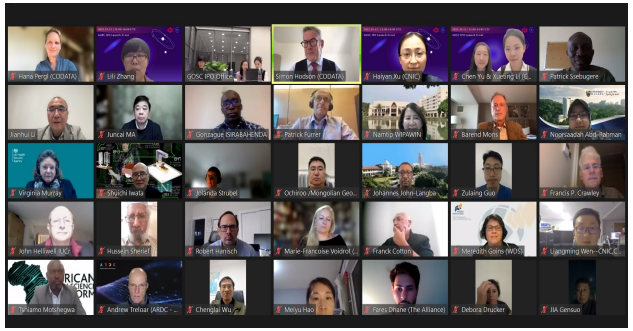
Simon Hodson
CODATA

Hana Pergl
CODATA

GOSC IPO CODATA Team

First IPO in CODATA's History

International Office located at CNIC



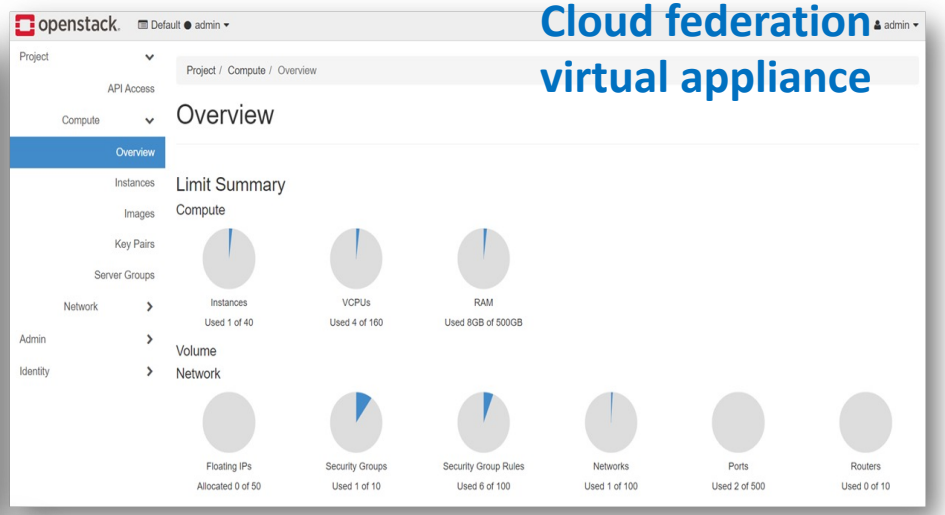
IPO Launch Event (10.12)

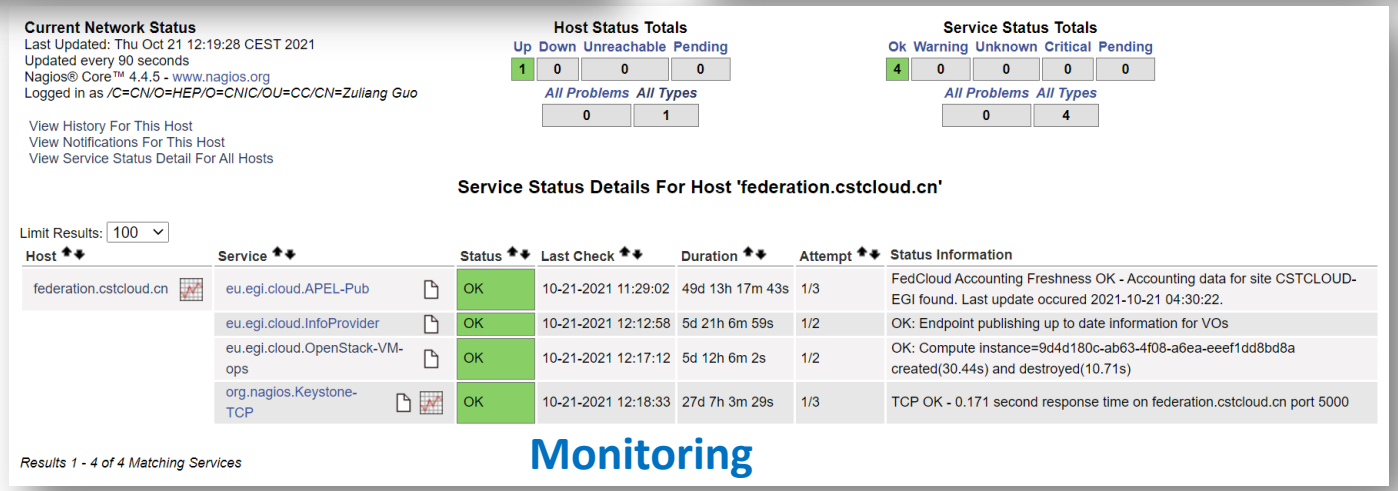
CSTCloud Federated with EGI

- **Technique solution has been verified , ready to provide service**
- **Shared resource is limited, need more resource…**

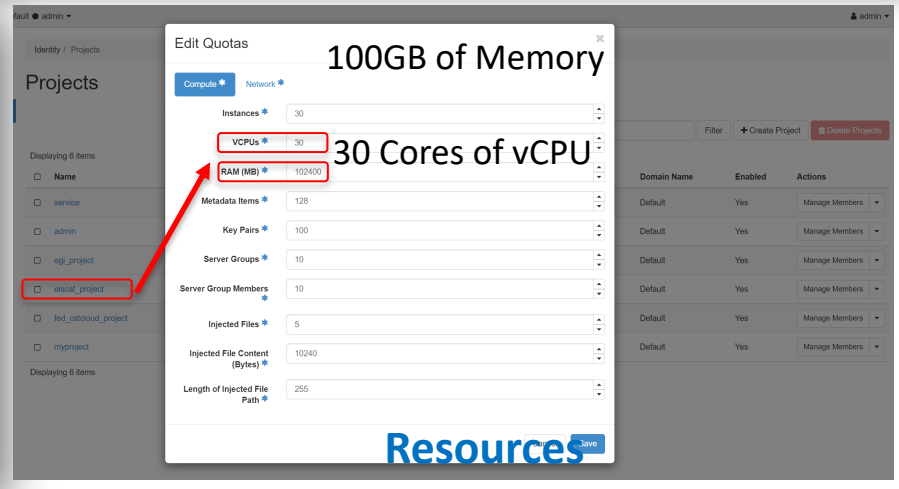
Secure AAI

Cloud federation virtual appliance



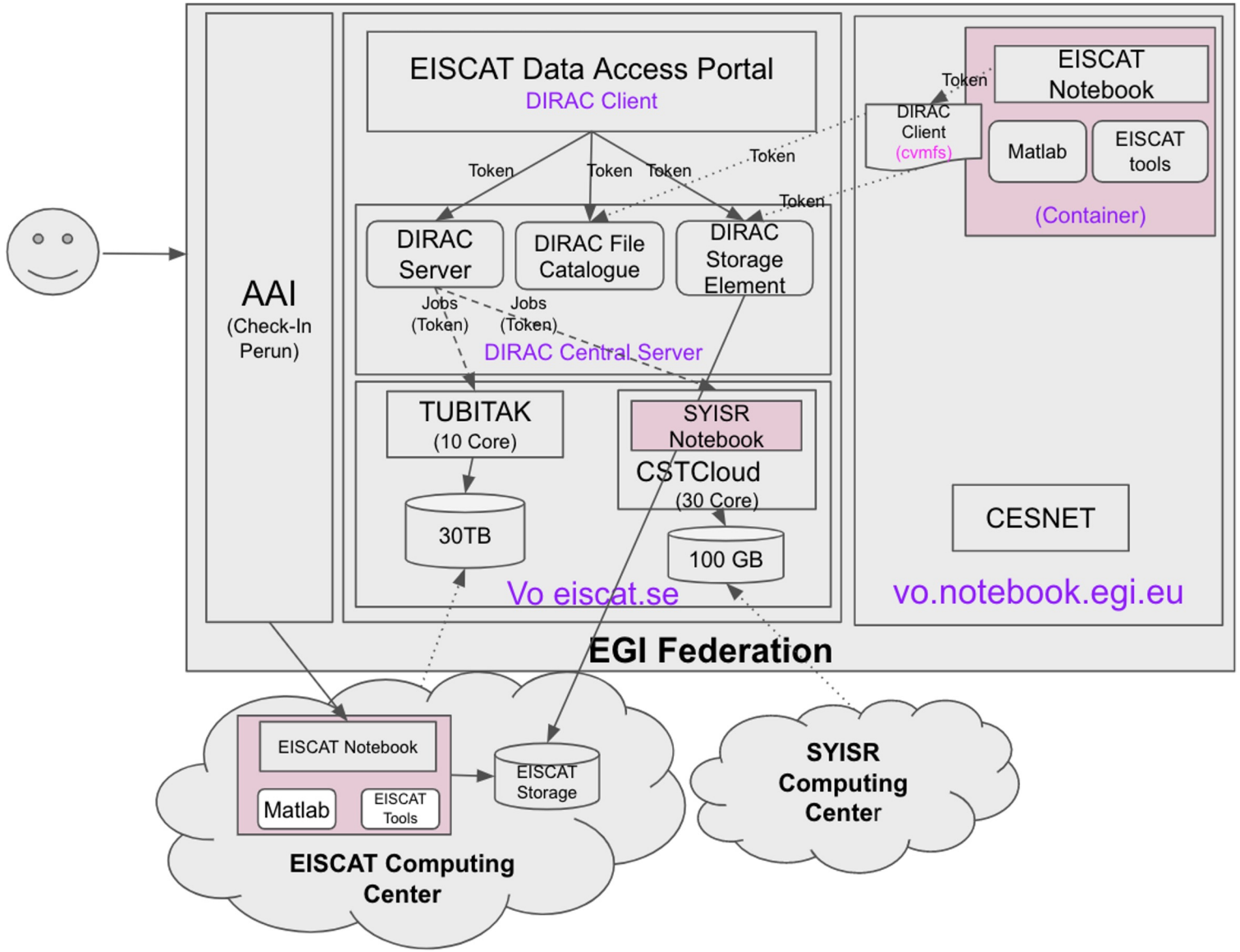


Monitoring



Resources

A Typical Use Case – 3D Radar Data Federation analysis

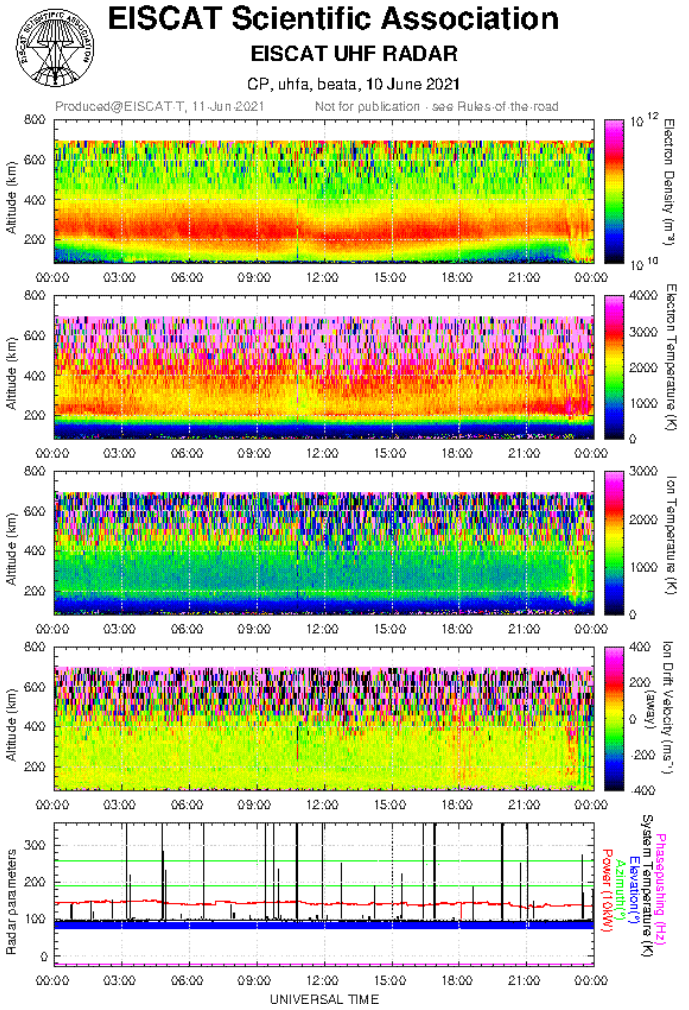


- **EISCAT Data Access Portal** (DIRAC)
  - User access via Check-In/Perun
  - Access token is passed to DIRAC File Catalogue (by DIRAC Client)
  - DIRAC File Catalogue return token with user information
  - (DIRAC Client) access DIRAC Storage Element with token
  - DIRAC Storage element enables search of EISCAT data (in the EISCAT storage)
  - DIRAC Server submits Jobs to Cloud
    - TUBITAK (10core, 30TB) + CSTCloud (30Core, 100GB)
    - Token access to the Cloud is ongoing
- **EISCAT Compute Center**
  - Jupyter notebook
    - User access via Perun
    - Matlab enabled
    - Not yet access to DIRAC Storage Element – cannot perform search, and have lower performance
  - EISCAT Storage: EISCAT data
- **EISCAT Notebook Container**
  - Enable EISCAT user access
  - Notebook+DIRAC image
    - dirac.egi.ei cvmfs repository is configured alongside eiscat cvmfs repository (eiscat.egi.eu)
    - DIRAC client is available via cvmfs
    - pass user access token to the DIRAC File Catalogue to get user information
    - Access the DIRAC Storage Element to search and retrieve data from the EISCAT storage
  - Will include Matlab
    - Token to use EISCAT license?
  - Will include EISCAT Tools

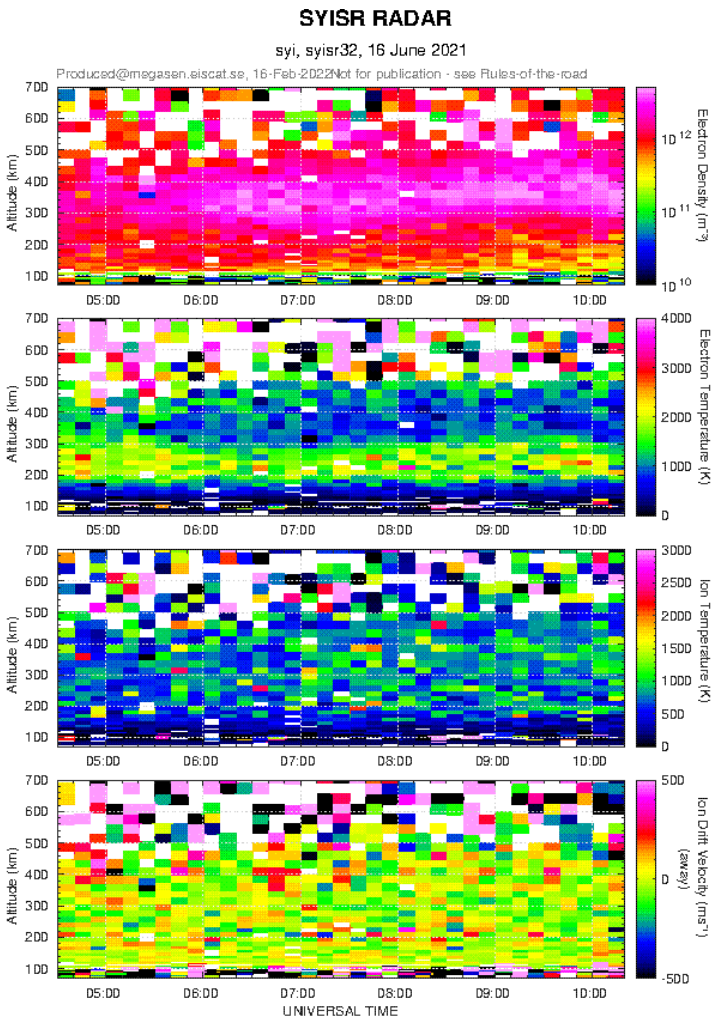
Analysis interoperability

- **Analysis (GUISDAP)**
  - Internally radar independent
  - EISCAT wrapper
    - Single beam setup
  - SYISR wrapper
    - Multi beam setup
      - Test also for EISCAT_3D

From the Use case group Report on the 2AHM meeting by Ingemar



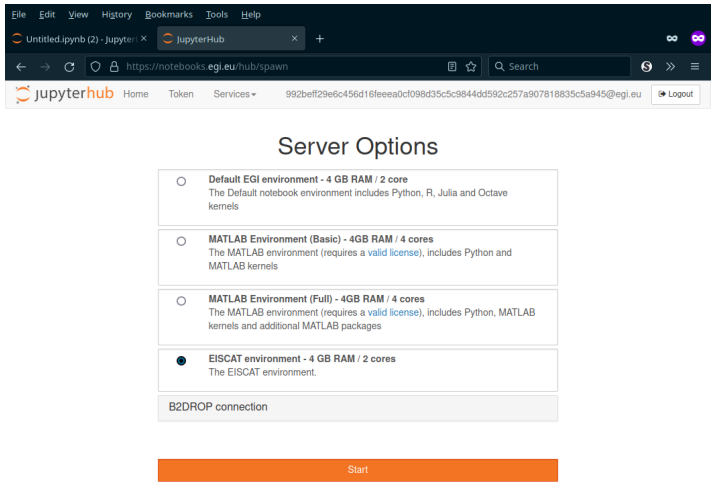
EISCAT analysis results



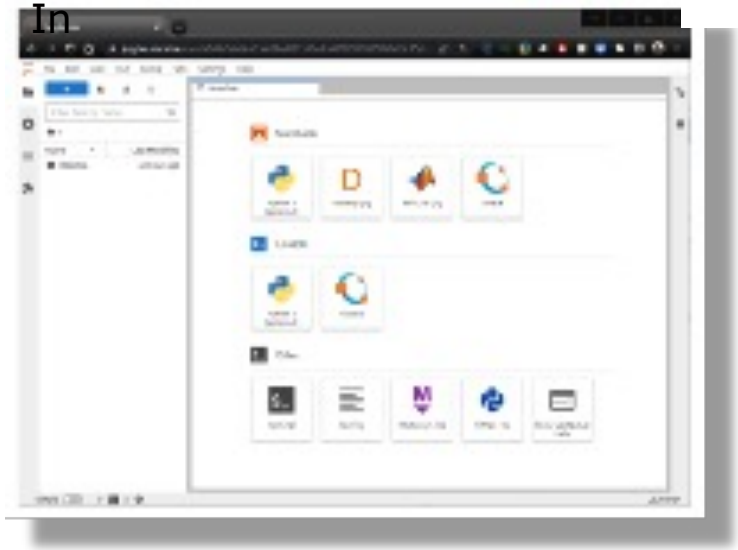
SYISR analysis results

Based on Jupyter notebooks

1. Login by EGI Check-In

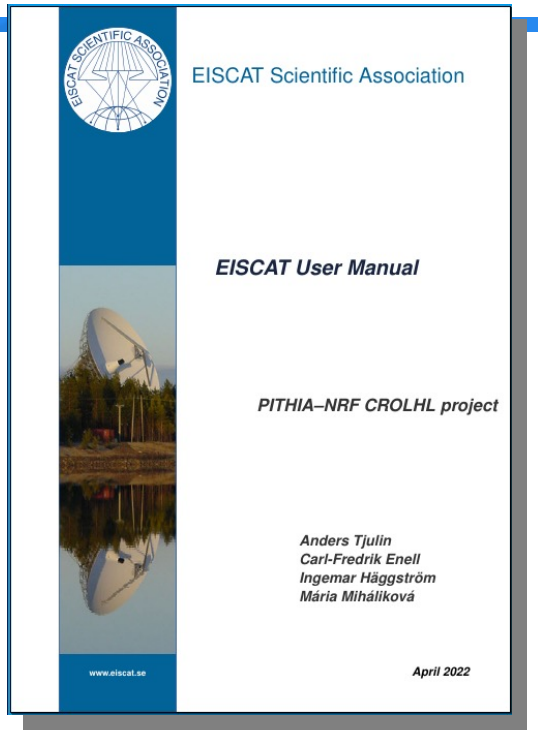


2. Select server



3. Jupyter Notebook landing page

•Jupyter Notebook Desktop workspace with GUISDAP application analysis running



User Manual

Notebook is in production:
https://notebooks.egi.eu/
https://jupyter.eiscat.se/
https://radardata.fed.cstcloud.cn/

From the Use case group Report on the 2AHM meeting by Ingemar

Call for International Partnerships

- Open mind, open framework, open collaboration for open data & open science.
- **Coming together is a beginning. Keeping together is progress. Working together is success.** *Henry Ford (1863 – 1947, American industrialist)*

Thank you!

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences