GACS Core: Creation of a Global Agricultural Concept Scheme

Thomas Baker[1], Caterina Caracciolo[2], Anton Doroszenko[3], Lori Finch[4], Osma Suominen[5], and Sujata Suri[4]

[1] Independent FAO consultant, Bonn, Germany
[2] Food and Agriculture Organization of the UN, Rome, Italy
[3] CAB International, Wallingford, UK
[4] National Agricultural Library, USDA, Beltsville, MD, USA
[5] National Library of Finland, Helsinki, Finland

Abstract. The most frequently used concepts from AGROVOC, CABT, and NALT – three major thesauri in the area of food and agriculture – have been merged into a Global Agricultural Concept Scheme, with 15,000 concepts and over 350,000 terms in 28 languages in its beta release of May 2016. This set of core concepts ("GACS Core") is seen as the first step towards a more comprehensive Global Agricultural Concept Scheme. In the context of a new Agrisemantics initiative, GACS is intended to serve as hub linking user-oriented thesauri with semantically more precise and specialized domain ontologies linked, in turn, to quantitative datasets. The goal is to improve the discoverability and semantic interoperability of agricultural information and data for the benefit of researchers, policy-makers, and farmers in support of innovative responses to the challenges of food security under conditions of climate change.

1 A shared concept scheme

The Food and Agricultural Organization of the United Nations (FAO), CAB (Centre for Agriculture and Biosciences) International (CABI), and the National Agricultural Library of the USDA (NAL) maintain separate thesauri about agriculture, food, and nutrition for indexing bibliographic databases. The AGROVOC Concept Scheme (created 1982)[6], CAB Thesaurus (1983)[7], and NAL Thesaurus (1990s)[8] are used to index, respectively, AGRIS (8 million records), CAB Abstracts (11.5 million), and Agricola (5.2 million).

Having collaborated in the 1990s on mappings and common classifications, the three organizations joined forces again in 2013 to explore the feasibility of creating a shared Global Agricultural Concept Scheme (GACS).[9] The project aimed at facilitating search across databases, at improving the semantic reach of their databases by supporting queries that freely draw on terms from any mapped thesaurus, and at achieving efficiencies of scale from collaborative maintenance.

[6] `http://aims.fao.org/agrovoc`
[7] `http://www.cabi.org/cabthesaurus/`
[8] `http://agclass.nal.usda.gov/`
[9] `http://agrisemantics.org/gacs`

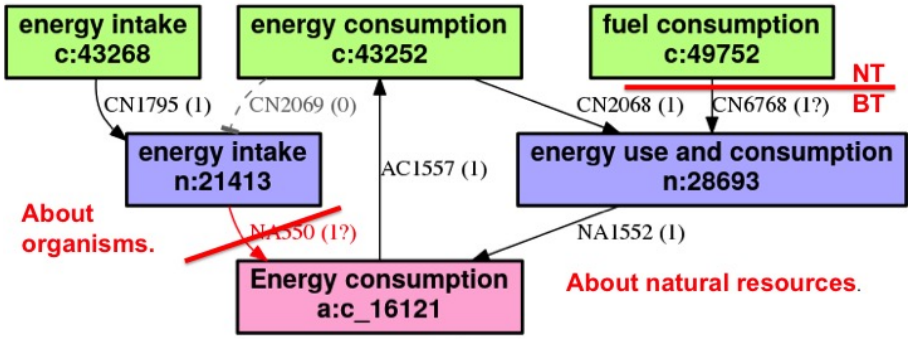

Fig. 1. Cluster of mappings between AGROVOC (a:), CAB Thesaurus (c:), and NAL Thesaurus (n:) flagged as a "lump"

2 Creating GACS Core

The process began in March 2014 with the formation of a joint GACS Working Group. After a preliminary analysis found that some 98% of the indexing fields in AGRIS used just 10,000 out of the 32,000–plus concepts in AGROVOC, mapping began with three selections of 10,000 most frequently used concepts. These were algorithmically mapped to each other, pairwise, by adapting the AgreementMakeLight ontology matching system[10]; mappings were verified by hand; a second algorithm checked for clusters of inconsistent mappings ("lumps"); the lumps were discussed online or in meetings; as a result of decisions taken, the mappings were corrected by hand (to remove mappings or to change their meaning); and the corrected mappings were used to generate new concepts algorithmically. Concepts in the new concept scheme were given URIs in a new namespace[11] and represented in RDF using the W3C standard, Simple Knowledge Organization Scheme (SKOS).[12] This initial set of core concepts is called GACS Core in the expectation that GACS will become more comprehensive in scope and less centralized in its maintenance.

Figure 1 shows a lump detected by algorithmic analysis of the manually verified mappings, the meanings of which are spelled out in Table 1. In this case, the working group determined that *energy intake* had to do with organisms and that *energy consumption*, along with the narrower *fuel consumption*, had to do with natural resources. By deleting the mapping `NA550`, redefining `CN6768` as narrow-to-broad, and letting the concept-creating algorithm pick the most popular labels, three new GACS concepts were created, with mappings back to their sources (see Figure 2).

[10] `https://github.com/AgreementMakerLight/AML-Jar`
[11] `http://id.agrisemantics.org/gacs/`
[12] `https://www.w3.org/TR/skos-reference/`

Table 1. Set of manually verified mappings (before correction)

ID	Source concept	Mapping	Target concept
AC1557	agro:c_16121	fully equivalent to	cabt:43252
CN2069	cabt:43252	not related to	nalt:21413
NA5507	nalt:21413	probably equivalent to	agro:c_16121
CN1795	cabt:43268	fully equivalent to	nalt:21413
CN2068	cabt:43252	fully equivalent to	nalt:28693
NA1552	nalt:28693	fully equivalent to	agro:c_16121
CN6768	cabt:49752	probably equivalent to	nalt:28693

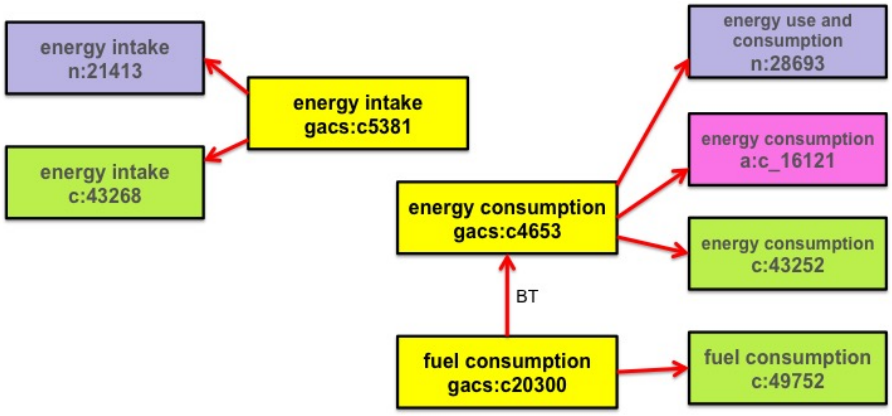

Fig. 2. Corrected mappings form concepts in GACS Core

3 Correcting GACS Core

GACS Core Beta 3.1, soft-launched in May 2016, provides 15,000 concepts labeled with 350,000 terms, some in more than twenty-five languages. This set of concepts is considered stable, with URIs that are not expected to change. The reconciliation of diverse source concepts into common GACS Core concepts, illustrated above, is largely complete. Some problems resulting from the integration process, such as overlapping labels, have been substantially fixed, though much detailed quality control remains to be done. During this test phase, implementers are encouraged to use GACS Core on an experimental basis and provide feedback.

The evolving editorial policies for GACS Core follow best practices of modern thesaurus design as per ISO 25964, "Thesauri and interoperability with other vocabularies": concepts, described with natural-language labels, clarified with definitions and scope notes, mapped to other concepts with associative and hierarchical relations, and organized into thematic groups. For Version 1.0, GACS Core must be cleaned and corrected with respect to the following:

Thematic groups. Many thesauri, including NAL Thesaurus, General Finnish Ontology (YSO), UNESCO, and STW Thesaurus for Economics (Germany) provide a thematic division of concepts into clusters orthogonal to the hierarchy of broader and narrower concepts. To provide this, the GACS working group revived an existing classification scheme that had been jointly developed by their predecessors in the 1990s, incorporated into the 1999 release of CAB Thesaurus, then set aside. Thematic group information gleaned from the 1999 CAB Classified Thesaurus (a separate thesaurus, soon to be re-released in electronic form) quite evenly covers 82% of the concepts in GACS Core, leaving circa 2,750 unassigned.

Custom relations. AGROVOC and CAB Thesaurus each use a few properties to specify the nature of a relation between two concepts beyond the generic thesaurus relations of broader, narrower, and related. Previous efforts to "ontologize" thesauri with such additional relations revealed practical obstacles to ensuring that the properties would be applied consistently, comprehensively, and maintainably. For GACS Core, the working group decided that custom relations must meet use cases salient enough to justify the effort. Two properties qualified: `hasProduct`, and `productOf`, for relating *fish (product)* to *fish)*, the organism.

Hierarchical relations. When concepts from the three sources were merged into GACS Core concepts, their hierarchical and associative relations were also merged. GACS Core has some 600 "top concepts," or concepts with no broader concept. Top concepts are typically meant to facilitate faceted browsing or the creation of microthesauri. Ideally, top concepts should fit on just a page or two. Likewise as a result of mapping, almost one third of the concepts in GACS Core ended up with more than one broader concept. While a certain amount of polyhierarchy may be inevitable, even desirable, best practice is to keep the hierarchy as simple and pyramid-like as possible. The working group will examine how similar thesauri define their top concepts and evaluate the use cases for top concepts in light of the thematic groups. Once a set of top concepts is agreed, along with a set of principles for assigning hierarchical relations, existing relationships will be carefully vetted, pruned, and adjusted.

Semantic types. Some thesauri differentiate concepts by type, such as *organisms* or *places*. Thesauri can use the hierarchies under top-level concepts to roughly group concepts of a given type (as with AGROVOC), though hierarchies may not follow the principle of general-to-specific (hyponymy) strictly enough to ensure that an "isa" ("type of") relationship would always hold; hierarchies may also contain "part of" (meronym) relationships. Type can be assigned to concepts using subject categories (as with CAB Thesaurus) or other type systems, such as the UMLS Semantic Network (as with NAL Thesaurus). While recognizing that semantic types could usefully clarify the meaning of concepts, provide transitive "isa" relationships, and pull together concepts from across the hierarchy, the GACS working group opted to explore the benefits of committing to types by starting with simple set of *Chemical*, *Geographical*, *Organism*, *Product*, and the generic *Topic*.

4 GACS extensions and modules

Almost one third of the concepts in AGROVOC (11,000) are now tightly coupled to GACS Core, leaving a long tail of circa 21,000 concepts that are both unmapped and less frequently used. Continuing to maintain this long tail under the AGROVOC brand is possible but poses problems if GACS Core is to be officially preferred. Users would face one set of GACS Core concepts and a larger set of AGROVOC concepts, with different URIs and browsable separately. As one possible solution, the unmapped concepts of AGROVOC could be assigned GACS URIs but not marked as being in GACS Core, creating an AGROVOC-based extension to GACS. GACS URIs would be promoted while AGROVOC URIs remained mapped to their GACS equivalents and thus usable in perpetuity.

Generalizing from this case, the notion of a GACS Extension could be defined as a set of concepts within the general scope of GACS but with no overlap to GACS Core. GACS Extensions would not be subject to the constraints of shared maintenance. GACS Extensions would provide a home for concepts pruned from GACS Core. Ideally, they would be searchable through a single interface with GACS Core simply by selecting from a menu.

Concepts of well-defined types, such as organisms, geographical names, or chemicals, could in principle be defined as GACS Modules, the maintenance of which could in principle be delegated entirely to other, more expert communities. Exploration of this option will begin with vocabularies for soil data[1].

5 Towards an Agrisemantics Ecosystem

GACS Core is intended to serve as a hub within Agrisemantics, an emerging community network of semantic assets relevant to agriculture and food security.[13] The Agrisemantics idea was explored in a July 2015 workshop, with support from the Gates Foundation[14], and elaborated in the Chania Declaration of May 2016[15], which looks towards an "ecosystem of linked data repositories, data management services and virtual collaboration environments to increase the pace of knowledge production for agricultural innovation."[16] This goal is currently being pursued by a new Agrisemantics Working Group of the Research Data Alliance (RDA).

Like other thesauri, GACS Core provides topics for tagging information resources from bibliographic abstracts, journal articles, and grey literature, to Web resources such as videos, podcasts, and courseware. Its topical concepts, such as *farmers' attitudes* and *family relations*, are fuzzy enough to accommodate the perspectives of a broad diversity of information seekers. In contrast, datasets for

[13] http://agrisemantics.org
[14] http://aims.fao.org/sites/default/files/Report_workshop_Agrisemantics.pdf
[15] http://blog.agroknow.com/?p=5067
[16] http://blog.agroknow.com/?p=5067

quantitative analyses, such as sensor readings and crop yields, are composed of data elements defined with precision and at a fine level of granularity. Datasets are typically defined in the context of a particular software application and serialized in formats specific to that application. Interoperability across datasets is limited by the sheer effort required to determine equivalences among differently named elements, then to extract those elements from a diversity of formats.

Semantic authority control of quantitative data elements

Between thesauri such as GACS Core and quantitative datasets lie ontologies—focused sets of concepts with precise definitions, global identifiers, and strongly typed semantic relationships. The Agrisemantics initiative proposes to test the idea that ontologies can provide a bridge between general-purpose thesauri and application-specific datasets. Ontologies can provide stable, global identity to concepts found under a diversity of local names and embedded in a diversity of software applications, in effect functioning as authorities for data elements, analogously to the library science notion of "authority control."

Semantic authority control for data elements could improve food security by supporting, for example, an analysis of the yield gap in sub-Saharan Africa. Such an analysis would need to draw both on crop-related datasets and on relevant research and multimedia resources indexed in bibliographic databases. A wheat data element, labeled 'GW' in a phenotype dataset, could be mapped to the concept 'grain weight' as defined and globally identified in the CGIAR Crop Ontology[2][17]. In turn, the Crop Ontology concept could be mapped to the broader concept 'Grain' in GACS Core. Searches could return not only datasets about grain weight, but references to published papers where the weight of the grain was studied.

In the context of Agrisemantics, GACS can serve as a hub for a richly linked network of thesauri and domain-specific ontologies, linked to innumerable quantitative datasets. By facilitating the integration of data and research results from many sources, such a semantic platform can support innovation in agriculture and contribute to the creation and management of sustainable food systems.

References

1. L'Abate, G., Caracciolo, C., Pesce, V., Geser, G., Protonotarios, V., Costantini, E.A.: Exposing vocabularies for soil as Linked Open Data. Information Processing in Agriculture 2(3-4), 208–216 (oct 2015), http://dx.doi.org/10.1016/j.inpa.2015.10.002
2. Shrestha, R., Matteis, L., Skofic, M., Portugal, A., McLaren, G., Hyman, G., Arnaud, E.: Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. Frontiers in Physiology 3 (2012), http://dx.doi.org/10.3389/fphys.2012.00326

[17] http://www.cropontology.org