CGIAR Platform for Big Data in Agriculture

IDM CoP

AGROVOC

Food and Agriculture Organization of the United Nations

Increasing Interoperability Between Food and Agricultural Systems: CGIAR & FAO Collaboration

Task Group and Curation Team Report

November 2021

*"If agriculture is going to take advantage of the digital revolution, it's critical that data resources can talk to each other through ontologies and standards – this is a necessary foundation before anything can meaningfully progress. Bringing CGIAR and FAO together ensures two major players in agricultural research and development are promoting common standards and approaches, for ourselves and for our large partner networks" – **Andy Jarvis, DDG Research Alliance Bioversity-CIAT***

*"FAO acts as a catalyst and a platform for leveling the playing field so that countries can make evidence-based decisions on the most appropriate technologies and innovations to adopt and adapt in sustaining their food security and nutrition. FAO can only achieve this by collaborating with the global scientific community of experts such as the CGIAR" - **Ismahane Elouafi, Chief Scientist at FAO***

Authors:

Task Group Co-chairs: Olatunbosun Obileye, IITA, Elizabeth Arnaud, Alliance Bioversity-CIAT,

Members: Enrico Bonaiuti (ICARDA),, Sara Jani (ICARDA), Soonho Kim, (IFPRI), Erica Saito (IFPRI), Marie-Angélique Laporte (Alliance Bioversity-CIAT), Jacqueline Muliro (WorldFish), Abenet Yabowork (ILRI), Imma Subirats (FAO), Kristin Kolshus (FAO).

Guests: Andea Turbati (FAO), Alan Orth (ILRI)

Team for the curation workflow test chaired by Sara Jani, ICARDA.

Center	Team members	Focal Points
Alliance Bioversity-CIAT	Maria Garruccio	
CIMMYT	Araceli Zúñiga for Spanish translations	Jesus Herrera de la Cruz
ICARDA	Sara Jani, Asma Jeitani	
IFPRI	Erica Saito	
ILRI	Abenet Yabowork	
IITA	Hafeez Adepoju	
WordFish	Saadiah Ghazali	Jacqueline Muliro

Table of contents

Recommendations of the Task Group

1. One CGIAR needs to strengthen its contribution to AGROVOC thus supporting the consolidation of the semantic landscape for labeling data in agriculture and food systems.

2. CGIAR centers should wait a bit till the affiliation process is complete so that the appropriate unit that will be responsible for AGROVOC can consume the Agreement since the timeline for the affiliation process is just some few months away.

3. The reality check performed for a selection of around 50 CGIAR keywords with no match in AGROVOC led to integration of missing terms into AGROVOC and a sub-concept schema for ONECGIAR was created to provide direct visibility to the set of concepts (https://agrovoc.fao.org/skosmosOneCGIAR/cgiar/en/). Based on the collaboration concrete results, The TG recommends that the term submission effort and collaboration with FAO continues with proper allocation of data managers' time and a training plan. Contribution to AGROVOC should be part of the data managers ToRs to concrete provide recognition of this role.

Results of the curation team for the test of the term submission workflow

A One CGIAR sub-schema was created in AGROVOC to provide an online visualization of the CGIAR terms that are integrated. The online schema will be populated with 36 fully defined CGIAR terms on the 2nd December 2021: https://agrovoc.fao.org/skosmosOneCGIAR/cgiar/en/

I. Introduction

To be Findable, Accessible, Interoperable, Retrievable (FAIR), data/metadata must be labelled with controlled vocabularies that are widely accepted by the community of the science domain, instead of creating specific project vocabularies that would lead the data sets to be siloed and hidden to the global community. Data/metadata labelled with such community-validated vocabularies can thus be interlinked, retrievable through a keyword-based search and produce a relevant results list.

The AGROVOC thesaurus is a very large-controlled vocabulary maintained by FAO. It offers a structured collection of agricultural concepts, terms, definitions and relationships which are used to unambiguously identify resources. It was established in 1980s as a multilingual structured thesaurus in all fields relating to the agriculture, forestry, fisheries, food, and other related fields such as environment. It consists of around 39,000 concepts with nearly 826,000 terms in up to 40 languages. AGROVOC is much used for adding keywords in the metadata of bibliographic references and data sets. 22 institutions are contributing to AGROVOC. It is the largest open thesaurus in agriculture sector which is open for everyone, and its impact is through providing the access and visibility of data across domains and languages. AGROVOC uses semantic web technologies, linking to other multilingual knowledge organization systems and building bridges between datasets.

The CGIAR data and publications are published in diverse repositories (see section 2) being mainly described with the CGIAR metadata core schema that recommends using keywords selected from the AGROVOC thesaurus. Therefore, all centers use AGROVOC as a source of keywords. ICARDA provides to FAO and partners a specific translation service from English into Arabic.

The Ontology Working Group of the Information and Data Management Community of Practice (IDM CoP) held meetings with FAO where a need was identified for a **defined and consistent process to submit the CGIAR terms.** Such a defined workflow will stimulate the submission of missing terms, improve the quality of terms submitted and used, and provide to CGIAR a visibility as contributor, acknowledging however that the task is staff time consuming.

The Task Group aims to define the importance of AGROVOC for CGIAR data publishing, the added value for the CGIAR to formally contribute to AGROVOC**,** and how to organize CGIAR contribution in a coherent workflow. The Task Group has now gained a better vision of AGROVOC objectives, its management with the editorial board and its curation tools. A formal collaboration could be materialized via a formal agreement with FAO, a participation in AGROVOC editorial community, and co-organized events.

II. The current usage of AGROVOC by CGIAR and needs for improvement

1. CGIAR Survey results

A survey of the Information and Data Management CoP was conducted in July 2021 to get the picture of the current use of AGROVOC by CGIAR. A total of 28 respondents answered the survey questions. Several answers were received per Center as the allocation of keywords is a team effort. All resulting diagrams are in Annex.

a) AGROVOC as the main source of keywords for all centers

100% of CGIAR Centers use AGROVOC as a source of keywords in various repositories (figure 1). *Note*: the only negative answer regarding the use of AGROVOC was from a newly appointed colleague from AfricaRice who indicated having no knowledge of AGROVOC. However, the AfricaRice Dataverse repository shows that AGROVOC is used as a source of keywords.

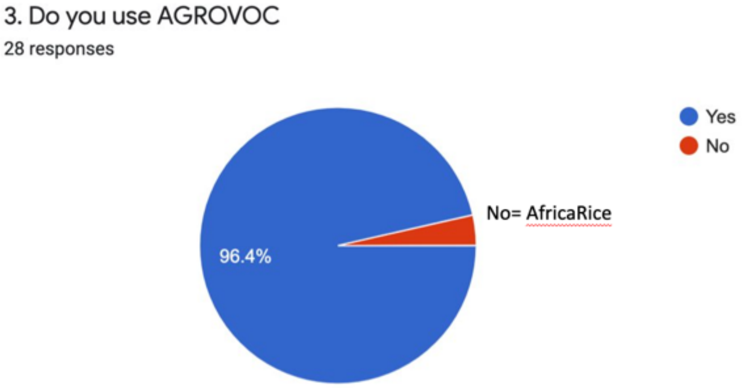


Figure 1:Response to question3 of the Task Group's survey

AGROVOC provides a very large source of concepts and controlled vocabulary aligned with the One CGIAR Core Metadata Schema.

AGROVOC is a resource to annotate metadata and harmonize the subject field (= keyword field). However, AGROVOC is not adequate to annotate data inside dataset. This annotation is key to leverage the full potential of datasets. This is the reason for expanding the collaboration to the ontologies developed by CGIAR.

Providing CGIAR research concepts to AGROVOC requires to standardize keyword allocation and submission processes. The Management of One CGIAR needs to understand the importance of implementing AGROVOC in our projects and publications, and that CGIAR needs to play a key role in the creation and maintenance of the semantics used for Agrifood research data and publications.

b) Important current contribution of some CGIAR centers to AGROVOC

Four CGIAR Centers (ICARDA, IFPRI, CIFOR and WorldFish) submit terms to AGROVOC and the contribution differs across centers.

It is worth mentioning that ICARDA is one of the major CGIAR contributor as the data management staff provides the Arabic translation of AGROVOC, aside the submission of ICARDA terms and term submission on the behalf of some centers through MEL. The translation into Arabic is a long-lasting activity performed in agreement with the global Memorandum of Understanding (MoU) signed between FAO and ICARDA. This is a service valued by FAO and by CGIAR partners in the Arabic speaking countries. ICARDA is a member of the AGROVOC Editorial Board.

WorldFish started to submit terms in December 2020 and is the most recent contributor in CGIAR.

2. Improving the contribution of CGIAR with a workflow and training of data managers

61.9% of the respondents that do not submit terms to AGROVOC indicated that they are not aware that they can send new words to AGROVOC while 28.6% indicated that they do not know who or where to send the new terms. 9.5% never heard about the FAO tools for term submission, e.g. Vocbench or the submission template.

Survey respondents requested training in how to use both the template and Vocbench, the online curation tool. CGIAR needs to define the training needs in such a way that FAO can allocate proper resources and staff time. The group of volunteers to prepare the 1st batch of concepts for submission forms the current target group for a focused training workshop in the use of AGROVOC curation tools and vocabulary structure

CGIAR will be a new use case for AGROVOC. FAO does not train trainers, but an option would be to develop micro-ionline courses (MOOCs) based on new editorial book. Part of the training would be based on the AGROVOC Editorial Guidelines 2020 (will be updated late 2021), http://www.fao.org/3/cb2328en/cb2328en.pdf

Proper training will support the engagement of CGIAR staff members to cover this curator role. OneCGIAR should create a process to train people when they start to work for CGIAR center.

3. CGIAR Repositories and Monitoring & Reporting systems enabling the use of AGROVOC

Dataverse	Dataverse allows user to insert aside a keyword coming from a controlled vocabulary (AGROVOC or an ontology) its hyperlink called A **Uniform Resource Identifier** (**URI**) so the term can easily be found and extracted by search engines and also be displayed in its source vocabulary. The URI allow to distinguish between a term selected in an existing semantic source from a newly created and local keyword. Dataverse does not implement lookups for existing standards, so users have to manually input both the keywords and the URIs. Therefore, most centers are now assigning the URIs to their keywords when preparing the data set upload template.
CKAN	allows the linking of keywords with URI. This could be either importation of the keyword, in this case, AGROVOC, or controlled vocabulary into its schema as well as manual entry of the keywords for each URI. To achieve this, the vocabulary needs to be presented in xml format. Currently, AGROVOC controlled vocabulary is in rdf which is not compatible with the current version of CKAN installed at IITA. The keywords are inserted manually. However, the system stores any new keywords and automatically provide auto complete when the same keyword is to be used in subsequent URI's. IITA is the CGIAR center using CKAN
DSPACE (CGSpace and other DSPACE installations in centers)	CGSpace is a repository of agricultural research outputs produced by different CGIAR centers, the CGIAR system management office, research programs and partners. It indexes reports, books, articles, presentations, videos, policy briefs, etc. Submitters use AGROVOC keywords while indexing research outputs in CGSpace. The most common AGROVOC keywords are compiled as controlled lists to select from for users making new entries. Users can also add new AGROVOC keywords by referring from the online AGROVOC thesaurus: https://agrovoc.fao.org/browse/agrovoc/en/. Submitters also use center specific keywords (might not be subset of AGROVOC) when they index research outputs.
Content DM	IFPRI's Institutional Repository (IR) is built on ContentDM from OCLC. The metadata (in qualified Dublin Core) and PDF are used to populate the Drupal-based primary corporate website, www.IFPRI.org. The IR record also populates WordPress-based project subsites. The metadata is also used to populate several outside repositories and portals such as Google Books, Repack, and SSRN. In partnership with other CG Centers, IFPRI has undertaken a process to align metadata across all repositories, referred to as CGCore. IFPRI incorporates a combination of controlled vocabularies, AGROVOC and CAB Thesaurus, as the primary subject keywords, and then supplements with Journal of Economic Literature codes (JEL) depending on whether the content is appropriate. Currently, three librarians that catalog in CONTENTdm are responsible for assigning keywords from AGROVOC and CAB in the repository records. In case a keyword is not covered by AGROVOC, we contact FAO and suggest inclusion of a new term. So far, all the terms sent to FAO were approved and included in AGROVOC.
Monitoring, Evaluation and Learning (MEL)	is an online Management Information System (MIS) to plan, manage, monitor, evaluate, report and share its activities and results. It is the first of its kind that included AGROVOC in its workflows and description of different level of information. An important component of MEL use is the storage of data and documents while providing exhaustive metadata for each item. The MEL platform introduced the use of AGROVOC to describe user profiles, blogs, outcome stories and capacity development events. The use of keywords plays a relevant role in the metadata on MEL, and it is a required component of the structure of our dataset. The innovative approach is not only to provide AGROVOC keywords in MEL or any other platform

	but to have a real-time connection with AGROVOC since it is updated monthly resulting in real-time update on external repositories (e.g., DSPACE, DATAVERSE) and increased interoperability for broader utilization of data across organizations in accordance with FAIR principles (https://www.openaire.eu/how-to-make-your-data-fair). Additionally, the MEL platform automatically reviews all keywords not previously recognized by AGROVOC and matches them with the new keywords provided by AGROVOC.

4. Linking ontologies developed by CGIAR

CGIAR has been a key player in the development of ontologies in agriculture for more than ten years, starting with the Crop Ontology project. The ontologies are used in databases and by field book apps to disambiguate data and provide clear meaning on variables and parameters recorded. The ontologies rely on well-established communities, that continuously provide new content and improve existing one in relation to new use cases. Although thesauri and ontologies have different goals at first when it comes to data management, they interplay in the context of the semantic web thanks to the semantic web standards. In the context of the collaboration FAO/CGIAR, the scope of AGROVOC in terms of data annotation in the metadata should be defined. The alignment of AGROVOC and CGIAR ontologies is a first step in that direction in the sense that if AGROVOC and ontologies are mapped and are seen are one big structure, terms from both sources can be used interchangeably.

A first draft of mapping AGROVOC to the Agronomy Ontology has been released in 2020, enriching AGROVOC with new terms coming from AgrO and providing a curated source for some term definitions in AgrO. Both standards now provide references to each other, raising visibility and awareness of each other in their respective communities.

III. Objectives of developing a collaboration framework with ONECGIAR perspective

The objectives of developing a framework for the collaboration between CGIAR and the FAO is to:

a.	Get more streamlined submission to AGROVOC from CGIAR.
b.	Improve AGROVOC content cover with terms from the CGIAR science domain and thus support CGIAR data labeling
c.	Support the indexation publications and data sets by GARDIAN and AGRIS
d.	Improve the keyword-based search returning both publications and data sets in repositories, in MEL, and GARDIAN and AGRIS
e.	Provide global visibility of CGIAR contribution to AGROVOC and of our stewardship role in the development of semantics for Agriculture.
f.	Remain a key player in the international community effort to develop semantics
g.	Enhance our stewardship role in the development of ontologies for agrifoods.
h.	Develop capacity building for partners for submitting terms and empower universities to do so for CGIAR science domain.
i.	Leverage a team of domain experts from the AGROVOC editorial community in terms of definitions, etc.
j.	Strengthen the AGROVOC coverage by adding necessary terms to AGROVOC. Reduce the duplicated efforts/costs to maintain domain terms for both FAO and in CG centers.
k.	Improve efficient use of our knowledge for internal and external users.

It will be necessary to include this work under the structure of One CGIAR. We learned that a Director of Data Management and Analytics position will be opened under the supervision of the Global Director for Digital Services and certainly this new set of services could include an AGROVOC-CGIAR collaboration. Data managers will have new reporting lines so communication action about the continuity of the collaboration will be important.

The CGIAR covers a vast area of knowledge that is relevant to AGROVOC and should be an active stakeholder. Not all CGIAR terms are included so there is room to collaborate more with AGROVOC to get our vocabularies synchronized within One CGIAR. This would mean one step closer to having interoperable data. This is for the betterment of the science community more than anything.

IV. An Agreement with FAO: marking CGIAR contribution as official and visible to high level management

ICARDA is the only CGIAR Center to have signed an MOU with FAO that includes the contribution to AGROVOC. One CGIAR should look at the broad framework and define what are the areas of interest for contribution that will benefit of the global community.

The suggestion made by the Task Group is that the process should avoid too many formalities as each center/team can submit terms to AGROVOC like any other institution. It is however recognized that in the perspective of securing an annual funding to cover few positions that include in their description the collaboration with AGROVOC a more formal collaboration framework can help.

A simple **Implementation Agreement** is possible. Imma Subirats, FAO, can look for examples of FAO implementation agreements example, much lighter than MoU. We should assess the value of the Implementation Agreement and keep the administration part light. We could add a general text in the agreement that covers the AGROVOC activities.

- **Option 1:** have a consolidated agreement emanating from One CGIAR to showcase the emerging entity, since, by January 2022, names of Centers may be changed to One CGIAR.
- **Option2:** each Center could have a short-term agreement or a broad agreement that could be adopted by One CGIAR soon with FAO/AGROVOC. It means identifying a contact person per CGIAR Centre who will follow up the agreement process.

The TG should take into consideration steps or short-term actions (e.g. training) that will facilitate the implementation agreement.

1. Value added of a formal collaboration and mutual benefits.

There will be mutual benefits for CGIAR and FAO. Several CGIAR research domains are partially represented in AGROVOC so the thesaurus can get new terms from the research ground and benefit from our research on domains involving use of vocabulary and ontologies. FAO can benefit from the linking to our ontologies which are more focused and maintained by a specific Community of Practice.

CGIAR is an important user of AGROVOC, which is a semantic resource fully maintained by FAO that develop necessary accompanying tools and provide training to users. CGIAR collaboration is not limited to the submission of terms. ICARDA provides a translation serviced of the AGROVOC terms in Arabic which opens the content for use by partners in Arabic-speaking countries. All keywords are submitted on the behalf of partners participating in the research project. CGIAR can participate to the training and capacity building of partners to contribute and use AGROVOC.

Merging CGIAR domain specific terms with AGROVOC will provide better research coverage thus supporting the indexing and labeling of research data and publication. Such a collaboration will reduce the maintenance costs of domain specific terms. The integration of our domain-specific terms will improve the interoperability of CGIAR metadata, across centers and domains, and with partners like FAO.

*"Much of the world's research is not published in English and may not be easily discoverable. … Adding new translated terms is a contribution to making research and datasets more discoverable, and to making national research more visible and accessible." ("*AGROVOC: Semantic data interoperability on food and agriculture*" FAO, 2020)*

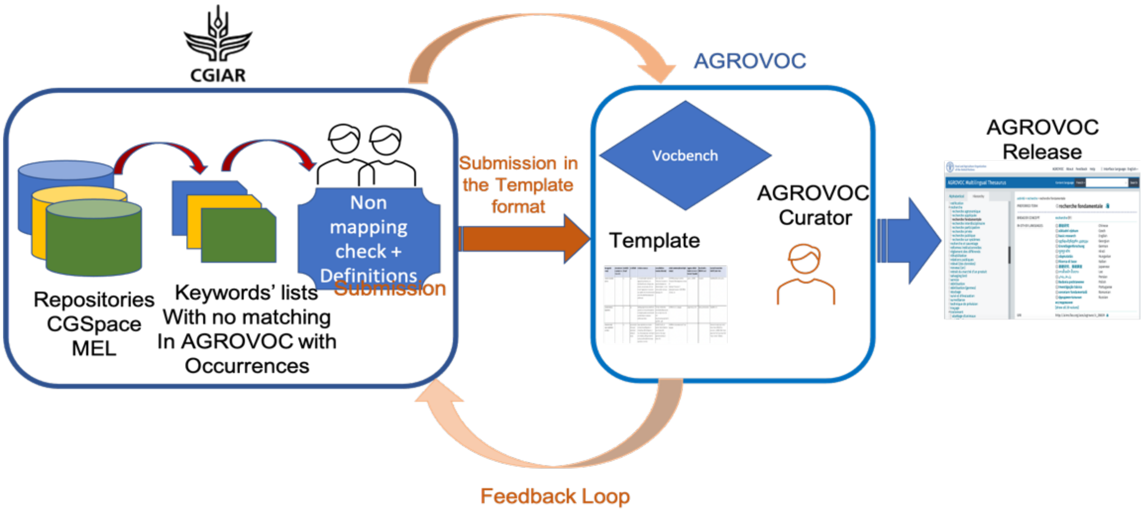
2. How CGIAR should increase its visibility as an official contributor?

Staff in many Centers believe that it is a good idea for CGIAR to collaborate with FAO in the submission of new terms to AGROVOC. Initially, two options of agreements were raised. These include center-level agreement in which the sample agreement template used for ICARDA with less admin could be adopted or a One CGIAR agreement which will be set by the new leadership in some few months. The existing collaboration of FAO with ICARDA still subsist wherein ICARDA is working at translating AGROVOC terms to Arabic.

However, FAO suggested sharing its Implementation Agreement template instead of generating Memorandum of Understanding per center. Although the Task Group members suggested the recognition of individual contributors, FAO's team explained that it only gives credit to organizations and not individuals. FAO did not indicate the willingness to add CGIAR as editorial board on the immediate. AGROVOC editorial community is very pragmatic and give visibility to institutions rather than to individuals. FAO is currently working at providing better visibility to contributing institutions with one page per contributor. Should the visibility be given to One CGIAR?

V. Identify Key Elements for the Workflow and Map Resources Needed

To improve CGIAR contribution and use of AGROVOC, there is a need **to define a clear term submission workflow within CGIAR** and enable the submission of quality concepts. An ideal solution mentioned was a possible integration of AGROVOC tools into the CGIAR repositories data management flow. A training campaign of data managers in the use of AGROVOC curation tools and term categorization is needed. Accessing focused domain within AGROVOC will ease the understanding of the thesaurus structure and improve the submission of concepts within the semantic hierarchy. The workflow that was discussed is represented in the hereunder schema:


CGIAR
AGROVOC
Non mapping check + Definitions
Submission
Repositories CGSpace MEL
Keywords' lists With no matching In AGROVOC with Occurrences
Submission in the Template format
Vocbench
AGROVOC Curator
Template
AGROVOC Release
Feedback Loop

To help validating the steps of the workflow, the Task Group proposed to run a **reality check by forming a curation team of 8 volunteers across centers** to select keywords the most used in repositories, appearing not being included in AGROVOC, and participate to the submission to the thesaurus. The TG also mentioned that there is a need to show concrete results in short term while we are in this period of One CGIAR taking shape. This work will enable assessing the effort needed to submission quality concepts.

Steps taken were:

Production of the keyword lists

a. ICARDA produced a list of all CGIAR own keywords with their occurrences, that do not have an exact or fuzzy match with an AGROVOC term. Keywords were extracted from CGSpace, MEL and centers repositories. IITA added keywords from CKAN and WorldFish from its Dataverse.
b. Out of this list, a short list of priority keywords were selected using criteria set by the TG.
c. Colleagues volunteered to take subsamples of this list and prepare it into the AGROVOC format using the Template.

Training to the curation team

a. Training in the production of quality definitions and in the use of the Term submission Template was provided by Kristin Kolshus, FAO Agrovoc curator.

b. Terms not matching any concept in AGROVOC were short listed for adding a quality definition and then were submitted.

A ONECGIAR Concept Schema created

a. The November release of AGROVOC integrated a dozen of CGIAR Keywords

b. A **Sub-Concept Schema for ONECGIAR** was created to provide a direct visibility of the CGIAR concepts in AGROVOC-
One CGIAR Skosmos URL is https://agrovoc.fao.org/skosmosOneCGIAR/cgiar/en/ and will populated after 12 December with the CGIAR concepts from the curation team, adding specific terms submitted by WorldFish, IFPRI and IMWI.

VI. Risks

A major risk identified is the uncertain of continuity due to the current One CGIAR affiliation exercise. Members of the Task Group, as well as respondents to the survey do not know the structure and the persons that will be responsible for AGROVOC in the new structure. So, the current plans may not take care of the future in One CGIAR

VII. Recommendations

4. One CGIAR needs to strengthen its contribution to AGROVOC thus supporting the consolidation of the semantic landscape for labeling data in agriculture and food systems.
5. CGIAR centers should wait a bit till the affiliation process is complete so that the appropriate unit that will be responsible for AGROVOC can consume the Agreement since the timeline for the affiliation process is just some few months away.
6. The reality check performed for a selection of around 50 CGIAR keywords with no match in AGROVOC led to integration of missing terms into AGROVOC and a sub-concept schema for ONECGIAR was created to provide direct visibility to the set of concepts (https://agrovoc.fao.org/skosmosOneCGIAR/cgiar/en/). Based on the collaboration concrete results, The TG recommends that the term submission effort and collaboration with FAO continues with proper allocation of data managers' time and a training plan. Contribution to AGROVOC should be part of the data managers ToRs to concrete provide recognition of this role.