Results-Oriented GEOSS: A framework for transforming Earth observation data to knowledge for decision making

This document has been submitted to the 48th Session of the GEO Executive Committee for decision

(Draft version – 6 June 2019)

1 INTRODUCTION

By decision of the 44th Session of the GEO Executive Committee, the GEO Secretariat (GEOSEC) established the Expert Advisory Group (EAG) composed of invited experts to develop a strategy document to advance a "Results-Oriented GEOSS". The main task of the EAG has been to provide expert advice on how best to design a GEOSS capable of addressing the EO needs of monitoring frameworks for global policy initiatives and conventions. Over the past nine months, the EAG held a number of strategy meetings, consulted with a variety of representatives from member states, participating organizations, the GEO programme board, Flagships, Initiatives, GEOSS EVOLVE, GEOSS Platform team members and others as appropriate to develop a proposed strategy.

The EAG, in consultation with Programme Board, present this update that includes a strategy for advancing a Results-Oriented GEOSS. In this strategy document, the EAG proposes an instrumental framework to advance a Results-Oriented GEOSS that includes foundational pillars, goals, objectives, and actions to transform the current data focused GEOSS to a knowledge-based GEOSS delivering decision-ready products and services. Recognizing the need for adjustments, EAG recommends this instrumental framework be considered an adaptive strategy that will incrementally improve overtime guided by tangible results, community feedback, lessons learned, improved processes and methods. To facilitate implementation of the framework, the EAG supports a limited proof of concept that includes development of a knowledge hub to facilitate the delivery of knowledge-based products and services.

2 MOTIVATION

As stated in the Group on Earth Observations (GEO) Strategic Plan 2016-2025, a central part of GEO's Mission is to build the Global Earth Observation System of Systems (GEOSS). GEOSS facilitates the sharing of environmental data and information collected from the large array of observing systems contributed by countries and organizations within GEO. GEOSS can make a substantial contribution to the societal understanding of Earth systems processes, interactions, responses, and causes to enhance decision-making. GEOSS can provide decision makers with *sound scientific data, knowledge, and tools to make informed decisions* for the health, welfare, and safety of communities under conditions of ongoing environmental change. GEOSS increases our understanding of Earth processes and enhances predictive capabilities that underpin sound decision-making: it provides access to data, information, and knowledge to a wide variety of users.

The Strategic Plan emphasizes the importance of involving all key stakeholders across the provider-user spectrum in a process of co-design to ensure that the GEOSS data is transformed into useable knowledge and information to address societal needs.[1] **This original vision of GEOSS as a "system of systems" continues to be essential to GEO.**

The Strategic Plan emphasizes the scope of GEO includes an "end-to-end process of identifying needs, ensuring the availability of data with which to develop information to address societal challenges, and transforming that information into knowledge through the generation of products and services for end-users."[2] The GEOSS Platform provides the foundational infrastructure and capabilities necessary to link and make Earth Observation(EO) resources available for discovery and access by the GEO community. However, key challenges for GEO remain *knowledge and technical diffusion*––the ability to transforming data into *fit-for-purpose* and *fit-for-use* knowledge-based products and services that support decision-making.

3 HOW CAN GEO ACHIEVE TANGIBLE RESULTS FROM GEOSS?

To make Earth observation resources decision-ready, complex data, models and other Earth observation information products from a variety of sources, in different formats, spatial/temporal scales and resolutions will require quality assurance, integration and aggregation for analysis, synthesis, and modelling to advance the development of knowledge-based tools and products for understanding and decision-making. Access to the underlying data, models, algorithms can accelerate co-design/production. GEO must adopt a results-oriented organization framework that supports *knowledge and technical diffusion* to commoditize EO data for evidence-based decision-making.

Given these challenges, this document from the EAG outlines a strategy to create a policy-driven, country-relevant open-science environment that transforms data into fit-for-purpose knowledge-based products and services, so that GEO Member states and Participating Organizations can better connect people, processes, and technologies to establish a next-generation GEOSS Platform that supports knowledge and technical diffusion to increase the value and use of Earth observations worldwide.

4 AN ILLUSTRATIVE EXAMPLE

GEO can assist Member States, particularly developing nations to increase the benefits they derive from Earth observation and provide support for decision-making. We consider the case where an expert in Cameroon is working to address a policy mandate for her government to assess tropical forest loss in her country, to support a system that complies with the best practices for Reducing Emissions from Deforestation and Forest Degradation (REDD+), and reports emissions to the United Nations Framework Convention on Climate Change (UNFCCC).

[1] GEO Strategic Plan 2016-2025: Implementing GEOSS: GEOSS, the Global Earth Observation System of Systems https://www.earthobservations.org/documents/GEO_Strategic_Plan_2016_2025_Implementing_GEOSS.pdf

She then enters the GEOSS knowledge hub. After being recognized by the system, she has access to a large network of resources provided by the GEO community. In her case, the GEO Secretariat has facilitated access to Copernicus service that has put together Sentinel and Landsat data for Africa. She has also access to MODIS data in the African Data Cube, which also hosts a data set of ground-truth data for the African tropical belt. Due to some restrictions by the in-situ providers, the ground-truth data set is not yet public, but GEOSS and the African Data Cube have secured its use by GEOSS-authenticated users.

She has access to the GEOSS knowledge hub and uses a detailed report on best practices for forest monitoring using big EO data done by Brazil's National Institute for Space Research (INPE). INPE has also built a state-of-the-art algorithm that uses deep learning methods for forest monitoring. This algorithm is available on a GEOSS repository of trusted and curated methods. She then uses GEOSS programming environment to combine the ground truth data with the MODIS data set, using INPE's algorithm. The result is a 20-year time series of forest change for Cameroon for the period 2000-2010. Next, she refines the result for the period 2012 onwards using the combined Sentinel-Landsat mosaic. After testing the results, she then uses the tropical biomass map provided by the European Commission (EC) Joint Research Centre to produce an emissions report that the government of Cameroon will submit to the UNFCCC.

5 OPERATIONALIZING THE CONCEPT

The scenario described above is not futuristic. Rather, it could shortly become reality in different ways. GEO and its collaborators should consider how knowledge about the Earth is created and disseminated. This issue is best framed from the perspective of reproducibility.

All successful Earth observation applications have different components: policy mandates, national priorities, satellite and in situ observations, collaborators, methods and algorithms, and deployment strategies. We need to be able to organize GEOSS to make sure that these components are visible and accessible and that others can diffuse relevant work for country relevant evidence-based decision-making. Doing this, we will capture a significant portion of the knowledge embedded in each EO application, lower barriers for developing countries to use EO and deliver tangible results. In so doing, we will transform GEOSS and the GEOSS Infrastructure from a discovery and access facility to an evidence-based decision platform––user interfaces with intelligent capabilities, analysis-ready data, co-creation capabilities and decision-ready services to produce evidence-based products and services that inform policy and decision-making.

The EAG recommends a whole systems end-to-end approach to address current gaps, to link the people, process and tools required to achieve a results-oriented GEOSS.

6 PILLARS TO ACHIEVE A RESULTS-ORIENTED GEOSS

The proposed results-oriented framework is based on four pillars that provide a consistent approach for GEO to identify, select, evaluate and implement actions towards a results-oriented GEOSS. These four pillars will underpin the GEO Work Programme with activities set in ***policy- and country-relevant*** objectives, using ***project-based*** methodology to deliver ***knowledge-based*** tangible results for decision-support.

A. Policy-Relevant:

A results-oriented GEOSS that supports decision-making must be policy-relevant along defined national/global policy objectives and societal benefit areas. These applications need to be carried out by the Member States, with support from the GEO community. The results-oriented approach will be particularly relevant to GEO Flagships and Initiatives that have policy-relevant mandates. Much progress on the Flagships and Initiatives will result from finding a trusted and consistent set of methods that can be applied to the big data sets that are openly available to inform policy-relevant decisions.

GOAL:	**Key Measures/Outputs:**
Engage, Advocate and Deliver Policy-Relevant knowledge-based services	a. Identify specific policy indicators where the application of EO has the potential to increase understanding leading to decision-making by member states. b. Delivery of policy-relevant and actionable products and services c. Policy related interventions and actions that result from member state application of EO

B. Country-Relevant

In a results-oriented approach, the involvement of the concerned Member state is necessary. In keeping with the principle of co-design, GEO needs to engage with member country institutions that provide in situ observations to ensure that these institutions benefit from full and unrestricted access to the results to which they contribute. Instead of having a large number of small-sized case studies, Flagships and Initiatives need targeted actions that support the development of country-relevant results. For example, instead of answering the question *"what data does agriculture monitoring need?"*, GEO needs to respond to the demand: *"what are the components of a successful agricultural monitoring application to support food security for GEO Member States and how can GEOSS organize knowledge-based services required to build it?"*.

GOAL:	**Key Measures/Outputs:**
Deliver fit-for-purpose country-relevant knowledge-based products and services	a. Member states advocate national open data policies b. Member states implement data sharing and management principles c. Member states advocate and engage in capacity development and communities of practice. d. GEO community deliver policy/country-relevant fit-for-purpose products and services that support member state understanding and decision-making

C. Project-Based

A results-oriented organization is better managed by having a portion of its work to be *project-driven*. In this context, we define projects as fixed-term activities that are approved in the GEO work programme, managed, properly financed and staffed by sponsoring GEO members and participating organizations so as to achieve policy and country-relevant results. Projects with fixed terms and clear objectives will provide a foundation to keep GEO focused on results and to ensure that those who finance them get what they expect. A main focus of the GEO projects will be to bring the benefits of EO to member states in particular developing nations.

GOAL: Improve management and accountability in the GEO Work Programme towards a result-oriented GEOSS	**Key Measures:** a. Approved in the GEO Work Programme that will be fundamental in both providing methodologies for solving problems and identifying potential end users. b. Properly managed, staffed, resourced and financed c. Clearly defined/aligned with policy/country-relevant indicators d. Defined schedule with clearly defined products and services e. Monitoring and reporting GEO governing bodies and the community

D. Knowledge-based

GEO has the mission of improving the capacity of all its Member States to use Earth observation data for decision-making. To achieve this goal, we need to go beyond data. Broadening the global access to knowledge-based products and services that support decision-making is essential for delivering tangible results. Achieving this goal requires combining best practices from the GEO community with long-term capacity development.

The GEO community produces many useful and relevant results, which they aim to share globally. Such resources will be of different types to reflect the entire information flow of the research and knowledge pertaining to a domain. Linked documents that contain relevant information for Earth observation applications that provide access to the underlying data, models, methods and algorithms are necessary to promote knowledge and technical diffusion for reproducibility, scalability, and co-design/co-production.

Engaging developing nations as providers and users of in situ data requires a change in attitude from the data analysis and modelling community. All models and analysis methods that use such data should be made public. In case of complex models such as numerical weather predictions or climate change models, at a minimum all results from these models that can benefit developing nations need to be public.

GOAL: Deliver fit-for-purpose	**Key Measures:** a. Member States adopt and implement GEO Data Sharing and GEO Data Management Principles

country-relevant knowledge-based products and services	b. Establishment of communities of practice for knowledge and technical diffusion c. Curated and linked data that contain relevant information to promote reproducibility, scalability, and co-design/co-production d. Advance the creation and co-development of analysis ready and decision-ready data and knowledge e. Establishment of knowledge-based products and services that lower the barriers for developing countries to use EO

7 ACTION PLAN

The EAG assessed the current state of capability, considered the proposed architecture from GEOSS Evolve and recommends the establishment of a GEO Knowledge Hub (GKH) to close current gaps in the GEOSS Platform to facilitate the implementation of the framework as proposed.

The primary goal of the GEO Knowledge Hub is to advance transformation of data into knowledge-based services for evidence-based decision-making. The GKH is envisioned to provide linkages to policy mandates, the work programme with related project and task workflow alignment, task monitoring, and implement knowledge diffusion and technical diffusion capabilities towards a result oriented GEOSS. Implemented, the GKH will effectively extend the GEOSS Platform to achieve the optimal state described in the Strategic Plan. The GKH will serve as a component of the GEOSS Platform, and its development is part of the GEOSS Infrastructure Development Foundational Task, which will be part of the GEO Work Programme 2020-2025. This proposal for this Foundational Task has been endorsed by the GEO Programme Board, GEOSS Evolve, and the GEOSS Platform.

The EAG recommends that the GEOSS Implementation Coordination Task Force carries out a proof of concept demonstrating the GEO Knowledge Hub using the GEOGLAM Flagship for the proof-of-concept. This demonstration will provide valuable lessons learned, best practices, recommendations on architecture, scaling and long-term implementation and operational requirements for the Knowledge Hub. The following governance structure and activities are proposed:

a) The GEOSS Implementation Coordination Task Force will oversee the development of the proof-of-concept to ensure that the Knowledge Hub will be consistent with the other components of the GEOSS Platform.
b) This proof of concept will be built in consultation with the GEO Programme Board, GEO Work Programme Leads, and external consultation as appropriate.
c) GKH will be developed using only free and open source software, drawing as much as possible on existing community solutions for Open Science and on accepted best practices for knowledge sharing.
d) Progress developing the proof-of-concept will be demonstrated at the 48th and 49th GEO Executive Committee meetings.

e) The proof of concept will be demonstrated at the GEO Plenary 2019, in Canberra, Australia. Based on the response of the GEO community to this proof of concept, the GEOSS Implementation Coordination Task Force will propose a plan for further development of the GEO Knowledge Hub, to be presented for decision to the GEO Executive Committee, at their March 2020 meeting.

8 CONCLUSION

To realize the GEO 2025 vision and maximize the use of Earth Observations, GEO must address knowledge and technical diffusion––the ability to transform data into fit-for-purpose and fit-for-use knowledge-based products and services that support decision-making. Implemented, this framework has the potential to support GEO member states' ability to advocate, engage and accelerate the delivery of tangible knowledge-based products and services for decision making. Recognizing the need for adjustments, EAG recommends this instrumental framework be considered an adaptive strategy that will incrementally improve overtime guided by tangible results, community feedback, lessons learned, improved processes and methods. To support the framework the EAG supports a limited proof of concept that includes development of a knowledge hub to facilitate the delivery of knowledge-based products and services.

The Annex includes a draft concept of operations, requirements and additional information related to the GKH. Additional refinements will be completed as part of the concept of operations. The final version that incorporates best practices, lessons learned and recommendations will be included at the conclusion of the proof of concept.

Annex A
Proposed Design of the GEO Knowledge Hub

This document is an annex to the final report of the Expert Advisory Group to the GEO Executive Committee. While the main report focuses on the broad issues related to the vision of a results-oriented GEOSS, this annex describes the proposed design of the GEO Knowledge Hub, a new component of the GEOSS Platform Infrastructure, whose need has been identified by the Expert Advisory Group.

1 INTRODUCTION

The GEO Knowledge Hub (GKH) is a digital repository providing access to knowledge required to build applications of Earth observations. The purpose of the GKH is to reveal all components of the application, including: (a) research papers and reports describing methods and results; (b) software algorithms and cloud computing resources used for processing; (c) in situ and satellite imagery data used; and (d) results for verification.

The Knowledge Hub is one component of the GEOSS Platform, and its development is part of the GEOSS Infrastructure Development Foundational Task, which will be part of the GEO Work Programme 2020-2025. This proposal for this Foundational Task has been endorsed by the GEO Programme Board. The GEOSS Implementation Coordination Task Force will oversee its development to ensure that the Knowledge Hub will be consistent with the other components of the GEOSS Platform.

1.1 Why is the GEO Knowledge Hub needed?

GEO has the mission of improving the capacity of all its Member States to use of Earth observation data for decision-making. To achieve this goal, we need to broaden the global access to knowledge. Achieving this goal requires combining best practices from the GEO community with long-term capacity development. The GEO community produces many useful and relevant results, which they aim to share globally. *Thus, several Flagships and Initiatives of the GEO Work Programme are calling for a knowledge hub as a centralised, efficient means for transferring knowledge and scaling-up applications developed by them*. In this way, the GEO Knowledge Hub will lower the barriers for developing countries to use the petabytes of big Earth observation data openly available.

1.2 Who is it for?

The Knowledge Hub will be useful to a wide range of stakeholders, from national experts needing to report on policy commitments, to individual end users and Small- and Medium-sized Enterprises (SMEs) seeking practical solutions to local environmental challenges. Activities of the GEO Work Programme will be fundamental in both providing methodologies for solving problems and identifying potential end users. Technical experts from research institutions may serve as intermediaries in assisting local end users to benefit from the resources of the Knowledge Hub. GEO intends to leverage the capacity development networks of its partners in a “training the trainers” approach.

1.3 What resources will it contain?

The contents of the *GEO* Knowledge Hub are *linked documents that contain relevant information for Earth observation applications that promote reproducibility, scalabity, and co-design/co-production*. Examples of *documents* include an HTML file, a PDF file (report of paper), a Jupyter Notebook, an R or python markdown file, a GitHub page, a repository entry linking to a dataset store with an assigned Digital Object Identifier (DOI), an Amazon Web Services (AWS) or other links to datasets, OGC service links for data, a video (see Figure 1). We also use the term *document set* to describe a set of documents linked to the same application.

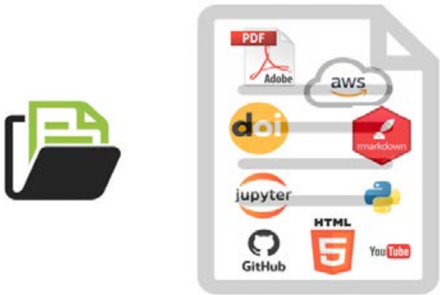

Figure 1 - Examples of documents in the GEO Knowledge Hub

As an example, consider Figure 2, which presents a document set whose five components describe the use of big data analytics for agricultural monitoring: (a) a journal paper that describes the algorithm and the methods; (b) the in situ data published in a repository; (c) an R-markdown file that describes and embeds the algorithm used for the work; (d) a list of images used for processing, stored in the cloud; (e) the results of the analysis. Dataset (b) links to paper (a), but the link to the dataset (b) is not available in the paper's (a) metadata. Neither (a) nor (b) provide a direct link to file (c). As for cloud data (d), there are no established standards for referencing it. Unless the results (e) are in a data repository, they are not available. A critical requirement of the Knowledge Hub is to support linkages between different parts of a document set and across documents.

With the GEO Knowledge Hub, users will have a single entry point to discover and access resources that have been developed by domain experts. Such resources will be of different types to reflect the entire information flow of the research and knowledge pertaining to a domain. The table below provides a summary of these resources.

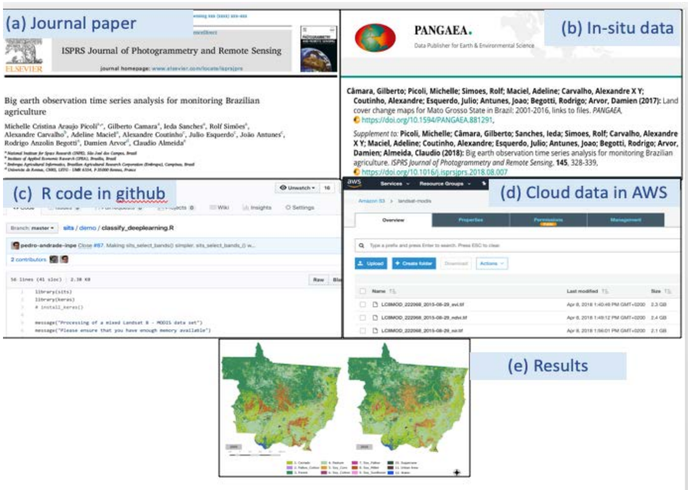

Figure 2 - A document set with five components

Table 1 – Types of knowledge resources organised in the GEO Knowledge Hub

Knowledge resource types	**Sources**
Publications	Scientific Journals and Reports provided by GEO Work Programme activities, with associated DOI (unique object identifier). Post-print copies will be stored in the Knowledge Hub.
Software code, models and tools	Open source software code provided by GEO Work Programme activities, preferably available in Github, with associated DOI. Backup copies stored in the Knowledge Hub.
Remote sensing data	Link provided by the GEOSS Platform (or similar indexing schema) to files in a cloud repository.
In situ data	Link provided by the GEOSS Platform (or similar indexing schema) to files in a recommended data repository, with associated DOI.
Ancillary data	Link provided by the GEOSS Platform (or similar indexing schema) to files in a recommended data repository, with associated DOI.
Output data and products	Link provided by the GEOSS Platform (or similar indexing schema) to files in a recommended data repository, with associated DOI.
Videos	Directly stored in the Knowledge Hub (preferably).
Other relevant documents (e.g., training material)	Directly stored in the Knowledge Hub (preferably).

1.4 How does it work?

A simplified Geo Knowledge Hub data flow diagram is presented in Figure 3, and described in more detail in what follows.

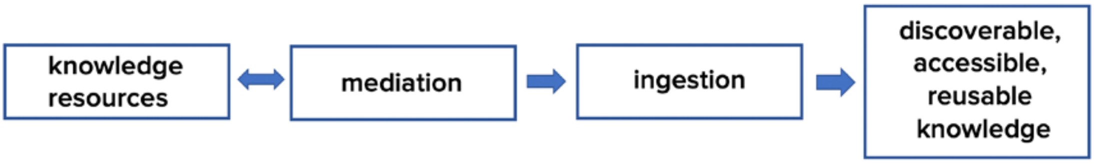

Figure 3 - Data flow on the GEO Knowledge Hub

The *knowledge resources* of the Knowledge Hub include information relevant to the activities of the GEO Work Programme (GWP). We expect each activity to contribute to the hub, sharing results and best practices. This practice will be beneficial to all involved. It will enhance the visibility of GEO's work, by having a unique focal point for the results of the GWP. It will help all those interested in EO to have a place to go to learn how best to use big imagery datasets. The Secretariat will work with the Programme Board as ensure that Flagships and Initiatives of the GWP contribute to expansion of the Knowledge Hub.

Mediation is the process of joint work by the GEO Secretariat and the Knowledge Hub contributors to ensure inclusion of trusted information. Since the GEO community needs reproducible best practices, information going into the Knowledge Hub needs to be verified and organised, and disperse parts of a document set have to be linked. The GEO Knowledge Hub team will interact with the authors so that the methods, data and software are consistent and follow the GEOSS Data Sharing principles, the GEOSS Data Management Principles and other applicable Open Science principles.

Data mediation for the GEO Knowledge Hub includes the following actions:

1. Once contributors to the hub (e.g., Work Programme activities) propose documents for inclusion, the mediator will contact the authors to request shareable copies of their methods, software, and data.
2. Interact with contributors to promote best practices to sharing, such as depositing in situ data on trusted repositories, making software available in github (or similar platforms) and assigning DOIs to software and to documents which do not have it. The moderator will draw on recommendations from Open Science promoters[3].
3. Promote the use of cloud computing solutions and encourage authors to provide versions of their results that work on the cloud.
4. Enter information about the links that connect the publication, software and data into the indexing system of the Knowledge Hub.

It is also useful to consider which tasks fall *outside* the mediator's responsibilities:

[3] Useful recommendations for reproducible research practices include: Sandve et al. (2013) "Ten Simple Rules for Reproducible Computational Research", PLoS Comput Biol, 2013; Morin et al., "A Quick Guide to Software Licensing for the Scientist-Programmer", PLoS Comput Biol 2012; Wilson et al., "Good enough practices in scientific computing", PLOS Comp. Biol., 2017.

1. Quality control on the data provided. All issues of data quality are the responsibility of the authors.
2. Testing reproducibility of the software provided by authors.
3. Conversion of software from one platform or language to another.

The *ingestion* process of the Knowledge Hub will be as automated as possible, but human intervention does the final checks. For best query results, it will include the full text of documents; this requires advanced text-based search capabilities. Given the need of GEO to provide open global access, the knowledge hub will store either open access papers or post-prints[4] of journal papers that are not open access.

Based on the above, we can derive the following *requirements* for the infrastructure of the GEO Knowledge Hub:

R1. Support efficient text-based search.

R2. Link document sets with different components (web pages, PDFs, links to DOIs, videos, Jupyter Notebooks, R markdown, URLs to data, GitHub pages, videos…etc).

R3. Use descriptors compatible with current search engine technologies and emerging solutions.

R4. Be integrated with the GEO website and GEOSS Platform and based on open source software.

R5. Include mediation services for data entry.

2 CLOUD COMPUTING

The cloud computing model is becoming the prevailing mode of work for most medium and large-scale EO applications. Cloud solutions archive large satellite-generated datasets and provide computing facilities to process them. By using cloud technologies, users can share big EO databases and minimise the time to data utilisation. This choice leads to optimised infrastructure investment and increases data and software sharing and reuse. The GEO Knowledge Hub will promote the use of cloud computing for Earth observation data analysis.

One of the key questions associated to applications of cloud computing for Earth observation data is their potential for reuse and reproducibility. At present, there are no standards that address the full issues of interoperability between EO cloud services, increasing the risk of dependency of incompatible proprietary solutions. Thus, the GEO community needs to promote and identify ways to achieve service interoperability between different platforms. GEO will work with OGC to develop open standards for doing Earth observation data analysis in cloud computing platforms. Meanwhile, GEO should promote application portability between different infrastructures, or at least algorithm portability between different cloud providers. This is best achieved by using open source solutions, such as the Open Data Cube.

[4] Most scholarly publishers allow researchers to share *post-prints* in public repositories. Post-prints have the same content as the journal paper minus the formatting. A detailed list of publishers' policies is available at `http://www.sherpa.ac.uk/romeo`.

Based on the above, we derive the following requirements for the GEO Knowledge Hub in relation to use of cloud computing facilities:

R6. Identify and promote solutions that describe big Earth observation data catalogues.

R7. Promote abstract description of methods used for cloud computing to facilitate interoperability.

R8. Support communities of practice to build packages that use open source scripting languages for big Earth observation data analysis in the cloud to promote application portability.

3 IN SITU DATA

This section presents a general discussion of the in situ component in the GEO Knowledge Hub. Although there is solid evidence that making public data to be open creates economic value, some data providers are reluctant to adopt such policy. The reasons for such behaviour are multiple, including:

a) Many countries lack national open data policies or restrict data availability for foreign users as they consider not to benefit from data sharing. Some fear others could use their data for private profit or for environmental-based market restrictions.

b) Many in situ observation networks are funded by time-limited research funds. Sometimes, the data is never released and may even be lost after the project finishes. Even when project data is shared, researchers and institutions use unreliable practices, such as creating short-lived websites.

In keeping with the principle of co-design, GEO needs to work with the country institutions that provide in situ observations to ensure that these institutions benefit from full and unrestricted access to the results to which they contribute. Engaging developing nations as providers and users of in situ data requires a change in attitude from the data analysis and modelling community. All models and analysis methods that use such data should be made public. In case of complex models such as numerical weather predictions or climate change models, at a minimum, all results from these models that can benefit developing nations need to be public. Engaging model developers will thus be an important responsibility for GWP activities. As these activities broaden their scope from local case studies to global products, they need to make sure that these results and products are openly available.

To design and build the in situ data component of the Knowledge Hub, we considered different kinds of in situ data: *(a) operational and/or continuous data collections; (b) data from census or resource assessments; (c) field data collected for research not involving satellite data; (d) field data collected for training, calibration and validation of satellite data*; *(e) data from citizen science and community activities*. Although the boundaries between these categories are fuzzy, they serve as a basis for the establishing requirements for integrating in situ data in the Knowledge Hub.

3.1 In situ data from continuous data collection services

This case is typical of meteorological, marine and hydrological data. Examples include the Argo program, which provides continuous data from 4000 drifting floats[5] distributed across the world's oceans, and national and regional authorities' hydrological networks[6]. Such data is collected by public institutions with specific mandates, which include curation, preservation and dissemination. Many such data collections predate Earth observation satellites. The main challenge for sharing these data sets arises from lack of sustainable public funding.

The Knowledge Hub can contribute to these services by providing metrics of how useful these datasets are, based on their use for producing relevant results. By showing results derived from their data organised and available in the Knowledge Hub, data collection services have further arguments to persuade national and international funding sources to fund their operations. This situation leads to two additional requirements:

R9. Promote the open sharing of modelling software in the Knowledge Hub.

R10. Work with the GEOSS Platform and other dataset search engines such that they access and promote repositories of continuous in situ data collection.

3.2 Data from census or similar field surveys

In most countries, public mapping and statistical agencies are responsible for census, surveys and resource assessments. These data sets are important for producing SDG indicators and for developing models for sustainable development. The combined use of EO and census surveys for improvement of SDG indicators and for projecting scenarios of sustainable development is one of the key areas of interest to GEO.

In terms of data access and its inclusion of these data sets in the Knowledge Hub, GEO has to encourage providers to make them available. Since these are official documents, it is not appropriate to manage them in a GEO database. However, GEO can promote innovative applications that use these data in novel ways. The more countries have access to open source models and methods that increase the value of their data, the more they will share it. Thus, the requirements for the Knowledge Hub and the GEOSS Portal for this data are like those presented in Section 3.1 above.

3.3 Data from field surveys unrelated to satellite observations

Case (c) includes field data used for studies in which space-based instruments are not the primary source of information about the Earth processes involved, nor are they the target. Typical cases is biodiversity data collection and chemical contaminants such as mercury. Many scientific projects related to global environmental change fall under this category. In this area, initiatives such as GBIF have achieved much progress. Here, the best approach to promote the use of accredited and trusted repositories, such as PANGEA, Oak Ridge DAAC, GBIF, and the Environmental Data Initiative (EDI). Similar to data journals such as "Nature Scientific Data", GEO should publish a list of recommended repositories. The GEO Knowledge Hub will also benefit from the search and discovery capabilities of the GEOSS Portal to ensure that data stored in these repositories are visible and accessible.

[5] http://www.argo.net/

[6] See, for example, the UK river flow archive at https://nrfa.ceh.ac.uk/

Sometimes, the institutions involved may prefer to make their data available through GEO. This leads to the need for GEO to build an in situ data repository to handle such data sets

These considerations lead to additional requirements:

R11. Ensure in situ data sets stored in accredited and recommended repositories are indexed.

R12. Build an in situ data repository to ensure long-term curation and preservation of data entrusted to GEO by its community.

3.4 Data from field surveys related to satellite observations

These datasets are associated with research papers and reports in most cases supported by short-term grants. They make up the bulk of the data used for innovative EO applications in forestry and agriculture. It has not been the practice for the EO research community to share such data sets. Many of these data sets are lost or are kept under close control of individual researchers.

GEO has to promote the practice that data associated to papers be deposited in long-term data repositories[7]. All GEO Work Programme Initiatives, Flagships and Community Activities need to incorporate data sharing as a best practice. For example, Figure 4 above shows a multi-part document that includes a DOI for data deposited in PANGEA[8]. Since data repositories offer a DOI with metadata, indexing them in the Knowledge Hub requires limited effort. This leads to an additional requirement:

R13. Provide links from the papers stored in the Knowledge Hub to the accredited long-term data repositories.

As stated above, it might be the case that some institutions prefer to use a GEO-supported data repository, leading to requirement (R11) outlined above.

3.5 Data from citizen science and innovative technologies

Given the challenges of in situ data collection by public institutions and researchers, the Knowledge Hub needs to support innovative methods of data collection, which include Citizen Science and new approaches such as mobile communication and sensor networks. For example, SDSN TReNDS and GPSDD (Global Partnership for Sustainable Development Data) have put together the "Value of Data" report[9], which highlights cases where non-conventional approaches can improve data collection for SDGs.

Most data collection in Citizen Science uses mobile applications to record in situ data and to transmit such data to a central repository. Alternative approaches use sensor networks or high-resolution imagery. All such approaches face similar challenges: (a) *How to convert spontaneous, unorganised contributions into trusted datasets?* and (b) *How to provide long-term repositories for data produced by many Citizen Science projects?*

Citizen Science initiatives are unlike any of the in situ data sources mentioned above. Data comes in irregular intervals with different levels of quality. Most institutions that promote Citizen Science are NGOs that do not have adequate means of building long-term

[7] For a list of accredited repositories in Earth and Environmental Sciences, see `https://www.nature.com/sdata/policies/repositories#envgeo`

[8] `https://www.pangaea.de/`

[9] `https://www.sdsntrends.org/valueofdata`

data repositories. GEO could provide a major service to the Citizen Science community if it were to build a long-term data repository. Thus, we have an additional requirement for the Knowledge Hub regarding Citizen Science data:

R14. Build a data repository for long-term archival, where needed, of Citizen Science data associated with in situ observations; otherwise link with existing Citizen Science repositories for seamless retrieval of data.

4 ANALYSIS-READY DATA

GEO intends to work with CEOS to promote generation of analysis-ready data. CEOS has been working to provide common specifications for analysis-ready data for land use as part of the CARD4L initiative[10]. The specifications include recommendations for surface reflectance and for radar backscatter. CARD4L products aim at a wide range of applications, including time series analysis and multi-sensor application development. They support rapid ingestion and exploitation via high-performance computing, cloud computing and other future data architectures.

GEO has identified the CEOS CARD4L specification to be useful in two contexts. The first and best option is that space agencies produce analysis-ready data for all their assets, following these specifications. This data would then be moved to cloud platforms. If that is not possible, GEO has to engage with the community to convert these specifications into open source software tools that can be deployed in the different cloud platforms.

This leads to a further requirement for the Knowledge Hub:

R15. Promote and disseminate open source software for building single and multi-satellite analysis ready data that work on EO cloud platforms, and supports the CEOS CARD4L specifications.

5 REQUIREMENTS OF THE GEO KNOWLEDGE HUB

5.1 Review of the user requirements

The requirements for the GEO Knowledge Hub set up in the previous sections are recalled below:

R1. Support efficient text-based search.

R2. Link document sets with different components (web pages, PDFs, links to DOIs, videos, Jupyter Notebooks, R markdown, URLs to data, GitHub pages, videos…etc).

R3. Use descriptors compatible with current search engine technologies and emerging solutions.

R4. Be integrated with the GEO website and GEOSS Platform and based on open source software.

R5. Include curation services for data entry, based on contributions from the GEO community.

[10] http://ceos.org/ard/

R6. Describe big Earth observation data catalogues.

R7. Promote abstract description of methods used for cloud computing to facilitate interoperability.

R8. Support communities of practice to build packages that use open source scripting languages for big Earth observation data analysis in the cloud to promote application portability.

R9. Include applications that use models and support the open sharing of modelling software.

R10. Work with the GEOSS Platform and other dataset search engines such that they access and promote repositories of continuous in situ data collection.

R11. Ensure in situ data sets stored in accredited and recommended repositories are indexed.

R12. Build an in situ data repository, managed by the Secretariat, to ensure long-term curation and preservation of data entrusted to GEO by its community.

R13. Provide links from the papers stored in the Knowledge Hub to the accredited long-term data repositories.

R14. Build a data repository for long-term archival, where needed, of Citizen Science data associated with in situ observations; otherwise link with existing Citizen Science repositories for seamless retrieval of data.

R15. Promote and disseminate open source software for building multi-satellite analysis ready data that work on EO cloud platforms.

5.2 Implementation requirements

The GEO Knowledge Hub also has to follow some implementation requirements. For reference, we present below the set of recommendations drawn up by Repositories Expert Group of the FORCE11.org to ensure proper data citation[11]. We consider these recommendations relevant to the implementation of the GEO Knowledge Hub. In the detailed implementation plans, we will consider if all these guidelines are pertinent

TABLE 1 - Guidelines for Repositories from FORCE11.org

Level	#	Guideline
Required	G1	All datasets intended for citation *must* have a globally unique persistent identifier that can be expressed as unambiguous URL.
	G2	Persistent identifiers for datasets *must* support multiple levels of granularity, where appropriate.
	G3	This persistent identifier expressed as URL *must* resolve to a landing page specific for that dataset.
	G4	The persistent identifier *must* be embedded in the landing page in machine-readable format.
	G5	The repository must provide documentation and support for data citation.

[11] Fenner, M. et al. (2016). A Data Citation Roadmap for Scholarly Data Repositories. https://doi.org/10.1101/097196

Recommended	G6	The landing page *should* include metadata required for citation, and ideally also metadata helping with discovery, in human-readable and machine-readable format.
	G7	The machine-readable metadata *should* use schema.org markup in JSON-LD format.
	G8	Metadata *should* be made available via HTML meta tags to facilitate use by reference managers.
	G9	Metadata *should* be made available for download in Bibtex and/or another standard bibliographic format.
Optional	G10	Content negotiation for schema.org/JSON-LD and other content types *may* be supported so that the persistent identifier expressed as URL resolves directly to machine-readable metadata.
	G11	HTTP link headers *may* be supported to advertise content negotiation options

6 PROOF OF CONCEPT

The EAG recommends that the GEO Secretariat carries out a proof of concept demonstrating the GEO Knowledge Hub. This demonstration will address the requirements for the Knowledge Hub as described above. It will be developed using only free and open source software, drawing as much as possible on existing community solutions for Open Science and on accepted best practices for knowledge sharing.

This proof of concept will be built in consultation with the GEO Programme Board, the GEO Work Programme, and with the GEOSS Implementation Coordination Task Force. The proof of concept will be demonstrated at the GEO Plenary 2019, in Canberra, Australia.

Based on the response of the GEO community to this proof of concept, the GEOSS Implementation Coordination Task Force will propose a plan for further development of the GEO Knowledge Hub, to be presented for decision to the GEO Executive Committee, at their March 2020 meeting.

Annex B

List of Members of the Expert Advisory Group and Contributors to the Document

Name	Affiliation
Ariane de Bremmond	Global Land Programme, University of Bern
Carolin Richter	Director, Global Climate Observing System (GCOS), WMO
Jessica Espey	TReNDS Project Director, UN Sustainable Develop Solutions Network
Laetitia Navarro	Executive Secretary, GEOBON
Lubia Vinhas	Associate Director for Earth Observation, National Institute for Space Research (INPE)
Marie-Françoise Voidrot	Europe Director, Innovation Programme, Open Geospatial Consortium
Yana Gevorgyan	Senior International Relations Specialist, National Oceanic and Atmospheric Administration (NOAA)
Anthony Milne	Professor of Geography and Remote Sensing, University of New South Wales
Baudouin Raoult	Principal Software Strategist, European Centre for Medium-Range Weather Forecasts
Christopher Justice	Chair, Dept of Geographical Sciences, University of Maryland
Edzer Pebesma	Co-director, Institute for Geoinformatics, University of Münster
Eldrich Frazier	Senior Expert, Federal Geospatial Data Committee (USA)
Fernando Ramos	Senior Researcher, National Institute for Space Research (INPE), Brazil
Geoff Sawyer	Secretary General, European Association of Remote Sensing Companies (EARSC)
Guido Schmidt-Traub	Executive Director, UN Sustainable Development Solutions Network
Gu Xingfa	Deputy Director General, Aerospace Information Research Institute (AIR), Chinese Academy of Sciences (CAS)
Imraan Saloojee	Stakeholder and Business Development Head. South African National Space Agency (SANSA)
Ivan DeLoach	Executive Director, Federal Geospatial Data Committee (USA)
Ivan Petiteville	Centre for Earth Observation (ESRIN), European Space

	Agency (ESA)
Joost van Bemmelen	Centre for Earth Observation (ESRIN), European Space Agency (ESA)
Junji Inoue	Executive Management Director, Remote Sensing Technology Center (RESTEC)
Matthew Hansen	Global Land Analysis and Discovery Group, co-director, University of Maryland
Massimo Craglia	Team Leader, Digital Economy Unit, Joint Research Centre (EC)
Meng Kimhor	National Institute for Statistics, Government of Cambodia
Michael Obsersteiner	Head, Ecosystems Services and Management, International Institute for Applied System Analysis (IIASA)
Mikko Strahlendorff	Space Adviser, Finnish Meteorological Institute
Noel Gorelick	Software Engineer, Google Earth Engine, Google Switzerland
Nuno Catarino	Payload Data Ground Segment Division Head, Deimos Engenharia
Osamu Ochiai	Associate Senior Administrator, Japan Aerospace Exploration Agency (JAXA)
Paolo Mazetti	Istituto Sull'Inquinamento Atmosferico (IIA), Consiglio Nationale di Recherche (CNR), Italy
Pedro Gonçalves	Founder and CTO, TerraDue CTL
Pierre Defourny	Earth and Life Institute, Université Catholique de Louvain
Robert Woodcock	Research Scientist and Consultant in Earth Data and Analytics infrastructures, CSIRO (Australia)
Stefano Nativi	Big Data Lead Scientist, Digital Economy Unit, Joint Research Centre
Steven Brumby	Chief Science Advisor, Descartes Labs
Wolfgang Wagner	Head of Dept of Geodesy and Geoinformation, Technische Universität Wien