Ontology Community of Practice sessions

During the open session, participants shared their expertise, objectives of joining the CoP and their needs. **Participants’ Objectives** are to understand (i) learn how to deploy ontologies in ICT projects (linking ontologies to the information systems or data portal they develop and how to develop ontology driven databases), (ii) how ontologies can support monitoring and reporting systems like MARLO that uses Dataverse metadata to describe deliverables, (iii) how ontologies can help with the Data management plans and Open Science Policy (FAIR data), (iv) how to improve the fielbooks for data collect by using ontologies, (v) use crop ontologies for linking characterization and evaluation traits to accessions.

There is no agreed set of standards for the data management across CGIAR. Participants seek help to organize the CGIAR data and ontologies appear like the necessary evil to achieve that. One long-term objective that was suggested would be that ontologies are not needed anymore but we did not discussed how they would be replaced. Participants need to access a consistent set of ontologies that are standard, accepted by the Agronomy domain, that can be combined or aligned, and to which they can contribute. Ontologies should support search functions of MARLO and automatic solutions should be developed for validation of vocabularies used in the repositories (aside CoP expert validation). The actions of the CoP should provide help with the ‘I’ of FAIR (Interoperability).

Participants also expressed specific needs like finding examples of linking the traits to genebank data (e.g. for Genesis portal); of using automatic, machine-to-machine, process; identifying standards for managing the Crop Wild Relatives species data.

<table>
<thead>
<tr>
<th>Group 1:</th>
<th>Group 2:</th>
<th>Group 3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bettina Heider (CIP, sweet potato germplasm evaluation), Hector Tobon CCFAS (ICT), David Abreu (CCFAS, ICT), Nora Castañeda (CROP TRUST, germplasm evaluation), Julian Sanchez (Univ Cauca, ICT)</td>
<td>Steven Sotelo (CIAT), Kevin Silverstein (U Minesota), Vicky Arciniegas (GBIF, SiB Colombia), Adriana Sanchez (independent consultant), Kostas Kastrantas (Agroknow)</td>
<td>Kelly Robins (Cornell Univ.), Steve Kemp (ILRI), Karen Smyth, Brian Lowe (Ontocale), Leroy Mwanzia (CIAT)</td>
</tr>
</tbody>
</table>
Group 2: Steven Sotelo (CIAT), Kevin Silverstein (U Minesota), Vicky Arciniegas (GBIF, SiB Colombia), Adriana Sanchez (independent consultant), Kostas Kastrantas (Agroknow)

Plenary discussions

1. Introduction to the objectives, workplan and composition of the Ontology CoP - Elizabeth Arnaud, Bioversity International, CoP Guru

The Ontology CoP Objectives and workplan were presented to the full audience and commented (Objectives: see website (http://bigdata.cgiar.org/ontology/).

We need Ontologies to have more focused return on queries. An agreed role of the Ontology is to enable FAIR data. The most important aspect is to achieve interoperability of the multi-disciplinary data. CoP needs to embrace from data collection to storage to analysis to access on semantic web.

Comments from the audience:

CoP role will be key in proposing a set of quality ontologies and validating metadata to improve their quality. We need to check if ontologies available in the relevant knowledge domains are sufficient for what we need. It is required to identify reference ontologies for each knowledge domain / each data type used in Agriculture. Which one are useful, where are the gaps and how can the community fill those? What about standards that have already been implemented in other areas? Will we be willing to use those? We need to reuse what exists. Correlations are already created between Ontologies in project like Planteome. We could all help identifying pieces of ontologies that could be useful. We don’t have all the technical information of all ontologies. AGROVOC is important for our repositories and still currently maintained by FAO: we need FAO as a partner in the CoP. The CoP should collaborate with FAO but also connect with the other ontologies that we really need. Discuss possible connections with the Global Agricultural Concept Scheme (GACS) that compile terms from AGROVOC, CABI, NAL and with other ontology groups.

What are the potential criteria to select an Ontology? What makes an Ontology the best standard? What are the standards by sub-domain? We should have list of criteria to help recommending Ontologies for various data types with a quality check. Could we conceive a look up service by terms and the app looks in ontologies for the terms needed? There are so many ontologies in each domain that it’s difficult to choose which one to use. For specific domains, the CoP has to decide which ones we want to use. This supposes to have an ontology switchboard linked to repositories. The ecosystems of Ontologies the CoP will recommend has to be consistent so we need to identify what in these Ontologies would make a long-term reference.

Agronomy, environment and socio-economic data is the focus of the big data platform. There are 2 other CGIAR platforms that take care of germplasm (Genebank
platform) and of crop, pre-breeding, genomics (Excellence in breeding platform). They both rely on the Ontology CoP in Big Data to provide missing ontologies they need.

**Add clearly in the objectives**, the necessary collaboration with Excellence in breeding platform and genebank platform.

**A priority gap to fill in is standards for household surveys and socio-economic data.** The socio-economic data CoP also needs our input. See page 7 the report of the session with the SED CoP,

**Boundaries of the Ontology CoP:** Will the CoP provide the means to share the data, or will it only work only on Ontology? The CoP should support ‘FAIRization’ of data. We need a hub that can be used as a source of concepts. First, the Ontology provides the data model, the domain model that enables accessing data annotated with the ontological terms and also to perform text and data mining. and then you have the data. This CoP will work more on knowledge modeling.

**Will the CoP share the technical info on how to use the Ontologies ?**

**Will the CoP work to get the ontologies into use?** Show example where ontology projects have achieved adoption by key users. Prove of interest by users, CGIAR partners. The use of controlled vocabularies: Users like MARLO could help validate the data annotated with the ontologies and propose new terms.

**How to make FAIR ontologies FAIR ?** Type a trait and see matching traits... we support definition of what should be best annotation tool.

**The CoP should provide a look up service for the whole Big Data Platform.** (e.g. Ontology look up service of the EBI)

**Organization and communication:** People can sign up from website, so we don’t know depth of involvement desired. We should have a light CoP advisory group, like 6 people? From other CoPs, from external partners. We don’t have money for the CoP as such, it’s part of module 1. The advisory group should be reference group for validating.

We should envisage the possibility to create working groups per knowledge domain, par technical issues, etc. This CoP connects with existing Crop ontology working groups like the CGIAR Ontology Working group, the DivSeek ontology Working group, IGAD-RDA, etc where we need to secure a representation keeping focus on the CoP objectives.

**2. The results of the Ontology CoP Survey that occurred from May to July 2017,** with already subscribed cop members were presented. The survey received 43 answers upon 84 members. At that time, the CoP was missing representatives of ICARDA, IMWI, ILRI. After the convention, we have contacts for ILRI and IMWI. See presentation at: *add link*
3. Lightning talks

**Medha Devare**, IFPRI, Module 1 leader – *Collecting agronomic data with an ontology-driven fieldbook* - Medha Devare presented AgroFIMS, the ontology-driven field book creator. It allows selecting parameters and creating a fieldbook with nested entries, download it for use and upload it back when filled in. It is based on the AgrO Ontology, which is a composed by terms of existing ontologies and new terms. For example, when you will choose a crop or crops, crop related traits will come from Crop Ontology. Users can create and download an Excel fieldbook to be shared across sites and collect standardized data right from collection point. Each parameter is identified with a unique identifier from AgrO. Download fieldbook and when you upload back it goes to a database. Value addition: standardization of data and one-click upload to the institutional repository. Test of AgroFIMS will perhaps starts with G-funded agronomy projects. Adoption will be stimulated through road shows, demos, and conferences, visits at CGIAR centers: we aim to have something to be tested by the end of 2017.

Q: What will be the connection with phenotype datasets or genotype information?

For example, there is a standard for the minimum information for phenotypic data (MIAPPE) that includes the observation variable format from Crop Ontologies and this could be a way to link to the Agronomy fieldbook.

Q: How is the tool dealing with duplications in Ontologies? The Ontology underpinning the fieldbook database is already a mash-up of existing ontologies plus new agronomic terms that are missing so there will be no duplication possible.

**Kevin Silverstein**, Operation Manager, International Agroinformatics Alliance – *An ontology mash-up framework for agriculture* - G.E.M.S ontologies for a Platform on Genetics by environment by management by socio-economics data – To aggregate data, we need a Dynamic metadata mashup model. We will have to solve the compliance problem of legacy data.

Comment: GEMS needs community collaboration and further discussion in a next event, for example PhenoharmonIS 2018.

**Steve Kemp**, Programme leader for animal biosciences, ILRI – *Managing livestock language - ontologies, vocabularies or useful anarchy?* - ILRI Biorepository is the long term storage system and associated informatics tools (Azizi.ilri.org ). Livestock are very different in terms of data. We collect biorepository core-data. We make use of random collections and of limited vocabularies. We hope that one day we can do natural language processing to skip ontologies step. ILRI does not impose rigid standards during project data management. Surveys are done using ODK on excel and the information system maps a field to a target
Comment: You should publish the list of terms that researchers use, and that ILRI has. We could cross the vocabularies and add a unique identifier by publishing it on the web. Bioversity can help with that step.

Second session for discussion groups - we formed 3 discussion groups to discuss topics inspired by the lightning talks.

The Group 1 discussed what Semantic standards are needed beyond crops. Currently, the CoP has a strong representation of crops. Participants mentioned that both, livestock & water communities should learn from established crop ontologies and use similar tools that are used for crop ontology (creation) for water & livestock. IWMI and ILRI will suggest ontology focal points.

Water: The group mentioned that there is a gap of Water data ontologies, particularly for the water demand side at all scales. The IMWi Water data portal is not consistent enough and will be revised which open an opportunity for collaboration. Concepts used by IMWi should be extracted into a list of terms to initiate the ontology. Water modelers’ feedback is needed.

Livestock and Fish: Need to create an ontology for livestock & for fish. For ILRI, ATOL is not a reference and the suggestion is to extract ILRI vocabulary used in their bio repositories and publish it as a starter for a standard.

Soil: standard for soil data are lacking so the creation of a Soil ontology was suggested. The effort initiated by the Soil data working group of the Interest Group for Agricultural Data (IGAD) in the Research Data Alliance was indicated.

Across domain discussion on concepts definition and context of use should be stimulated within the CoP – e.g. water on definition of dirty water, grey water vs brown water sued by livestock scientists vs water scientists definitions. Water modelers needed (livestock?)

The Group 2 discussed the use of Ontology-driven fieldbooks. Participants indicated that they use fieldbooks but the issues reside with the use of tablets in the field (light, size, …). However, spreadsheets like AgroFIMS proposes are not convenient for large experiments. Question was asked on the use of bar codes in the AgroFIMS fieldbook? The principle of using an ontology for creating fieldbook appeared valid for Is the principle valid for other experimental data collect or surveys as controlled vocabularies already exist (e.g. descriptors for genebanks). Centers like CIP have already developed Data dictionaries (sweet potato/potato) that could be useful for AgroFIMS.

The Group 3 discussed the need for Ontology mash-up and Semantic web for Agriculture. To get agricultural data on the semantic Web (RDF, Linked Open Data,
ontologies, etc), collaboration of different groups of expertise is necessary and would enable to split the work:

- computation expertise
- knowledge domain expertise
- plan for long term maintenance
- ongoing funding identified

Regarding an ontology mash-up: the CoP should identify and recommend a hierarchy for grouping similar terms and link to underlying ontologies. CoP should develop rules for dealing with inexact and/or incomplete matches to terms. If no match, then have a process to create an entry of proposed term. Ontologies and mash-up of ontologies will help the platform on Big Data in Agriculture by enabling F.A.I.R. data!

How do we get people to use ontologies?
- Tools have to be user friendly
- Ontology concept compliance should be required by funding agencies and journals
- Ontologies should provide the direct value upon search
- Helps to have champion’s model / ontology use behavior

Priority discussion topics: During the open CoP session, participants could rank on a paper board the priority topics suggested by the respondents of the Ontology CoP survey from June to July and they added their topics.

Topics per higher ranking order:

1. Arriving at mutual understanding of what (tools, approaches) need to be in place to do this and how it could be practically applied?
2. How do we develop a killer application using Ontologies in CGIAR?
3. Role of Ontology enabling the work of other CoPs
4. Review of trait dictionary, usage and submission of CO
5. Synchronizing and opening UP O developments efforts into sustainable e-infrastructure
6. Future of Ontologies in CGIAR and broader international research community

New topics added:
- Ontology look up service for Agriculture (contribute or create)
- Develop a water ontology

Friday – Closed Ontology CoP session based on results of the Wednesday session

On the COP list of domain ontologies are:
- Plants, with Crop Ontology
- Livestock (ILRI) – extract the vocabulary developed by ILRI and publish it. Community will contribute as necessary. ATOL will be checked for content. – Need a curator from the domain
- Water - need a curator from the domain
- Fish - need a curator from the domain, from Fishbase?
- Soils - need a curator from the domain – contact AfSYS team, e.g. Markus Walsh

A Farming System Ontology could result from the above mentioned ontologies with addition of the Socio-Economics. Develop higher-level concepts that can support deriving conclusion on productivity.

Test applicability of Crop Ontology Trait Dictionary template to other domains, like livestock; Also look at the AgrO development model with processes to see if it is possible to model it this way and re-use existing generic ontologies.

Stimulate cross-domain discussions to precise definitions and their context of application.

Actions
1. ILRI (Steve Kemp) and IMWI (David Wiberg) will nominate CoP focal point for their institutions.
2. Identify new CoP members to cover Soils – invite AfSYS – CoP chair
3. Revise the CoP objectives and action plan according to the discussion results – CoP chair
4. Extract knowledge domain- specific lists of concepts and publish as such for a start
   a. Livestock with ILRI
      i. List of ILRI concepts covers Africa, Asia and Latin America and includes cows, goats, chicken, bees, etc. ILRI will contact FAO for engaging team maintaining ATOL.
   b. For FISH in collaboration with WorldFish (Fishbase)
   c. Water with IMWI (Water data portal)
5. Call for external help to consolidate the lists
6. For Plants - bring the priority topics to the CoP, e.g. Discuss links to genebank traits, Participatory Varietal selection

Proposed timeline for getting actions done: PhenoHarmonIS 2018 (April-May)

Discussion with Socio-Economic Data CoP

The SED CoP will define:

1. **100 standards questions** following a hierarchy in sequence. Upon this, **20 will be the ‘MUST HAVE’**.
   - These 100 questions will be available online
   - Tutorials for enumerators won’t be added at this global level because they are very context-specific.

2. **Metadata set for questions**

   Each question will be linked to a set of metadata and the minimum list of metadata to attach to a newly added question

   I. **LABEL**: e.g. number of households, members of the household, number of adult male, number of Adult female, under 15 which means teenager, child, baby
   II. A **definition** will be given for the question and the definition Identifier will appear in the metadata
   III. **Terms in the question will use ontology terms** – each important concept will use a SOCIO ontology Identifier.
IV. The computation method will be indicated e.g. for calculating the variables on Households from the underlying data.

V. Conversion methods: store the most common ones; indicate the formula (inches, m, etc). Be flexible enough to accommodate local measurements.

VI. Open data status - What can be open or not and what cannot be disclosed or not – will be indicated in the metadata

VII. The value format - text, controlled vocabulary, unit of measurement when possible

VIII. Indicate which language was used to ask the questions. It has great importance to know what translation was done and how? translation into local languages are often done on the fly.

The CoP will provide recommendations on how the collected data can be converted given who will be collecting (e.g. region where metric system if different)

The questions will have to be downloadable into tablets

UPLOAD the surveys into a repository

A file may go beyond the 100 standard questions.

Annotation of the data and of the questions

- Data can be Raw data or Analyzed data
- upload the file that can be an Excel or an interview record, SPSS file, R file
- Protocol of the survey
- Protocol of the analysis
- Conversion method used
- Language used for the Translation
- And add anything that will make the data set useable and comparable

Metadata will have Keywords linked to ontologies. Ontologies will be applied at the data level. Metadata will be searchable and structured = F.A.I.R.

Action:
I. Start from the questions and extract the concepts
II. Take an example and test the full workflow from survey to upload in Dataverse.
III. Guidance from the ontology COP will be necessary

Participants of the Ontology Session, Wednesday

<table>
<thead>
<tr>
<th>LAST NAME</th>
<th>FIRST NAME</th>
<th>INSTITUTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robbins</td>
<td>Kelly</td>
<td>Cornell</td>
</tr>
<tr>
<td>Mwanzia</td>
<td>Leroy</td>
<td>CIAT</td>
</tr>
<tr>
<td>Cruz</td>
<td>Paola</td>
<td>CIAT</td>
</tr>
<tr>
<td>Name</td>
<td>Last Name</td>
<td>Affiliation</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Kemp</td>
<td>Steve</td>
<td>ILRI</td>
</tr>
<tr>
<td>Galcano</td>
<td>Carlos</td>
<td>Corpoica</td>
</tr>
<tr>
<td>Nyakundi</td>
<td>Fridah</td>
<td>CIAT</td>
</tr>
<tr>
<td>Lowe</td>
<td>Brian</td>
<td>Ontocale</td>
</tr>
<tr>
<td>Laporte</td>
<td>Marie-Angélique</td>
<td>Bioversity International</td>
</tr>
<tr>
<td>Arciniegas</td>
<td>Vicky</td>
<td>SiB Colombia/GBIF</td>
</tr>
<tr>
<td>Sanchez</td>
<td>Adriana</td>
<td>Independent Consultant</td>
</tr>
<tr>
<td>Silverstein</td>
<td>Kevin</td>
<td>University of Minnesota</td>
</tr>
<tr>
<td>Sanchez</td>
<td>Julian</td>
<td>University of Cauca</td>
</tr>
<tr>
<td>Heider</td>
<td>Bettina</td>
<td>CIP</td>
</tr>
<tr>
<td>Abreu</td>
<td>David</td>
<td>CCAFS/CIAT</td>
</tr>
<tr>
<td>Tobon</td>
<td>Hector</td>
<td>CCAFS</td>
</tr>
<tr>
<td>Castaneda</td>
<td>Nora</td>
<td>Global Crop Diversity Trust</td>
</tr>
<tr>
<td>Sotelo</td>
<td>Seven</td>
<td>CIAT</td>
</tr>
<tr>
<td>Silverstein</td>
<td>Kevin</td>
<td>Univ. Minesotta</td>
</tr>
<tr>
<td>Arciniegas</td>
<td>Vicky</td>
<td>GBIF, SiB Colombia</td>
</tr>
<tr>
<td>Kastrantas</td>
<td>Kosta</td>
<td>Agroknow</td>
</tr>
<tr>
<td>Smyth</td>
<td>Karen</td>
<td>Livestock Data for Decision</td>
</tr>
</tbody>
</table>

Paper boards
Tweets sent during the sessions