GEO-XVI - 6-7 November 2019 GEO-XVI-6.2(Rev1)

Proposed Design and Proof of Concept of the GEO Knowledge Hub

This Document is submitted to the GEO-XVI Plenary for decision.

GEO Plenary will take note of the proposed proof of concept of the Knowledge hub. GEO Plenary is expected to decide on whether to delegate authority to the GEO Executive Committee to oversee its further development. Should the GEO Plenary so decide, the GEO Secretariat, in consultation with the GEOSS Implementation Development Task Team and GEO Programme Board, 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.

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 a given 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 GKH is one component of the GEOSS Infrastructure and its development is included in the GEOSS Infrastructure Development Foundational Task, which will be part of the 2020-2025 GWP. The 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 GKH will be consistent with the other components of the GEOSS Infrastructure.

1.1 Why is the GEO Knowledge Hub needed?

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 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 GWP 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 GKHwill lower the barriers for developing countries to use the petabytes of big Earth observation data openly available.

1.2 Who is it for?

The GKH 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 GWP 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 GKH. 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 GKH are *linked documents that contain relevant information for Earth observation applications that promote reproducibility, scalability, 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.

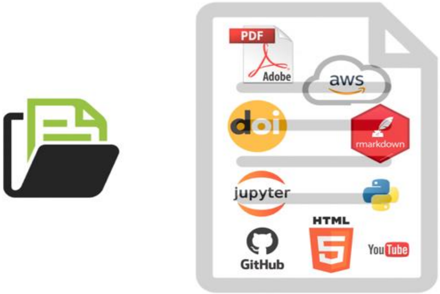
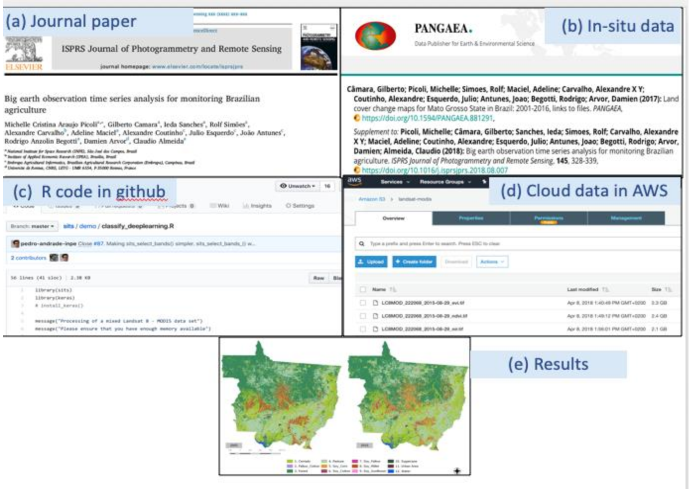


Figure 1 - Examples of documents in the GEO Knowledge Hub

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; and (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 GKH is to support linkages between different parts of a document set and across documents.

With the GKH, 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.



ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Big earth observation time series analysis for monitoring Brazilian agriculture

Michelle Cristina Araujo Picoli, Gilberto Camara, Ieda Sanches, Rolf Simões, Alexandre Carvalho, Adeline Maciel, Alexandre Coutinho, Julio Esquerdo, João Antunes, Rodrigo Anzolin Begotti, Damien Arvor, Claudio Almeida

PANGAEA.

Data Publisher for Earth & Environmental Science

Câmara, Gilberto; Picoli, Michelle; Simoes, Rolf; Maciel, Adeline; Carvalho, Alexandre X Y; Coutinho, Alexandre; Esquerdo, Julio; Antunes, Joao; Begotti, Rodrigo; Arvor, Damien (2017): Land cover change maps for Mato Grosso State in Brazil: 2001-2016, links to files. *PANGAEA*, https://doi.org/10.1594/PANGAEA.881291,

Supplement to: Picoli, Michelle; Câmara, Gilberto; Sanches, Ieda; Simoes, Rolf; Carvalho, Alexandre X Y; Maciel, Adeline; Coutinho, Alexandre; Esquerdo, Julio; Antunes, Joao; Begotti, Rodrigo; Arvor, Damien; Almeida, Claudio (2018): Big earth observation time series analysis for monitoring Brazilian agriculture. *ISPRS Journal of Photogrammetry and Remote Sensing*, **145**, 328-339, https://doi.org/10.1016/j.isprsjprs.2018.08.007

Branch: master ▾ sits / demo / classify_deeplearning.R

pedro-andrade-inpe Close #87. Making sits_select_bands() simpler. sits_select_bands_() w...

2 contributors

```
56 lines (41 sloc) | 2.38 KB

library(sits)
library(keras)
# install_keras()

message("Processing of a mixed Landsat 8 - MODIS data set")
message("Please ensure that you have enough memory available")
```

Amazon S3 > landsat-modis

Name	Last modified	Size
LC8MOD_222068_2015-08-29_evi.tif	Apr 8, 2018 1:40:49 PM GMT+0200	3.3 GB
LC8MOD_222068_2015-08-29_ndvi.tif	Apr 8, 2018 1:49:12 PM GMT+0200	2.4 GB
LC8MOD_222068_2015-08-29_swir.tif	Apr 8, 2018 1:56:01 PM GMT+0200	2.1 GB

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 GWP activities, with associated DOI (unique digital object identifier). Post-print copies will be stored in the GKH.
Software code, models and tools	Open source software code provided by GWP activities, preferably available in Github, with associated DOI. Backup copies stored in the GKH.
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 GKH (preferably).
Other relevant documents (e.g., training material)	Directly stored in the GKH (preferably).

1.4 How does it work?

A simplified GKHdata flow diagram is presented in Figure 3, and described in more detail in what follows.

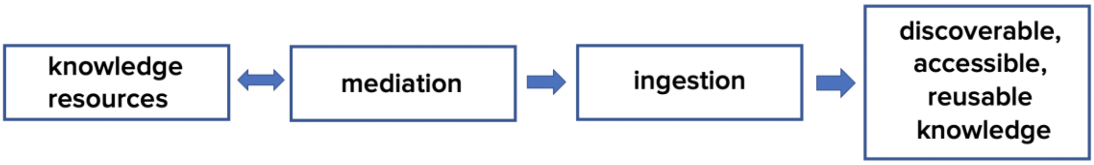


Figure 3 - Data flow on the GEO Knowledge Hub

The *knowledge resources* of the GKH include information relevant to the activities of the GWP. We expect each activity to contribute to the Hub, sharing results and best practices. This practice will be beneficial to all involved, by enhancing the visibility of GEO's work and providing 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 EO datasets. The Secretariat will work with the Programme Board to ensure that Flagships and Initiatives of the GWP contribute to expansion of the GKH.

Mediation is the process of joint work by the GEO Secretariat and the GKH contributors to ensure inclusion of trusted information. Since the GEO community needs reproducible best practices, information going into the GKH must be verified and organised, while disperse components of a document set have to be linked. The GKH 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 GKH includes the following actions:

1. propose documents for inclusion: once contributors to the Hub (e.g., GWP activities), 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[1].

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 GKH.

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

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 GKH will be as automated as possible, but human intervention is required for 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 GKH will store either open access papers or post-prints[2] of journal papers that are not open access.

Based on the above, we can derive the following *requirements* for the infrastructure of the GKH:

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

[1] 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.

[2] 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`.

optimised infrastructure investment and increases data and software sharing and reuse. The GKH 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.

Based on the above, we derive the following requirements for the GKH 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 GKH. 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 GKH, 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 often fluid, they serve as a basis for the establishing requirements for integrating in situ data in the GKH.

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[3] distributed across the world's oceans, and national and regional authorities' hydrological networks[4]. 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 GKH 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 GKH, 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 GKH.

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 GKH, 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 GKH and the GEOSS Portal for this data are like those presented in Section 3.1 above.

[3] `http://www.argo.net/`

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

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 GKH 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. 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[5]. All GWP 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[6]. Since data repositories offer a DOI with metadata, indexing them in the GKH requires limited effort. This leads to an additional requirement:

R13. Provide links from the papers stored in the GKH 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 GKH needs to support innovative methods of data collection, which include Citizen Science and new approaches such as mobile communication and sensor networks. For

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

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

example, SDSN TReNDS and GPSDD (Global Partnership for Sustainable Development Data) have put together the “Value of Data” report[7], 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 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 GKH 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[8]. 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 GKH:

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.

[7] https://www.sdsntrends.org/valueofdata

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

5 REQUIREMENTS OF THE GEO KNOWLEDGE HUB

5.1 Review of the user requirements

The requirements for the GKH 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.

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 GKH 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 GKH also has to follow certain 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[9]. We consider these recommendations

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

relevant to the implementation of the GKH. 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.
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 has recommended that the GEO Secretariat carry out a proof of concept demonstrating the GKH. This demonstration will address the requirements for the GKH 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 is being built as a joint effort of the GEO Secretariat, the GEOSS Platform and GEOGLAM teams. The proof of concept, to be showcased at the GEO-XVI Plenary 2019, in Canberra, Australia, will consist of a short demonstration of a user scenario from GEOGLAM, leveraging the GKH functionalities.

The demonstration will show how a hypothetical user, such as a scientific officer in a Ministry of Agriculture, can discover, access and reproduce a solution, presented as a knowledge package, to identify crop areas using the Sen2Agri tools from GEOGLAM/UCL.

As part of that knowledge package, users have access to a cloud-based working environment where they can run the code, combining cloud stored analysis ready remote sensing, in situ data and models to reproduce the proposed solution made openly available by the knowledge provider.

Based on the response of the GEO community to this proof of concept, the GEO Secretariat, in consultation with the GEOSS Implementation Development Task Team, will propose a plan for further development of the GKH, to be presented for decision to the GEO Executive Committee, at their March 2020 meeting.