The Global Open Science Cloud: Vision and Initial Successes

GOSC INTERNATIONAL PROGRAMME OFFICE

Updated by August 25, 2023

Editors

Yin Chen, EGI Foundation
Lili Zhang, Computer Network Information Center, CAS
Jianhui Li, Computer Network Information Center, CAS
Simon Hodson, CODATA, the Committee on Data of the International Science Council Paul F. Uhlir, Consultant
Xueting Li, Computer Network Information Center, CAS
Hana Pergl, CODATA, the Committee on Data of the International Science Council

Authors

Incoherent Scatter Radar Data Fusion and Computation
Ingemar Haagstrom, EISCAT Science Association
Xinan Yue, Institute of Geology and Geophysics, CAS
Junyi Wang, Institute of Geology and Geophysics, CAS
Xiaoli Zhang, Computer Network Information Center, CAS

Open Reproducible Raw Diffraction Data for Access in Pandemics
John R Helliwell, University of Manchester
Genji Kurisu, Osaka University

Biodiversity and Ecology Information Platform for Camera Trap Data
Joe Miller, Global Biodiversity Information Facility
Yingchao Piao, Computer Network Information Center, CAS

SDG-13 Climate Change and Natural Disasters
Monthip Sriratana, National Research Council of Thailand
Gensuo Jia, Institute of Atmospheric Physics, CAS
Bapon Fakhruddin, Green Climate Fund
Zhang Ying, Computer Network Information Center, CAS

Sensitive Data in Population Health
Jildau Bouwman, TNO
Lei Liu, University of Fudan
Lauren Maxwell, Heidelberg University
Francisca Oladipo, Thomas Adewumi University

Technical and Infrastructure
Happy Sithole, Council for Scientific and Industrial Research, NICIS
Ze Luo, Computer Network Information Center, CAS
Haiming Zhang, Computer Network Information Center, CAS

Citation

Yin Chen, Lili Zhang, Jianhui Li, Simon Hodson, Paul F. Uhlir, Xueting Li, Hana Pergl, Ingemar Haagstrom, Xinan Yue, Junyi Wang, Xiaoli Zhang, John R Helliwell, Genji Kurisu, Joe Miller, Yingchao Piao, Monthip Sriratana, Gensuo Jia, Bapon Fakhruddin, Zhang Ying, … Haiming Zhang. 2023. The Global Open Science Cloud: Vision and Initial Successes. GOSC IPO, 25 Aug 2023. Zenodo. DOI:https://doi.org/10.5281/zenodo.8296517.

Contents

Executive Summary **2**

1. Rationale **3**

2. Vision and Mission **3**

3. Target Stakeholders **4**

4. The Challenges **5**

5. GOSC Governance **6**

6. Initial Successes **8**

Showcase of the GOSC funding model 8

Aligning with international projects – WorldFAIR and GOSC 8

Showcase of a Common Technical Framework 9

Showcase of implementation of the GOSC technical infrastructure 9

Showcase of GOSC for global science in PDBj and its adjunct data archive, XRDa 10

Showcases of GOSC for SDG-13 disaster mitigation 10

Showcase of GOSC for global radar science data sharing 10

7. Conclusions and Future Directions **11**

Executive Summary

The Global Open Science Cloud (GOSC[1]) vision is of an international collaborative and federated research environment that aims to facilitate scientific collaboration by providing a secure platform for researchers to access, store, share, and analyse data. The initiative started in 2019 and received seed funding from the Chinese Academy of Sciences. It has been designated as a key project for the CODATA[2] Decadal Program 'Making Data Work for Global Grand Challenges', and has gained support from EGI.eu[3], a leading European e-Infrastructure. The mission of GOSC is to link existing initiatives, creating a robust network of trusted research e-Infrastructures that connects research resources and stakeholders, enabling innovative science discovery in the evolving global open science environment.

The GOSC platform will be built on open standards, open source software, and open data, and is designed to be secure and compliant with data protection regulations. It aims to provide access to data from multiple sources and enable data interoperability, thereby facilitating data sharing, aggregation, and analysis across multiple research domains and geographies. The ultimate goal of GOSC is to promote international collaboration and access to scientific resources, contributing to solving global challenges and realising a world where science has no boundaries.

The potential beneficiaries of GOSC include researchers, institutions and organisations, funding agencies, policy makers and governments, industry and the private sector, citizen scientists and the public, and international collaborators. Researchers can use the platform to accelerate their research, access data from diverse sources, and promote open and reproducible research practices. Institutions and organisations can leverage the platform to promote a culture of open science, enhance research collaboration, and leverage shared resources. Funding agencies can use the platform to promote transparency, accountability, and impact of research supported by their funding. Policy makers and governments can facilitate knowledge exchange and evidence-based policy making. Industry and the private sector can access scientific data and tools for research and development, innovation, and commercialization purposes. Citizen scientists and the public can access scientific data and educational resources to promote scientific literacy and engagement. International collaborators can use the platform to foster global research collaborations and cultural exchange.

However, there are several challenges associated with the creation of a Global Open Science Cloud, including technological development, data interoperability, competition among major players, inclusivity and diversity, cultural and legal differences, data privacy and security, data governance policies, and funding for long-term sustainability. To address these issues, CODATA has established working groups and case study groups focused on specific topics and use cases related to GOSC. These groups work together to identify and address key challenges, develop best practices, and provide guidance and recommendations for GOSC implementation. The groups operate in an inclusive, open, and transparent manner, regularly reporting on their activities and outcomes to the wider GOSC community.

The Global Open Science Cloud has the potential to advance the way scientific data and resources are shared and accessed, and how global collaboration happens. However, addressing the challenges associated with its creation and ensuring inclusivity, interoperability, data privacy, and sustainability are crucial for its success. The collaborative efforts of stakeholders from different disciplines, regions, and sectors will be essential in realising the vision of a truly global and open science platform. The achievements of GOSC so far, including successful collaborations, funded projects, and the development of a common reference framework, demonstrate its potential and progress towards its goals.

[1] GOSC: https://codata.org/initiatives/decadal-programme2/global-open-science-cloud/

[2] CODATA: https://codata.org/

[3] EGI: eu: www.egi.eu

1. Rationale

Science is increasingly global and international cooperation is essential. The Square Kilometre Array (SKA) project is building the world's largest radio telescope, the scale of which represents a huge leap forward in both engineering and research toward future science and requires an international effort to accomplish. The COVID pandemic has highlighted the crucial need for international scientific collaboration to develop diagnostics, vaccines, and treatments in order to tackle health emergencies. This will leave its mark on research collaborations for years to come. According to the 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, there are 17 Sustainable Development Goals (SDGs), which is an urgent call on all countries to take joint action to deal with the impacts of climate change, the preservation of our oceans and forests, the ravages of infectious diseases, the improvement of health and education, and other imperatives. Science today is facing grand challenges that require global cooperation to resolve.

Around the world, "Open Science" is rapidly growing. As scientific research becomes highly data-driven and dependent on computing, there is an increasing need to share data, software, and infrastructure to reduce duplication and increase economies of scale. Digital infrastructures are developed at the institutional, national, and regional levels. For example, the US Open Science Grid (OSG) was established in 2004 and the construction of the European Open Science Cloud (EOSC) started in 2015. These initial developments were followed by the Chinese Science and Technology Cloud (CSTCloud) in 2017, the African Open Science Platform and the Australia Research Data Commons (ARDC) in 2018, and the Malaysia Open Science Platform in 2019. These Open Science platforms represent a novel chapter in the history of research and scientific infrastructures. They are changing the ways in which the world's scientists traditionally have done their research, leading to more innovative research methods and discoveries. Similarly, they have changed the ways in which professors teach and students learn. As these platforms increase and facilitate access to knowledge as well as to the storage and sharing of data, they also improve our capacity to respond more effectively to the numerous challenges confronting society.

The direction of this trajectory is clear. We have to do all we can to optimise the use of Open Science platforms globally and we should increase collaboration among the different platforms that are being established. UNESCO and the International Science Council (ISC)[4] have highlighted the need for interconnected Open Science platforms that serve global development and ensure adequate benefit sharing between high and low-and-middle income countries. This policy was best articulated on a global scale in UNESCO's Recommendation on Open Science, which was published in December 2021. [5]

It is in this context that the Global Open Science Cloud (GOSC) was initiated. The aim is to encourage dialogue and cooperation between Open Science platforms to facilitate alignment on governance, policies, technical, and interoperability concerns. Through the coordination of related investments, GOSC intends to support the global movement to bridge cross-domain and cross-geography divides and move towards a global data commons. The ultimate goal is to promote international collaboration and access to scientific resources, and to realise a world where science has no boundaries.

2. Vision and Mission

The idea of a "Global Open Science Cloud" initiative was first proposed during the September 2019 CODATA Beijing Conference. It generally refers to the co-design and co-development of a collaborative research environment for inclusive research around the world. Supporting the Open Science movement described by UNESCO, the GOSC Initiative has received seed funding from the Chinese Academy of Sciences for GOSC's international alignment. It has also been designated one of the key projects for the CODATA Decadal program (CODATA 2021[6]). Later in 2020, EGI.eu joined the discussion, which is the European leading e-Infrastructure that federates computer resources

[4] For ISC see https://council.science/actionplan/open-science/; see also the UNESCO Recommendation on Open Science, footnote below.
[5] UNESCO Recommendation on Open Science, https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en
[6] CODATA Decadal program: https://codata.org/initiatives/decadal-programme2/

and services. The first GOSC workshop was organised during the EGI conference 2020, which invited world-class science communities, e-Infrastructures, policy makers, and funding bodies to jointly discuss the current landscape and requirements for GOSC. The results were published in a GOSC landscape report[7]. After a series of meetings and workshops, the formal formation of working groups was agreed within CODATA. The successful launch event on 28th June 2021 welcomed participants from all over the world. The GOSC International Project Office (IPO) was established by the Chinese Academy of Sciences and CODATA in October 2022, which marked another significant milestone.

According to the vision of GOSC is "*the cooperation and alignment between Global Open Science Cloud activities in a robust network of trusted research e-Infrastructures, connecting research resources and all stakeholders to enable innovative science discovery in the dynamically evolving global open science environment*". GOSC is a new, open, and collaborative global research environment. It is a platform that enables researchers from all over the world to securely access, store, share, and analyse data. GOSC facilitates the sharing of data and knowledge, enabling researchers to collaborate more effectively, and to quickly and easily access data from anywhere in the world. In essence, it seeks to make international scientific collaboration easier in order to deliver excellent research resources that can be used to solve global challenges such as those caused by COVID19 and those set forth in the UN Sustainable Development Goals.

GOSC aims to create a platform that is built on open standards, open source software, and open data. It is designed to be secure and compliant with data protection regulations. GOSC provides a secure and trusted environment for researchers to store and share data, and to collaborate on research projects.

The GOSC platform should be able to provide access to data from multiple sources and enable data interoperability. This will enable data sharing, aggregation, and analysis across multiple platforms. Data should be organised in a way that is discoverable and reusable, while maintaining privacy and security. This will ensure that data is accessible to the right people at the right time.

The GOSC platform will provide access to a wide range of tools and services for data analysis, visualisation, and sharing. These tools should be easy to use and accessible to all users. The GOSC should also provide access to cloud computing services, such as storage and processing, to enable the analysis of large datasets and the development of new applications.

The GOSC will improve digital infrastructure connectivity. It will join existing regional platforms, such as EOSC[8], EGI[9], AOSP[10], CSTCloud[11], ARDC[12], and the Canadian Digital Research Alliance[13] etc., to form a worldwide highway that supports data transmission across continents and regions. The objective is to confederate digital resources, coordinate their provision, and co-deliver services so as to support international scientific research needs for global access, collaborative data analysis, and federated data processing.

There is thus a major opportunity for international research cooperation and alignment, with a wide range of stakeholders and organisations invited to work together. It will also contribute to developing countries and involve initiatives in Africa and Latin America like La Referencia. The GOSC will do that in partnership and collaboration with global organisations such as UNESCO, CODATA, RDA, WDS, and other interested parties.

3. Target Stakeholders

Stakeholders can benefit from developing and using the GOSC platform in various ways, including the research e-infrastructures that provide consolidated foundations in supporting the physical and virtual sharing of research resources and services for open science; service providers who make various research resources available to different communities following interoperable solutions; funding agencies who provide policy and fundings to manage sustainability, and who drive the level of openness and interconnectivity of the linked research facilities; researchers as well as the citizens who use open science and contribute to research resources generation and sharing; and end users who help explore fully the value of GOSC in various disciplines. More specifically, these stakeholders include:

[7] The GOSC Landscape paper: https://zenodo.org/record/5575275#.Y8VVXOzMLj0
[8] EOSC: https://eosc-portal.eu/
[9] EGI: eu: www.egi.eu
[10] AOSP: https://aosp.org.za/
[11] CSTCloud: https://www.cstcloud.net/
[12] ARDC: https://ardc.edu.au/
[13] Digital Research Alliance of Canada: https://alliancecan.ca/en

1. Researchers: Researchers from various disciplines, such as the natural sciences, social sciences, medical sciences, engineering, and humanities, can use the GOSC platform to share, access, and collaborate on scientific data, tools, and resources. This can help researchers to accelerate their research, access data from diverse sources, collaborate with peers from around the world, and promote open and reproducible research practices.

2. Institutions and organisations: Academic institutions, research organisations, and various scientific communities can use the GOSC platform to facilitate data sharing, collaboration, and innovation among their members. This can help institutions to promote a culture of open science, enhance research collaboration, and leverage shared resources for their scientific missions.

3. Funding agencies: Research funding agencies and organisations may be interested in using the GOSC platform to promote Open Science principles and facilitate access to research outputs supported by their funding. This can help funding institutions to promote transparency, accountability, and impact of research funded by public or private funds.

4. Policy makers and governments: Policy makers and government agencies can use the GOSC platform to promote open and collaborative science, facilitate knowledge exchange, and address global challenges, such as climate change, health crises, and sustainable development. This can help governments to promote evidence-based policy making, foster innovation, and support international scientific collaborations.

5. Industry and the private sector: Industry and private sector organisations can benefit from the GOSC platform by accessing scientific data and tools for research and development, innovation, and commercialization purposes. This can help industries to leverage scientific knowledge and resources to drive technological advances, develop new products or services, and create value-added solutions.

6. Citizen scientists and the public: Citizen scientists, science enthusiasts, and the general public may be interested in using the GOSC platform to access scientific data, educational resources, and engage in scientific activities. This can help promote scientific literacy, citizen engagement in research, and the democratisation of scientific knowledge.

7. International collaborators: Researchers and institutions from different countries may be interested in using the GOSC platform to collaborate on global research projects, share data and expertise, and foster international scientific collaborations. This can help promote cross-border research collaborations, cultural exchanges, and mutual learning among different scientific communities.

4. The Challenges

There are many hurdles to overcome. There are technological challenges in the development of platforms that will work equally well for natural and social sciences. There's a challenge of interoperability between data from different fields and disciplines and from different regions. There's competition among the major players that can work against the very principle of openness and access. And there's a specific need to encourage the creation of Open Science platforms in the Global South, and to position science systems and scientists in those countries at the cutting-edge of data-intensive science. Ultimately, the implementers of GOSC need to understand what is beneficial to the world and societies. Will all open science platforms be collaborative and interconnected? We certainly do not need platforms that generate new silos.

The requirements associated with creating a Global Open Science Cloud, include but are not limited to the following:

1. Data integration: GOSC needs to ensure that data from different sources can be seamlessly integrated and shared across the GOSC system. However, bringing together data from diverse sources in different formats and with different levels of quality can be a significant challenge. Lack of standardisation among the data and methods used by different scientific disciplines can make it difficult to share and utilise data effectively.

2. Data privacy and security: GOSC needs to ensure that sensitive data is protected and that access to it is controlled, while also allowing for data sharing and collaboration.

3. Data governance: GOSC needs to establish clear policies and procedures for data management, ownership, and access to ensure its integrity and utility.

4. Technical infrastructure and interoperability: GOSC aims to build and maintain a technical infrastructure that can handle the scale and complexity of the data, and ensure that different systems and platforms can communicate and share data seamlessly.

5. Funding and sustainability: GOSC will require significant funding to develop and maintain the infrastructure and ensure its long-term sustainability.

6. Inclusion and Diversity: GOSC needs to ensure that it is inclusive and accessible to researchers from diverse backgrounds and countries, especially those from low- and middle-income countries.

7. Cultural and Legal differences: GOSC needs to address the cultural and legal differences among different countries and regions to ensure that data sharing and collaboration are in compliance with all relevant laws and regulations.

5. GOSC Governance

The GOSC Initiative will be developed by Working Groups and Case Studies. All these groups are to promote collaboration and cooperation among stakeholders and to develop solutions to the challenges of creating a Global Open Science Cloud.

Working Groups are focused on specific topics or issues related to GOSC, such as data interoperability, technical infrastructure, governance, and policies. These groups are composed of experts and practitioners in the relevant fields, and they work together to identify and address key challenges, develop best practices, and provide guidance and recommendations for GOSC implementation.

Case Study groups, on the other hand, are focused on specific use cases or pilot projects that demonstrate the value and impact of GOSC. These groups work closely with researchers, data scientists, and other stakeholders to identify and test solutions to real-world problems and to gather evidence of the benefits of GOSC for different domains and applications.

Both the Working Groups and Case Study Groups are open to participation from interested parties. They are set up to be inclusive, open and transparent in their operations, and regularly report on their activities and outcomes to the wider GOSC community.

The four Working Groups are:

Governance and Sustainability

Long-term and sustainable visions and missions, guiding principles, governance models, funding mechanisms, and rules of participation are essential for GOSC to effectively connect with academics and worldwide research infrastructures. The main task of the Governance and Sustainability Working Group is to review and share knowledge on the above issues from various Open Science practices and initiatives around the world, and select the best match for the GOSC Initiative while coordinating the implementation process.

Policy and Legal

The Open Science practices are increasingly standardised through organisational, national, regional, and even transcontinental policies and principles. However, the fragmentation of these policy documents poses a fundamental challenge for the Open Science Cloud development across regions. To tackle this challenge, the Policy and Legal Working Group aims to conduct a comprehensive review of Open Science policies and practices in the context of operational platforms, focusing primarily on exploring possible agreement on policy and legal interoperability among Open Science platforms, which would then be implemented in selected GOSC case studies.

Technical Infrastructure

Research infrastructure is indispensable for the GOSC development. By building on and improving existing federation capabilities and interoperation frameworks, the Technical Infrastructure Working Group primarily focuses on achieving technical interoperability and connectivity between and among global e-infrastructures in order to support the need of cross-border research collaboration for researchers and the research industry. The discussion topics include: network connectivity and protocols, secure Authentication and Authorization Infrastructure (AAI), mechanisms for federation of computing, data management, and other services.

Data Interoperability

Human beings are facing fundamental challenges that call for an interdisciplinary approach to integrate facts from and across traditional domain boundaries. Following the guiding principles of Findable, Accessible, Interoperable, and Reusable (FAIR), the Data Interoperability Working Group aims to break down the silos that could hinder data exchange, fostering cooperation and possible alignment in the field of data interoperability, and encouraging global cooperation and alignment among Open Science clouds/platforms/commons in the area of data interoperability. The discussion topics include but are not limited to identifiers, semantic services, rigorous contextual and provenance metadata, analytical tools, and virtual research environments.

Five Case Studies are currently under investigation, and they are:

Incoherent Scatter Radar Data Fusion and Computation

EISCAT-3D[14] is a next-generation incoherent scatter

[14] EISCAT-3D: https://eiscat.se/eiscat3d-information/

radar system developed by the EISCAT association in Europe. It is a world-leading infrastructure, located in the Fenno-Scandinavian Arctic, that studies how the Earth's atmosphere is coupled with space at high latitudes. The Sanya Incoherent Scatter radar (SYISR) is also an incoherent scatter radar under design and construction by the Institute of Geology and Geophysics, Chinese Academy of Sciences (IGGCAS). The SYISR is located in the East Asia region, at a low latitude. The two radar systems complement each other and their collaboration on data sharing and analysis may greatly facilitate new research discoveries. However, there are no existing digital infrastructures currently supporting such cross-regional data sharing, so this Case Study involves infrastructure providers from both Europe (EGI FedCloud) and China (CSTCloud) together to investigate service solutions. This is an excellent scenario for GOSC, as it allows many tests to be conducted for cloud federation and service development.

Biodiversity and Ecology Information Platform for Camera Trap Data

Recent technological innovations have led the way for novel methods of detecting and monitoring species, often endangered animals. Camera traps use infrared movement detection to take an image of the animal and then AI is used to identify it taxonomically. In the past 10 years, trigger camera and camera trap technology have been widely used in nature reserves because it is non-invasive and has other advantages over traditional survey methods, especially for the monitoring and research of terrestrial large and medium-sized animals. According to statistics, camera trap technology has been used for wildlife monitoring and research in nature reserves, and images and videos of hundreds of rare and endangered species have been obtained.

Followed by a large number of camera deployments and massive amounts of image and video data, there are some challenges to share and use these data, including how to effectively store, organise, and manage massive amounts of camera trap data. The challenge continues in how to intelligently analyse these very large amounts of camera trap data from multiple camera trap sources into global research data infrastructures for easy reuse and citation, while maintaining needed data security and privacy requirements. This project, led by the Global Biodiversity Information Facility (GBIF), describes a biodiversity community project to mobilise camera trap data more easily into the open science realm. The project describes the expansion of the GBIF data model by CameraTrapDP to incorporate the details of the camera-trapped data following the FAIR principles. For instance, the model allows better interoperability among different camera trap installations. The outcomes of this Case Study are expected to be best practices for standards and specifications as well as alignment of international partners for sharing camera trap data.

SDG-13 Climate Change and Natural Disasters

In response to the UN Sustainable Development Goals (SDGs) and the UNESCO recommendation on Open Science, this Case Study primarily focuses on research regarding climate change and natural disaster, shouldering the key mission of co-building a better and more sustainable human community with stakeholders around the world. Supported by multiple sources of data, models, algorithms and tools, this Case Study intends to test policy and technical interoperability through regional demonstration on natural disasters, extreme weather events, and climate change. The Case Study seeks to study the large-scale patterns and heterogeneity of climate change and natural disasters through a dynamic monitoring and risk assessment model system, achieve high-precision comprehensive assessment and seasonal prediction on the temporal changes and spatial patterns of extreme climate events and natural disasters, and investigate the causes and frequency of extreme regional climate and related natural disasters, offering science-based support for responding to and mitigating disasters. Expected deliverables may include metadata and database federation, cooperative development of the online Computing and Processing Toolkit for SDG-13 indicators, and exploration of cloud federation techniques for SDG-13 on-demand data processing and analysis.

Currently several examples are being prepared and shared (e.g., in Thailand, Malaysia, Indonesia, and the Pacific region). The examples will showcase climate change impact assessments for sectors and be used for better understanding risk probability.

Sensitive Data in Population Health

Reusing Real World Observations (RWO) and health data for research, health innovation and policy are key to better health management, pandemic preparedness and imminent cost savings. However, the generally accepted notion that ‘citizens should be in control of the reuse of their personal data’ remains a paper mantra unless we design and implement a user friendly, trusted, and sustainable environment that allows the realisation of that ambition. Performing GDPR-compliant research will be entirely dependent on solving the trusted data federation challenge. This Case Study explores how innovative technology can change the current methods in care and improve the crucial influence of healthy citizens and patients in an advanced globally interoperable health data system. The resulting apporach will be relevant for future initiatives that need to reuse sensitive data of individuals. For example,

the currently limited or non-existent level of reuse of critical data on infection, viral spread, and the post-vaccination period as well as long term effects related to COVID-19 will severely hamper preparedness for future SARS-CoV-2 related problems, including the emergence of new variants. Reuse of sensitive RWO is potentially of much wider use than just for COVID-19. Hence, this Case Study aims to

contribute significantly to the generic abilities of the global society to tackle future health issues.

Open reproducible raw diffraction data for access in pandemics

The quest to find medical treatments based on 3D structural data derived from protein crystal structure analysis has a long history. The current COVID-19 Pandemic, following vaccine treatments, can be supplemented by a drug treatment, or the use of both can of course be envisaged. Searches for lead compounds to drugs require as precise as possible protein molecular models and protein with bound ligand crystal structures. Reproducibility of such data sets is paramount. Ideally, with the approach being followed here, there would be a single point of contact for definitive molecular models. This Case Study aims to manage raw diffraction data and specifically to link to the X-ray Diffraction raw data archive 'XRDa' at the Institute for Protein Research, Osaka University in Japan. XRDa is currently specifically tailored towards Asian depositors. The scope is protein crystal structure diffraction experiments and analyses. The proposed approach would seek to avoid multiple versions of a protein model derived from a single raw diffraction data set.

6. Initial Successes

GOSC delivers promising results. This section reports a number of success stories that have been achieved by GOSC Working Groups and Case Studies.

Showcase of the GOSC funding model

The funding model for GOSC development can be challenging. This is largely due to the fact that regional or national funding agencies often prioritise local projects and have limited resources for global partners.

To address this, EGI.eu (a coordinating body of EGI Federation which is an European e-Infrastructure for Cloud federation) included GOSC as a project objective in a Horizon 2020 project proposal, EGI-ACE (2020-2023) and outlined collaboration activities with CNIC (the Computer Network Information Center, a Chinese Academy of Sciences research institution that operates the CSTCloud). CNIC contributes 30 Person Month (PM) to EGI-ACE without claiming any contributions from the European Commission. Instead, CNIC submitted a proposal for the GOSC Initiative to China's funding agencies and obtained matched funding.

In order to ensure continued development, EGI.eu and CNIC have signed a Memorandum of Understanding (MoU) to define long-term collaborations, including cloud federation, joint service development, and dissemination. CNIC has received positive feedback from the Chinese Academy of Sciences for an additional 5 years of funding support. Supporting global collaboration is a long-term mission of EGI, and will be included in future projects such as the EGI Flagship and EOSC, which is part of the Horizon Europe Working Program.

Aligning with international projects – WorldFAIR and GOSC

Building on recent CODATA activities, 'WorldFAIR: Global cooperation on FAIR data policy and practice',[15] was successfully funded in June 2022 by the European Commission through its Horizon Europe Framework Programme.

Led by CODATA, with the Research Data Alliance association as a major partner, the WorldFAIR project works with a set of case studies to advance implementation of the FAIR data principles, in particular those for data interoperability, and to develop a set of recommendations and a framework for FAIR assessment in a set of disciplines, or cross-disciplinary research areas. The project explores features of an emerging Cross-Domain Interoperability Framework (CDIF) with 11 case studies from the physical, social, agricultural and environmental sciences and the cultural heritage sector, and prepares FAIR Implementation Profiles, appropriately adapted to each (cross-)discipline area. This will lead to, and help inform, a fuller mapping of current best practices and emerging solutions and initiatives for FAIR data in these domains.

WorldFAIR runs for 24 months from 1 June 2022. The project is a collaboration between 19 partners around the world including prominent research institutions and scholarly organisations from Africa, Australasia, Europe, and North and South America. This work is an essential part of CODATA's Decadal Programme 'Making Data Work for Global Grand Challenges', which is sponsored by ISC as part of its Action Plan.

WorldFAIR aims to make a major contribution to thinking about FAIR implementation and interoperability in the context of EOSC and worldwide. GOSC and WorldFAIR share aspects of their methodology, in particular the emphasis on Case Studies, which will be mutually enriching. GOSC is exploring the use of FAIR Implementation Profiles and will review, contribute to and aim to implement the CDIF. Such international alignment is essential for GOSC and the partner Open Science Clouds and platforms.

[15] WorldFAIR project: https://worldfair-project.eu/

Showcase of a Common Technical Framework

A common technical framework for GOSC has been proposed by the Technical Infrastructure Working Group, which aims to capture the best practices of existing Open Science Platforms to inform future design and implementation. It provides a structure for organising and classifying different types of research resources and services, and can serve as a common language and reference point for different Open Science platforms.

The GOSC common technical framework consists of five different layers: 1) the network layer, which provides a consolidated foundation for internet interoperability between different GOSC stakeholders; 2) the computing layer, which focuses on computing interoperability to enhance open science cloud infrastructures and services; 3) the data layer, which offers optimal solutions to increase the ability to generate value from effective big data analytics for valuable scientific insights; 4) the software and technical layer, which plans to design, build, and operate a multi-site cloud-based facility and resources to support research across applications, services, and systems; and 5) the community layer, which connects the technology with the community and explores sustained business models and mechanisms for multilateral cooperation. The governing rules and alliance for secure, trustworthy, and sustainable resource accessibility will also be essential to drive the running of a GOSC research ecosystem.

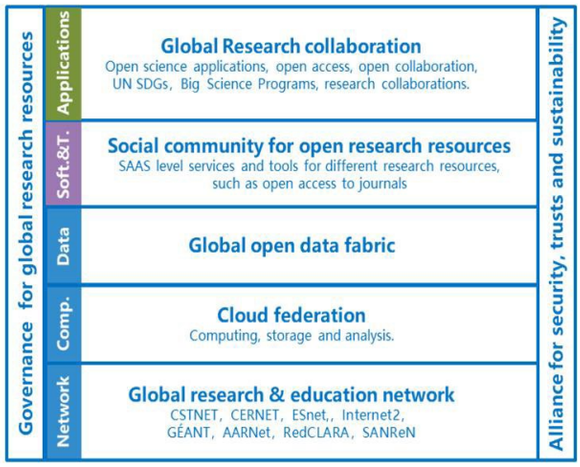

Fig. 1. The Global Open Science Cloud Common Technical Framework [16]

We foresee that AI-powered technologies can be used in the implementation of the GOSC common technical framework. For example, AI has the potential to revolutionise various aspects of scientific research, including data analysis, knowledge discovery, and decision-making, and can therefore play a significant role in advancing the goals of the GOSC platform. AI-powered technology can also bring significant benefits to the GOSC federation by improving resource management, service placement, fault detection and recovery, security and compliance, Service Level Agreement (SLA) management, and cost optimization. These capabilities can help the GOSC federation achieve better performance, increased reliability, enhanced security, and improved cost-efficiency, ultimately leading to a better user experience and increased customer satisfaction.

Showcase of implementation of the GOSC technical infrastructure

Integration of China's CSTCloud with EGI Federation was achieved in 2021, and CNIC became the first EGI Cloud

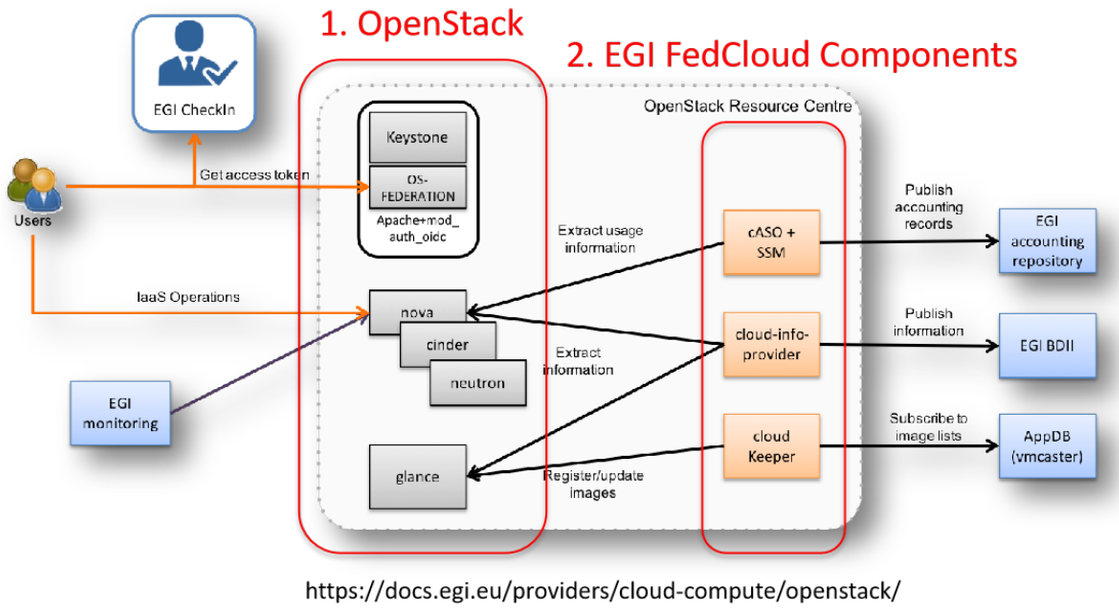

Fig. 2. The architecture of integration of CSTCloud OpenStack cluster with EGI Federation

[16] Figure from https://www.cstcloud.net/gosc/

provider outside Europe. EGI Federation is an European infrastructure for data-driven, computing intensive, and exabyte-scale processing. During the past 15 years, EGI has been delivering open solutions to advanced computing and data analytics, supporting hundreds of international user communities reaching a scientific impact of about 2,000 open access publications per year, and about 71,000 registered users worldwide.

Operated by the Computer Network and Information Center (CNIC) of the Chinese Academy of Sciences (CAS), the CSTCloud is a national infrastructure for CAS scientific communities and China's top research. CSTCloud provides computing facilities for Chinese advanced research projects including CASEarth, CAS space science missions, and research related to big facilities or observation stations such as the Five-hundred-meter Aperture Spherical Telescope (FAST) and the Large-High-Altitude Air Shower Observatory (LHAASO).

This achievement in the integration of the projects contributes to the GOSC by realising the cross-regional cloud connection. It provides opportunities for testing resource federation and service delivery.

The integration work has been encountering many challenges -- different technical environments, different development cultures, limited documentation, no previous examples, and the like. It has taken a number of months of effort by people from several organisations and teams. The EGI-CSCloud federation takes a consolidated step forward in the exploration of infrastructure-level collaboration and provides useful experiences.

Showcase of GOSC for global science in PDBj and its adjunct data archive, XRDa

Using a process similar to peer review, IUCr and PDBj have established a collaboration within the CODATA Case Study to make the medical-protein crystal structure definitive versions of data files that are FAIR (Findable, Accessible, Interoperable, Reusable) and of high quality. PDBj launched XRDa, https://xrda.pdbj.org/, a data archive for raw diffraction images, to evaluate raw data, processed structure factors, and derived protein molecular models. This provides a platform for scientific reproducibility and definitive reusability and can be beneficial to the entire macromolecular crystallography field beyond medical pandemics.

The method used to make a selected protein structure deposition 'definitive' involves a comprehensive review of its data using several metrics, especially the diffraction resolution limit and scrutiny of unmodelled electron density. The PDBj Director can be advised by the reviewer of any necessary remediation of the deposition files and who can contact a depositor to seek revisions. This procedure is exactly analogous to a journal editor, with their chosen referees, in the peer review stage.

This Case Study is including covid-19 and other medically significant proteins[17,18,19.]

Showcases of GOSC for SDG-13 disaster mitigation

Supported by the GOSC testbed, an Open-Science platform for SDG-13 research in Southeast Asia is in progress, providing remote sensing monitoring of the history and near real-time situation of disasters such as fire, drought and flood.

To facilitate the use of services and resources, this SDG-13 platform has been integrated with the unified user authentication system, CSTCloudAAI (https://aai.cstcloud.net/), to facilitate data accessibility. Pre-stored and multiple-sourced datasets are ready for use. A SDG13 workbench is also developed, providing container cloud-based software and tools for online data analysis and Spark clusters also ready to facilitate the workflows. Visualization display of meteorological forecast data and information are also available to support scientific research and assist decision-making in the collaborative virtual environment.

Currently, based on converged data resources, tools and services, this SDG-13 data platforms seek to explore open-science ways of resources collecting and inclusive services delivery for disaster prediction and early warning based on Southeast Asian collaboration. Beyond this regional use case, further work will highlight worldwide engagement, particular collaboration with developing nations and regions, to better support research on climate change and assist decision making across domains and regions.

Showcase of GOSC for global radar science data sharing

The EISCAT–SYISR data access portal, developed by the Incoherent Scatter Radar Data Fusion and Computation Case Study Group, is a cutting-edge platform built on top of the cross-region cloud federation achieved by EGI and CSTCloud. This innovative system is configured with a dedicated community resource space within the federation, known as the Virtual Organisation (VO), which serves as a hub for seamless collaboration among radar science researchers. Cloud resources from both EGI, offering 10 Core CPU and 10TB storage, and CSTCloud, providing 30 Core CPU and 100GB storage, are allocated to this VO, ensuring ample computing power and storage capacity for the research community.

To enable user access, data management, and data analysis, a suite of e-Infrastructure services has been configured within the VO. This includes EGI Check-In for

[17] Brink, A. and Helliwell, J.R. (2022) IUCrJ 9, 180-193.
[18] Hanau, S. and Helliwell, J.R. (2022) Acta Cryst. (2022). F78, 96–112.
[19] Helliwell, J R (2021) Acta Cryst F77, 388-398.

authentication and authorization, DIRAC-based Workload Manager for efficient job submission and data management, and Notebook services for data analysis. These services work in harmony to create a seamless and user-friendly environment for researchers, enabling them to effortlessly access, manage, and analyse data across the connected cloud resources.

This remarkable development showcases the immense potential of a cross-region cloud federation in supporting international research collaborations. The connected cloud resources are now accessible to users from both Europe and China, which was not possible before. This has paved the way for greater collaboration and progress in research and development on a global scale.

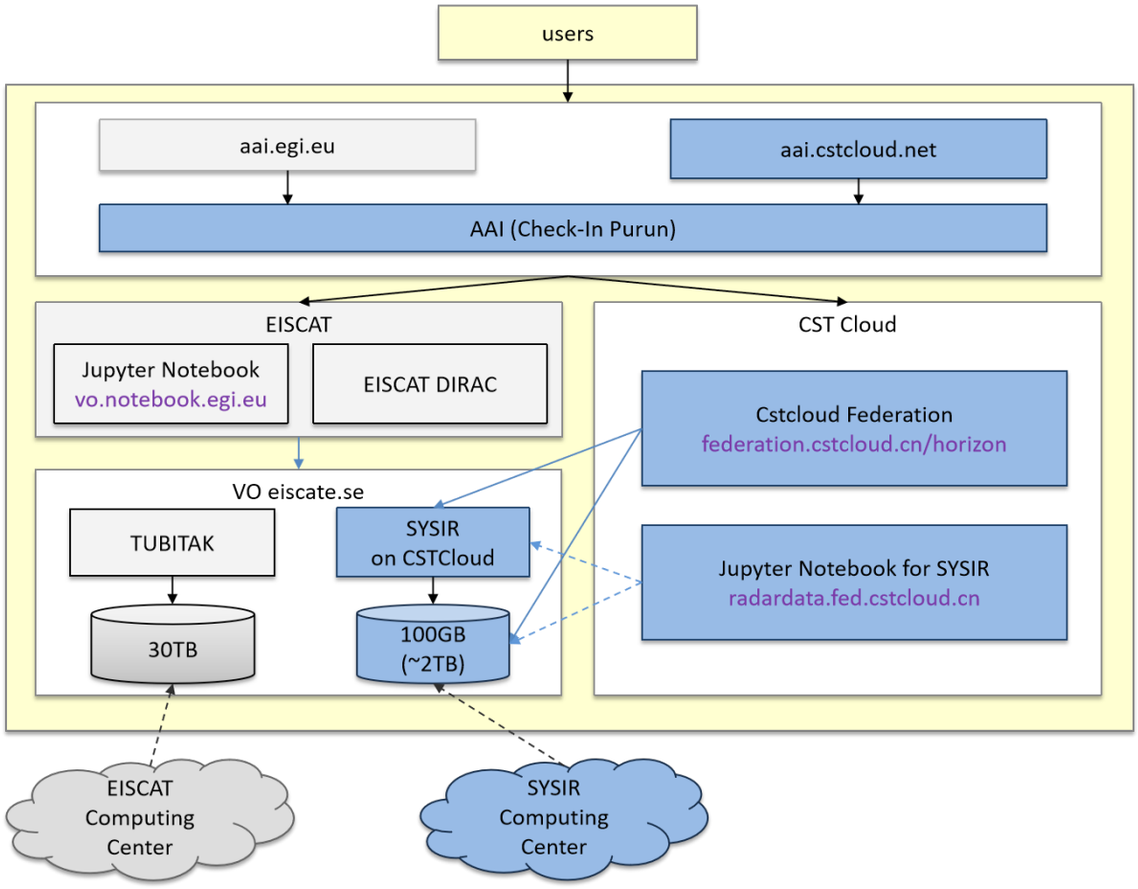

Fig. 3. The architecture of EISCAT–SYISR data access portal

7. Conclusions and Future Directions

The Global Open Science Cloud is an ambitious project with the goal of creating a global network of data and computing infrastructure resources that scientists can use to share and analyse data. The success of GOSC will depend on a number of factors, including funding, support from governments and research organisations, and the ability to effectively collaborate with other organisations and initiatives. However, the concept of GOSC aims to provide an unprecedented level of access to data, tools, and computational resources to researchers around the world, which has the potential to greatly advance scientific discovery and collaboration.

Future development of the GOSC may focus on implementing several key strategies, including:

1. Building a robust infrastructure: GOSC should invest in building a robust, scalable, and secure infrastructure that can handle large amounts of data and provide reliable access to researchers and other users.

2. Developing useful and user-friendly tools: GOSC should develop useful and user-friendly tools and applications that make it easy for researchers and other users to access, analyse, and share data.

3. Exploring GOSC testbeds: based on the robust infrastructure and tools, testbeds will be developed jointly by the GOSC community to help implement disciplinary showcases. These demonstrations will in turn help testify the usability and robustness of interconnected research facilities of different levels, such as the federation in the network layer, the computing later, the data layer, the software and technics layer, and the community layer (see figure 1). Such testbeds provide portraits for the GOSC full picture and help validate the efficiency and effectiveness of transparency, interconnectivity together with trustworthiness of the open science clouds.

4. Establishing partnerships: GOSC should establish partnerships with other organisations and institutions, such as research institutions, universities, and private companies, to expand its reach and impact.

5. Promoting open science and data sharing: GOSC should actively promote open science and data sharing principles and practices to encourage researchers and other users to take advantage of the platform's resources and services.

6. Providing training and support: GOSC should provide training and support to researchers and other users to help them make the most of the platform's resources and services.

7. Encouraging community engagement: GOSC should encourage community engagement and participation by providing opportunities for users to contribute to the platform and share their own research and data.

8. Continual improvement: GOSC should be focused on continually improving the platform, by soliciting feedback from users, and incorporating new features, functionalities and technologies.

9. Compliance and security: GOSC should ensure compliance with relevant laws and regulations and implement robust security measures to protect user's data and keep it confidential.

10. Innovative funding model: GOSC should explore innovative funding models, such as public-private partnership and subscription-based services, to ensure sustainability of the platform in the long run.

The development of GOSC requires significant funding. And we can utilise several strategies to build up resilient business model, such as to work with government agencies, national funding bodies, non-for-profit foundations and international funding orgnaizations to secure funding for infrastructure, research and development, and other initiatives. Or to partner with commercial enterprises, such as technology firms and research organisations, to jointly invest and operate the platform. GOSC can also seek sponsorship from organisations that align with its mission and values, such as scientific associations, research institutions and philanthropic organisations. It may also leverage its reach and influence to generate advertising and branding opportunities for organisations that align with its mission; organise and participate in innovation competitions to attract support from investors and other stakeholders.

Acknowledgment

This work is supported by National Key R&D Program of China (2021YFE0111500), National Natural Science Foundation of China (72104229), CAS Program for Fostering International Mega-science (241711KYSB20200023) and European Commission (EGI-ACE, 101017567). We wish to thank all the members of the CODATA GOSC Steering Group, Working Groups and Case Studies for providing valuable ideas in developing the GOSC Initiative.

Contact:
GOSC IPO: gosc_ipo@cstnet.cn
CODATA Secretariat: info@codata.org