GEOGLAM In-situ Guidance and Workplan

Table of Contents

Note on Contributors

This report was written by the GEOGLAM in-situ data working group, with inputs from the GEO community and other key stakeholders.

Editors (in alphabetical order): Hendrik Boogaard, Florian Franziskakis, Sven Gilliams, Ian Jarvis

1 Introduction

The purpose of this report is to present the key findings and actions developed during the GEOGLAM in-situ workshop, November 2022.

Starting in 2021 GEOGLAM decided to see what the community could do to provide enhanced coordination of in-situ data to support EO based agricultural monitoring. By 2022 this evolved to a vision for a harmonized global in-situ reference data system to support R&D and operational implementation. The scope of the activity was to support key GEOGLAM priorities expressed in the development of the GEOGLAM Essential Agriculture Variables (EAV's). A draft work plan and guidance document was drafted in early 2022 (Link). From the guidance the following principles of the in-situ activity were articulated:

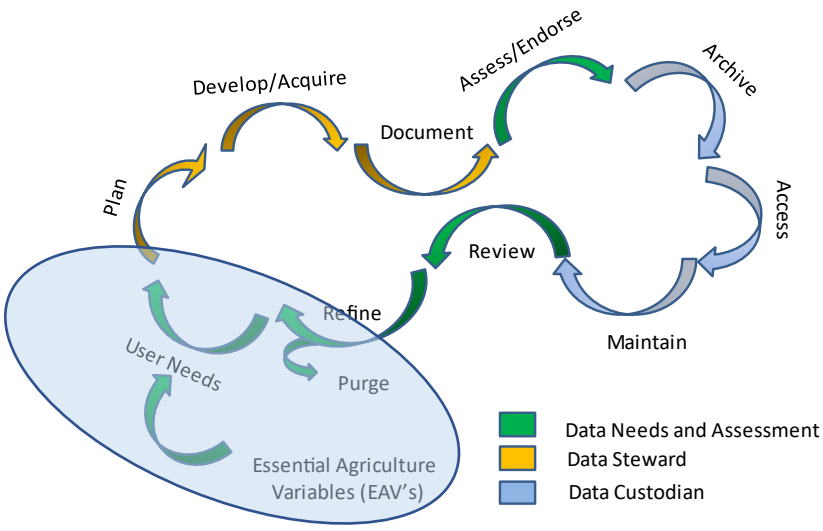
- GEOGLAM will coordinate in-situ data within a management life cycle that is driven by requirements, and explicitly defines the roles of data stewards, custodians, in-situ working group, and the GEOGLAM Secretariat.
- GEOGLAM will adopt a Federated-Hybrid approach. Wherever possible we will link to data managed by the data stewards. Where a sufficient level of curation is not available, GEOGLAM will store and curate data that helps address GEOGLAM requirements.
- GEOGLAM will make data open by default but where required will include the ability to limit access to sub-sets of authorized users.
- GEOGLAM will promote Interoperable data structures to ensure that in-situ data can be accessed and utilized as seamlessly as possible. Further GEOGLAM will create harmonized versions of in-situ data that address EAV requirements.
- GEOGLAM will make quality assessment integral to the life cycle and use this to inform endorsement.
- GEOGLAM will promote shared collection tools, collection standards and best practices.
- GEOGLAM will work with partners to develop a data access portal with application programming interface(s) (API) to seamlessly search and access data of interest based on thematic, spatial, and temporal requirements (as driven by EAV's).
- GEOGLAM will utilize existing IP and infrastructure wherever possible (AgroStac, ML Hub, GEO Knowledge Hub, etc.).

The working group decided to take the proposal forward for broader community input, to improve and refine the plans and get buy-in. To this end, the first GEOGLAM In-situ Coordination Workshop was organized at WMO in Geneva on November 21-22, 2022. The agenda is attached to this document as Appendix A.

During the workshop much of the time was set aside for feedback. This contributed to several improvements and broader community consensus on the way forward and the required actions to implement the technical and institutional aspects of a data life-cycle management process for open and free agricultural reference data. The scope included both retaining legacy data and supporting the acquisition and sharing of future collections. The underlying report aims to align actions towards implementation and identify remaining gaps requiring incremental funding and/or technical support.

To organize discussion, the breakouts were parsed into three broad thematic groupings based on the in-situ data life-cycle, and this report reflects this structure. The groups were User Needs; Data Acquisition; Data Usability; and Data Structure and Maintenance Infrastructure. The results are summarized in the next chapters and Appendix B.

2 User Needs



2.1 Introduction

The GEOGLAM in-situ coordination initiative is guided by the requirements for the implementation of the Essential Agriculture Variables (EAVs). In turn, EAV priorities are driven by the requirements for the major policy drivers of GEOGLAM. Foremost, these drivers will include United Nations Framework Convention on Climate Change (UNFCCC) monitoring needs for adaptation and mitigation; United Nations Office for Disaster Risk Reduction (UNDRR), particularly early warning; and the communities' contribution to the UN Sustainable Development goals (SDGs).

In the context of GEOGLAM, EAVs are state and change variables that track essential aspects of agricultural production and thus are fundamental to a wide range of applications across multiple policy dimensions. The EAV mapping and attribute variables are presented in Figure 1 (more details can be found at AgVariables.org).

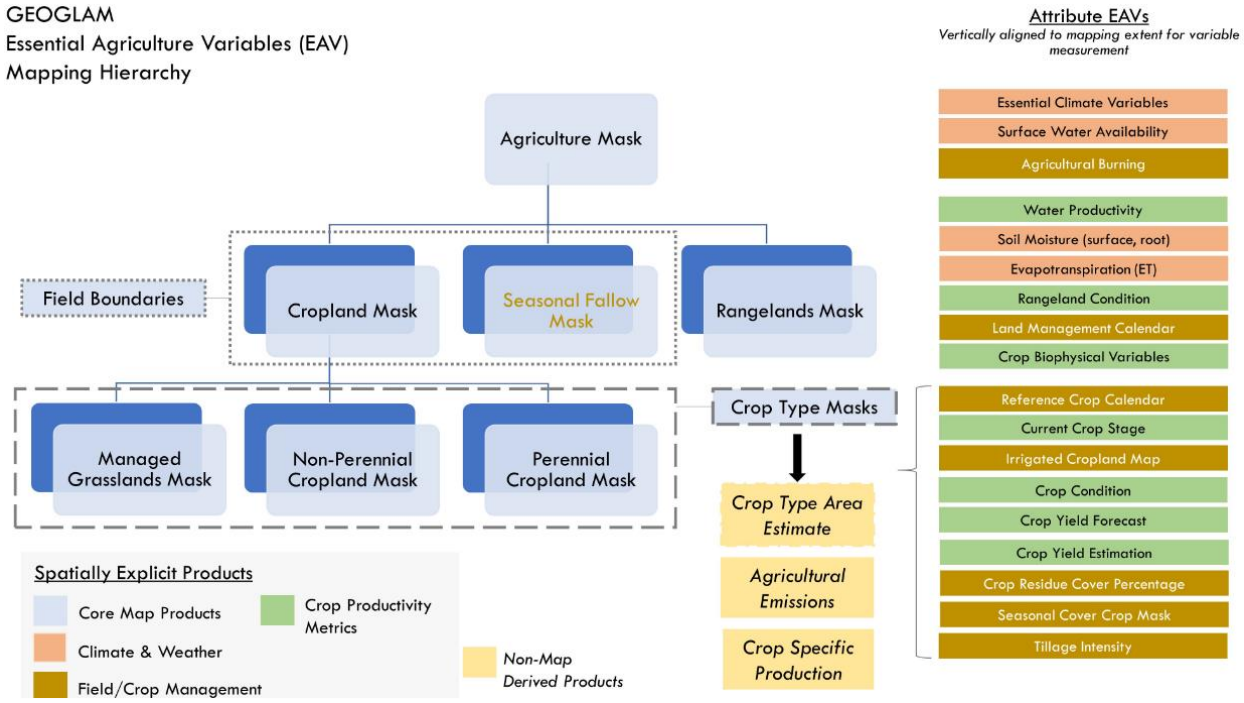


Figure 1. Essential Agriculture Variables (AgVariables.org)

2.2 Prioritization of the Essential Variables

During the November 2022 in-situ workshop, there was general agreement for the policy driven EAV approach. In the discussions, a consensus was established around a set of priority state and change EAVs that would be initially supported, which included:

1. Crop map;
2. Crop condition;
3. Field boundary;
4. Crop damage[1];
5. Crop yield including crop phenology.

[1] Weather related, conflict, pest, and disease, includes early warning response

2.3 Gaps Assessment

GEOGLAM has defined EAV's to be inherently scale independent. EAV scale(s) are conveyed by the spatial and temporal needs outlined by the policy requirement, and in turn define in-situ data needs. Consequently, the first step in EAV implementation is to document EAV requirements corresponding to priority policy needs.

Once the requirements are clearly understood, the second step is to conduct a gaps assessment. The gaps assessment links current information and capacity capabilities to the requirements and identifies where gaps remain. Key aspects of the gaps assessment include:

- **Earth observation (EO) data requirements**: Identify EO needs, link needs to existing open data resources, and identify where gaps exist.
- **Open-source analytical tools**: Identify existing best practices, and open-source tools for operational EAV generation. Identify where gaps exist in tool development due to technical and knowledge limitations.
    - Use technical gaps to prioritize application development;
    - Use knowledge gaps to inform research priorities.
- **In-situ data needs**: Identify the in-situ data needs based on operational requirements, link to existing open in-situ data resources and identify gaps. Use gaps assessment to develop in-situ data strategies.
- **Sustained systematic analytics**: Identify institutional requirements to sustainably deliver EAV information products. Identify gaps in current institutional capacity and prioritize actions for mitigation.

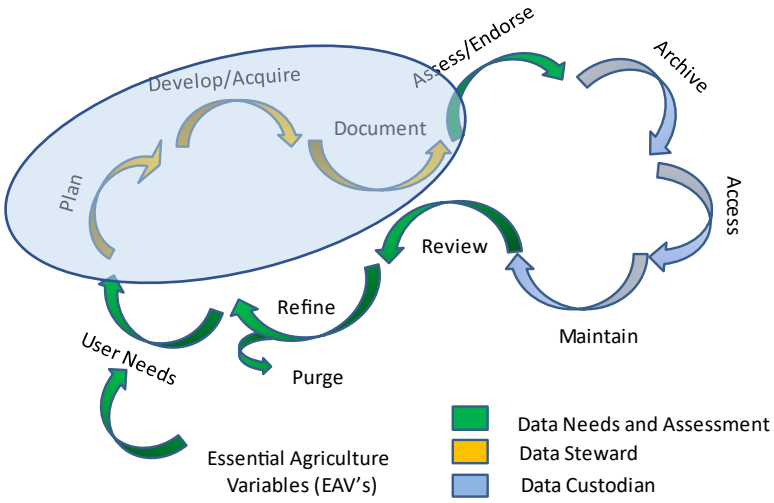
2.4 Actions Required to Address User Needs

The in-situ workshop identified three actions to address user needs and inform the in-situ data coordination strategy:

1. **Action UN.1**, **Identify priority policies and corresponding EAV needs**: There was general agreement to initially pilot the process using the EAV requirements linked to one policy driver. The UNFCCC climate mitigation policy driver was selected. Consequently, initial work will be based on the needs for the UNFCCC-Agriculture Forest and Other Land Use (AFULO) input into the Global Stocktake (GST), and Nationally Determined Contributions (NDC) inventories and change monitoring. The priority EAV will be crop type mapping at the global to national scale. This was selected because:
    - Monitoring agricultural land use/cover for climate mitigation provides an intersection of priority policy needs with the top priority EAVs;
    - Builds on work already undertaken in the WorldCereal project; and
    - Contributes to the work of the Committee on Earth Observing Satellites (CEOS), AFOLU Roadmap.

2. **Action UN.2**, **Document EAV Requirements**: Record the crop mapping EAV requirements to address UNFCCC climate mitigation and adaptation needs.
3. **Action UN.3**, **Gaps assessment**: Conduct a gaps assessment including in-situ requirements with an initial focus on crop mapping EAVs to address UNFCCC, GST and NDC needs. It was estimated that this would require dedicated support for coordination (Estimated 1 FTE Post Doc, approximately $60k/year).
4. **Action UN.4**, **EAV stewardship**: Identify additional experts to support the development of the EAVs by becoming EAV stewards, and follow-up with interested participants who self-identified during the breakout sessions.
    - **Action UN.4b**, **Rangelands**: In discussions about stewardship, the gap in Rangelands was flagged as particularly acute and a specific action was identified. Interest and leadership for a GEOGLAM rangeland working group was also identified.

3 Data Acquisition (plan, develop and acquire)



3.1 Introduction

There are many well established open-source field collection tools that are used by the community. Standardizing these tools would be beyond the scope of this initiative, but work is required to develop best practices for data collection and standards for the variables collected (i.e., crop legends, collection methods, measurement protocols, etc.). Additionally, back-end data structures are not standardized, adding to the challenge of post collection data harmonization. Consequently, harmonizing data from multiple sources (a

goal of our initiative) takes a great deal of manual effort and expert interpretation. Understanding that there is no value in standardizing collection tools and applications, the next best thing is to develop interoperability of the existing tools to reduce the need for human intervention in the harmonization process. The workshop identified several data acquisition actions to support the process of data harmonization.

3.2 Actions Required to Address Data Acquisition

To build on existing efforts and to collect more in-situ data, a framework under the GEOGLAM umbrella is needed to proactively inform data collection efforts that enhance data consistency and interoperability. Key actions include:

- **Action DA.1, Communication to funding organizations**: Prioritize new data collection needs and better communicate these needs to funding organizations, and the GEOGLAM community at large.
- **Action DA.2, Data acquisition best practices**: Document existing best practices for data acquisition and create a "one stop shop" for guidance. This includes:
    - Building on and reinvigorating the existing JECAM guideline process;
    - Creating and documenting consistent variable definitions and collection methods; and
    - Developing back-end data structures for collection tools that are interoperable.
- **Action DA.3, The case for open data**: Develop stakeholder communications to convey the value proposition for sharing in-situ data.
- **Action DA.4, In-situ sharing best practices**: Create and assemble guidance on engaging stakeholders on open in-situ data sharing (e.g., government, academia, farmers). Includes guidance on data rights, sharing policies and agreements for best practices.
- **Action DA.5, In-situ registry of field campaigns**: Develop a registry of in-situ activities so the community can share their field collection intentions, including:
    - Where and when field activities are taking place;
    - What variables are being collected.

Additional cross-cutting actions for consideration:

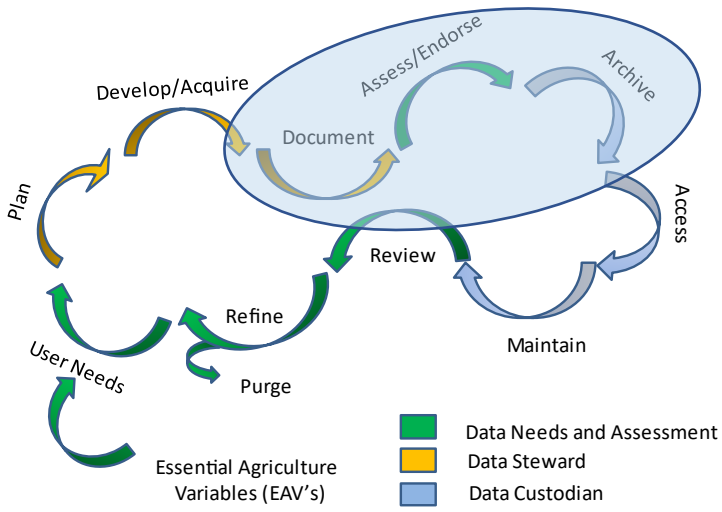
What are the key frontiers that need to be advanced or evolved to support coordinated in-situ data development?

- Admin support, meetings, inviting engagement and building the coalition:
    - Need support for coordination. GEOGLAM needs money to initiate and leverage;
    - Develop a proposal for potential funders;
    - Raise this in EUROGEO (e.g., inclusive financing, early warning systems).

Game-changers: what could be done if there were no limits to resources?

- Explore the interest in GEO developing a coordinated, cross GEO Work Programme approach towards a terrestrial monitoring framework. This initiative would:
    - Integrate and harmonize in-situ data across communities (i.e., forests, wetlands, mountains, agriculture);
    - Develop efficiencies, promote data sharing and a long-term vision for in-situ data to support LC/LU state and change monitoring;
    - Support policy drivers associated with climate adaptation and mitigation, land degradation, and early warning.
- Develop tools to access anonymized, aggregated data from public and private sector entities. This would break down barriers to data sharing from sources that must ensure the clients/constituents remain anonymous.

4 Data Usability (document and archive)



4.1 Introduction

Global research tends to continually adopt a more open approach towards open science and the sharing of research data as a public good. However, the operational use of open, published data sets is limited due to several factors including limited access, unclear and proprietary formats, inadequate data standardization, unknown data quality, inconsistent and/or incomplete metadata, unclear (re)use of data policies. In other words, the (meta)data does not always fully comply to the Findable, Accessible, Interoperable,

Reusable (FAIR) standards. While larger initiatives may have resources and facilities that eventually reach FAIR standards, it is different for individual researchers and smaller initiatives as they often lack the right expertise and technical solutions to publish data according to FAIR standards. In such cases, the second-best option is that data is published in a way that the data is findable and, if possible, accessible. We cannot expect the whole community to publish data in a fully "FAIR" manner. However, there is the need for a community specific solution to further harmonize in-situ data according to FAIR standards and make it ready for use (see Figure 2).

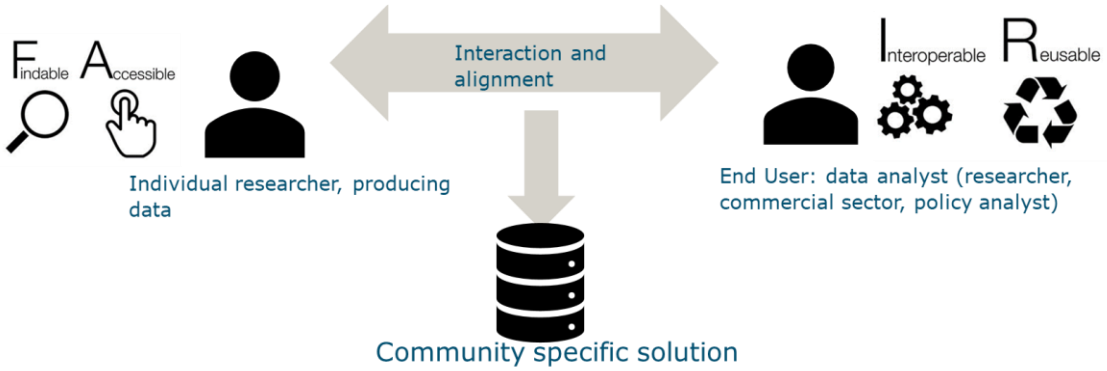


Figure 2. Top et al., 2022, Cultivating FAIR principles for agri-food data (https://doi.org/10.1016/j.compag.2022.106909)

During the November 2022 in-situ workshop, all participants acknowledged the importance to review and further harmonize data for certain applications (e.g., the European Space Agency (ESA) funded WorldCereal project and AGROSTAC, developed within Agricultural Model Intercomparison and Improvement Project (AGMIP) and GEO-related projects). In support of crop type and cropland maps, WorldCereal built an open, harmonized reference data repository at the global extent, ready for model training and product validation. The repository contained data from different sources including the GEOGLAM-JECAM sites, the Radiant MLHub, the Future Harvest (CGIAR) centers, LACO-Wiki, IIASA's Geo-Wiki and data from individual project contributions. Recent data were collected, harmonized, and annotated. Currently, the repository holds approximately 75 million harmonized observations of which a large share is public. The repository can be used for either the calibration of image classification, deep learning algorithms or the validation of Earth observation generated products, such as global cropland extent and crop type maps. The AGROSTAC initiative collects and harmonizes georeferenced open data surrounding key agronomy observations such as crop type, phenology, biomass, yield, and leaf area. Published, open data sets are screened for key observations, and selected data goes through a dedicated data curation procedure. In this procedure, metadata are checked

and completed using all information available in the data files, supporting documents and associated publications. Data are converted into standard units, and phenology events are mapped to the BBCH scale.

The participants stressed the importance of independent review and quality assessment of open, published data. This can be realized through publishing in repositories that support review, and data journals. Publication and review might become part of future research and service contracts. Resources for review depend on the intensity and depth of the review. To ease the review task, automatic tools could be developed and applied: out-of-bounds, outlier detection, checks of spatial, temporal, and thematic accuracy. In the specific case of WorldCereal, users are asked to apply a quality assessment before uploading and sharing data, and automated checks are applied during the upload. Finally, a moderator is foreseen who will review uploaded data before sharing with the community.

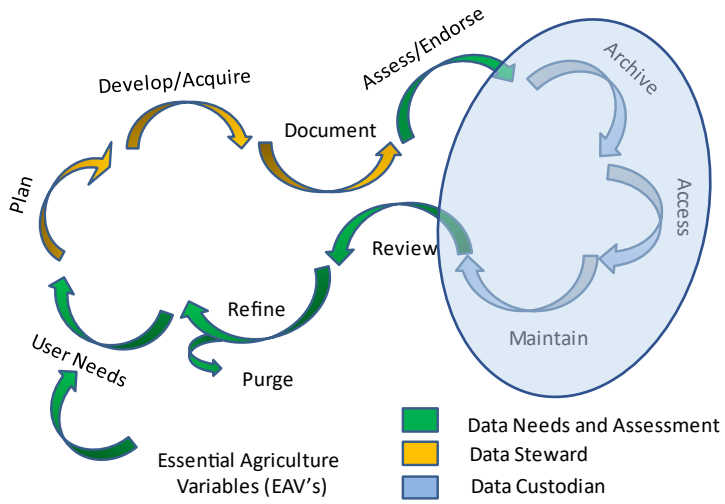
4.2 Actions Required to Address Data Usability

To stimulate and support the re-use of in-situ data we have defined the following actions.

- **Action DU.1**: Institutionalize, sustain, and update the WorldCereal reference data repository. The currently harmonized data can be accessed via a public repository: https://zenodo.org/communities/worldcereal-rdm and https://worldcereal-rdm.geo-wiki.org. Both ways give access to the harmonized data sets but do not give access to the API rest interface which facilitates 1) flexible data retrieval e.g. data for a certain crop or region and 2) data upload which can be integrated with other applications or workflows. We recommend transferring and sustaining the WorldCereal reference data repository, including the API rest interface, in a new infrastructure supported by GEOGLAM. It requires ad-hoc and running costs estimated around several k Euros.
- **Action DU.2**: Encourage the community to publish land cover and crop type data following the open science and open data principles and more specifically the FAIR principles and the GEO Data Sharing and Data Management Principles (GEO-DMP). These principles need to be clearly introduced and explained including aspects like standards (JECAM guidelines, WorldCereal protocol, vocabularies or ontologies such as AGRO, see also Chapter 3), data curation (e.g. tools for quality control), data publication (suggested repositories, see also Chapter 5), and data management plans. See also:
  - "Revised GEO Data Sharing and Data Management Principles" (link).
  - www.go-fair.org
  - fairaware.dans.knaw.nl
- **Action DU.3**: Explore business models to strengthen and sustain harmonization hubs (data holdings) such as WorldCereal and AGROSTAC that review and harmonize open and published data. This requires active networking, promotion, and acquisition.

- **Action DU.4**: Further develop methods and protocols to automate and support the review and harmonization of open and published data. As example WorldCereal published a protocol to harmonize reference data (see 10.5281/zenodo.7584463). Furthermore, notebooks and pipelines can be developed to harmonize data of on-going field surveys (e.g., AAFC Crop Inventory, COPERNICUS-GEOGLAM, JECAM, NASA Harvest, World Food Programme (WFP), Digital Earth Africa).

5 Data Structure and Maintenance Infrastructure (archive, access and maintain)



5.1 Introduction

During the workshop, technical aspects surrounding where to store in-situ data and how to manage it were discussed. Many partners either used a distinct solution or didn't have one that satisfies their needs. Moreover, the need for institutions to build trust with the people on the ground collecting the data was seen as a key issue to solve to advance in-situ data sharing.

A proposed solution was presented during the workshop with three main components:

- CKAN[2] structure to store the data
- STAC[3] to describe assets and facilitate their cataloging
- Databases of harmonized data from various sources (e.g., WorldCereal, AgroSTAC) in support of EAVs

[2] Comprehensive Knowledge Archive Network (CKAN) is an open-source open data portal
[3] SpatioTemporal Asset Catalogs (https://stacspec.org/en)

This ideal architecture was estimated to be feasible and coherent, with the need to assess how to efficiently bind the different components and test their robustness. The need to work with open-source and free tools as much as possible in the spirit of GEO was also highlighted during the workshop. This would lead the way for other activities of the GEO Work Programme to advance on the topic of in-situ data sharing and enhance the GEOSS Infrastructure.

There is a clear need to address trust and transparency issues by raising awareness on the GEOGLAM work and why we are tackling this issue of in-situ data sharing. Ranking EAVs based on their sensitivity, starting with the least sensitive, could be a way to kick start the activity. Through capacity development targeting priority countries on data management practices, trust issues could be reduced. This could lead to institutions being able to manage the data, leveraging CKAN and allowing GEOGLAM to link the data through the CKAN Harvester and the GEO Discovery and Access Broker (GEO DAB).

To strengthen national level coordination, there is a need for more engagement with National GEO offices, to link directly with agriculture ministries or any other relevant agencies responsible for managing the in-situ data related to agriculture. As part of future efforts to incentivize data sharing, an interesting approach would be to provide support (e.g., financial reward) to share data at different scales (e.g., country, region, local and smallholder farmers). Engaging with key international organizations such as the WFP or the United Nations Food and Agriculture Organization (FAO) would be key as they are already working at the country level.

As part of the AFOLU roadmap, stronger ties with CEOS and individual Spaces Agencies are also important to ensure the captured data helps validate satellite data. There is a strong need to approach funders either from governments, international organizations or private donors that could have an interest in supporting in-situ coordination activities, aside from data storage. For this we need to make the added value of GEOGLAM clear.

Finally, we need to partner more with organizations that are active in citizen science programs, as well as with universities and youth communities, to develop innovative ways of collecting data and make this a common global effort where everybody can contribute to GEOGLAM.

5.2 Actions required to address data structure and maintenance infrastructure

The following actions were defined to host and advance open data sharing:

- **Action DSMI.1**: Perform a data inventory and assess the level of data openness in countries and organizations, including data infrastructure capabilities.
- **Action DSMI.2**: Implement a demo distributed architecture with hosting capability using the open source CKAN solution under the neutral GEO/GEOGLAM umbrella,

with an implementation of STAC, and common API to query meta data and the data itself (i.e. using WorldCereal in-situ data as a demonstrator).

- **Action DSMI.3**: Help develop local/national/regional CKAN repositories where necessary and link them with the GEOGLAM one, using the CKAN Harvester. In the federated spirit, assess how to leverage the GEO DAB and to link other catalogs.

Appendix A: GEOGLAM In-situ Data Workshop Agenda

GEOGLAM In-situ Coordination Workshop

Draft Agenda V Final

Geneva, WMO building Room C1 (-1 level), November 21-22 2022

Supporting Documents. Agenda; Draft In-situ Guidance and Workplan: https://www.dropbox.com/sh/x46gk90zowfq18u/AAB2t1TcnI1Y5auBnsDZtmVaa?dl=0

The first morning will consist of scene setting presentations to introduce the in-situ data life cycle, followed by updates on progress made to date. The remainder of the meeting will consist of breakout discussions and report back to inform consensus around the next steps.

To promote discussion and consensus building the meeting is primarily intended to be face-to-face. Notwithstanding this, the first session of scene setting presentations, and the final report back and discussion session will be streamed and recorded for those that cannot make it in person.

Monday November 21

9:00am - 1:00pm

1. Roundtable Introductions (30 min)

2. Introduction to GEOGLAM In-situ Coordination and the Data Lifecycle (DLC). Overview of the guiding principles for the in-situ initiative; and going the "extra mile" to develop and curate harmonized data sets that address the requirements for generating Essential Agricultural Variables (EAVs – AgVariables.org). It will also provide an overview of the DLC approach and the draft GEOGLAM Guidance and Workplan (circulated in advance of the meeting). Ian – 25 min.

3. Highlight progress made to date on addressing DLC components. Present Progress made by:
    a. **Best Practice Review**. (Hendrik and Sander, 30 min)
    b. **Data needs and Users**. GEOGLAM Essential Agriculture Variables as a driver for in-situ data requirements. (Sven – 10 min)

c. **Data Development/Acquisition.** Best practices for data collection and management (IIASA CropObserve, others TBD) (30 min)
d. **Data Usability (quality and policies).** Explore what processes are required to assess data quality, also discuss how we as a community can review and endorse in-situ data sets. (Sander + open discussion 15 min)
e. **Legacy Data.** WorldCereal in the development of a Global reference database, and data quality processes for crop mapping-Leveraging WorldCereals. Hendrik – 10 min

*__Approx 11:00 Coffee Break 30 min__

f. **Data structure and maintenance infrastructure.** Present a proposal that provides a unified approach to architecture: (GEOGLAM Architecture Working Group, 50 min)
    - Overview of GEOGLAM high-level architecture
    - AGROSTAC API/Platform (Wageningen)
    - STAC (Radiant Earth)
    - CKAN (UMD/Harvest)

4. Additional Discussion and Set up of the Afternoon Breakouts

2:00pm-5:00pm

5. **Breakout discussions.**

Note: Face to face discussions only, virtual presence not possible. Virtual attendees can submit their written input to the topics/questions by November 28th and they will be integrated with the results of the breakout groups.

Format: We will divide the attendees into 4 breakout groups. Breakout will have rotating discussions using a "World Cafe" approach:

- There will be four stations, one for each topic related to the key gaps to implementing a sustainable in-situ data lifecycle for agriculture.
- Each station will have a rapporteur and someone from the organization committee as a resource.

- Breakout groups will rotate through the topic stations.
- Round #1 breakouts will be given 50 min, then rotate to the next station.
- After round 1 each round will be 30 min and breakouts are directed to check/validate what was said by previous groups, then provide additional ideas/input. And so on.

Topics Stations:

1. *Data needs and users*
2. *Data development/acquisition*
3. *Data usability including quality and policies*
4. *Data structure and maintenance infrastructure*

Questions to be addressed at each station by the breakout groups:

a. *Define the gap (what is missing)*
b. *High level action (s) to address*
c. *Identify interested contributors (sub-working group)*
d. *Business models and resources required (time, funding)*

Tuesday November 22

9:00am – 12:00

6. Plenary feedback working groups, discussion, and conclusions (Rapporteur/Organizer for each topic)

   Open points for discussion

   Conclusions

12:00pm - 1:00pm

7. Summary of findings and actions/next steps and wrap-up

Appendix B: Overview of Actions

Action	Status
Action UN.1, Identify priority policies and corresponding EAV needs	Complete
Action UN.2, Document EAV Requirements	On-going
Action UN.3, Gaps assessment	Planned - not yet resourced
Action UN.4, EAV stewardship	In process
Action DA.1, Communication to funding organizations	In process
Action DA.2, Data acquisition best practices	In process, via workshops e.g. CEOS LPV/GEOGLAM
Action DA.3, The case for open data	Planned - not yet resourced
Action DA.4, In-situ sharing best practices	In process, via workshops e.g. CEOS LPV/GEOGLAM
Action DA.5, In-situ registry of field campaigns	On-going
Action DU.1: Institutionalize, sustain, and update the WorldCereal reference data repository	Exploring GEO platform
Action DU.2: Encourage the community to publish land cover and crop type data	In process, via workshops
Action DU.3: Explore business models to strengthen and sustain harmonization hubs	Planned - not yet resourced
Action DU.4: Further develop methods and protocols to review and harmonize open and published data	Planned - not yet resourced
Action DSMI.1: Perform a data inventory and assess the level of data openness in countries and organizations	Planned - not yet resourced
Action DSMI.2: Implement a demo distributed architecture with hosting capability	Planned - not yet resourced
Action DSMI.3: Include local/national/regional CKAN repositories where necessary	Planned - not yet resourced