Global Agricultural Concept Scheme

A Hub for Agricultural Vocabularies

Tom Baker
Independent FAO consultant
Bonn, Germany

Caterina Caracciolo
Food and Agriculture Organization of the UN (FAO)
Italy, Rome

Elizabeth Arnaud
Bioversity International
Montpellier, France

***Abstract*— Thesauri are used to tag semi-structured documents, texts, while more complex semantic structures are used to describe (annotate) scientific data. We are creating a Global Agricultural Concept Scheme (GACS) by mapping AGROVOC, CABT and NALT – three major thesauri in the area of food and agriculture, with a beta release in May 2016. We see GACS as a hub linking user-oriented thesauri with semantically more precise domain ontologies linking, in turn, to datasets about food and agriculture, in order to make that data more interoperable and reusable**

Keywords*—*thesauri, ontologies, food, agriculture, GACS, AGROVOC, CABT, NALT, Crop Ontology

I. Global Agricultural Concept Scheme

The Food and Agricultural Organization of the United Nations (FAO), CAB International (CABI), and the National Agricultural Library of the USDA (NAL) have long maintained separate thesauri about agriculture, food and related topics -- the AGROVOC Concept Scheme[1], CAB Thesaurus, and NAL Thesaurus – for use in indexing their respective bibliographic databases:: AGRIS (8 million records), CAB Abstracts (8.3), and Agricola (5.2). the AGROVOC Concept Scheme, CAB Thesaurus[2], and NAL Thesaurus[3]. The thesauri provide globally identified concepts for use in automated indexing and retrieval, subject description, natural language processing, and translation.

Having previously collaborated on mappings and common classifications, the three organizations resolved in 2013 to explore the feasibility of pooling their most frequently used concepts into a jointly maintained Global Agricultural Concept Scheme (GACS). GACS was seen as the first step towards improving the coherence and interoperability of agricultural data – a vision explored in a July 2015 workshop on "Agrisemantics"[4], with support from the Gates Foundation, elaborated in the Chania Declaration[5] of May 2016, and pursued by an Agrisemantics Working Group that is forming within the Research Data Alliance initiative.

GACS Core Beta 3.1[6], soft-launched at the Open Harvest workshop of May 2016, provides 15,000 concepts formed by mapping and merging the most frequently used concepts from the three source thesauri. GACS Core concepts are labeled in multiple languages, with some in more than twenty-five languages. The soft launch opened a period of testing and feedback in preparation for the next phase of its development, which will begin in circa October 2016. GACS Core Beta 3.1 presents a set of concepts that is considered to be fairly stable, with URIs that are not expected to change (see an example of concept in GACS in Fig. 1). Problems resulting from the integration process, such as overlapping labels, have been substantially fixed, though much detailed work remains to be done, notably the specification of a common hierarchical structure. During this test phase, implementers are encouraged to use GACS on an experimental basis and provide feedback.

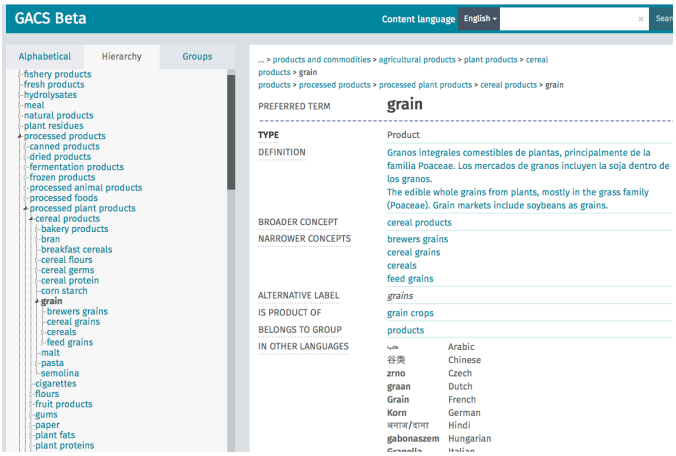

Fig. 1 A concept in GACS

In the next phase of development, the scope of GACS will be broadened beyond the core. Concepts from some of the source thesauri that were not included in GACS Core may be given an id.agrisemantics.org URI in a GACS Extension to be maintained by their original owners or, optionally, in collaboration. The notion of GACS Module anticipates a

[1] http://aims.fao.org/agrovoc
[2] http://www.cabi.org/cabthesaurus/
[3] http://agclass.nal.usda.gov/
[4] http://aims.fao.org/sites/default/files/Report_workshop_Agrisemantics.pdf
[5] http://blog.agroknow.com/?p=5067
[6] http://agrisemantics.org/gacs

longer-term need to devolve maintenance of distinct types of concepts, such as organisms or geographical names, to communities of experts.

II. Semantic Assets for Food and Agriculture

Information relevant to food and agriculture encompasses data collected on factors ranging from yield and climate to demographics and markets., Information is presented in forms ranging from narrative texts (policy, technical, and scientific documents) through structured datasets (empirical data). Information may be graphically visualized, e.g., plotted onto timelines or maps, or plugged into models for nowcasting or for forecasting trends. All types of data, from the analytical to the empirical, are required for achieving sustainable food systems.

Thesauri provide concepts for indicating the overall topic of information resources, usually semi-structured texts such as bibliographic abstracts, journal articles, but also videos and courseware. Empirical data is composed of data elements with precise definitions at defined levels of granularity. Datasets are typically serialized in formats specific to a particular software application, and their individual data elements are named within the context of that particular application. Interoperability across datasets is hampered by the sheer effort required to determine equivalences among differently named elements, then to extract sets of comparable elements from a diversity of applications and formats. Ontologies, focused set of related concepts specified with precise definitions and global identifiers, are increasingly used to "annotate" data. However, ontologies too may embody ad-hoc semantics in different degrees, and are usually totally disconnected from the world of thesauri, so preventing a seamless access to "hard" and soft data alike.

III. Linking Thesauri to Data via Ontologies

The more fuzzily defined, globally identified concepts of general-purpose, search-oriented thesauri and concept schemes, such as GACS, may be mapped to the more precisely defined, globally identified, domain-specific, application-oriented ontologies and, from there, to locally defined data elements embedded in software-specific databases. An unbroken chain may be formed linking the most general concepts to the most specific data elements. Semantic authority control for data elements facilitates the re-use of datasets, and links from precise ontologies to search-oriented concepts facilitates the discovery of those datasets.

One path to data interoperability is to use appropriately defined ontologies – i.e., ontologies that not only enable the extraction of data from a database (process often called "data annotation"), but that can also situate data within the appropriate "context" -- a modeled set of data about the time and place of its collection along with any additional elements required for its correct interpretation. Another path is to place those ontologies in a network with other semantic assets, including the thesauri and concept schemes used to express the "topicality" of information resources. Such an integration of semantic assets may support, for example, an analysis of the yield gap in sub-Saharan African countries by providing well-connected data elements across a diversity of cropwheat-related datasets from databases and repositories along with multi-media information, and relevant literature from main bibliographic databases like AGRIS, CABI and NAL with the goal of improving food security.

The Agrisemantics vision points in two directions: on the one hand, to turn GACS into a more extensive network of thesauri and concept schemes to ensure the appropriate coverage for our domain of interest. In particular, we are going to test the notion of a GACS Extension on the example of AGROVOC. On the other hand, we aim at establishing tools and methodologies to connect GACS and its constellation of "extensions" to multiple domain-specific ontologies.

The first ontology we will be working with is the Crop Ontology [1], which supports data comparison and interpretation at a higher granularity by providing a means for annotating data element with trait measurement method and unit or scale. (See Fig. 2)

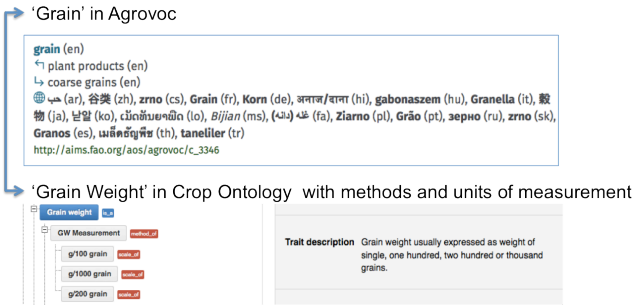

Fig. 2 Mapping from thesaurus to ontology

More specifically, a wheat data element labeled with the code "GW" in a phenotype dataset can be mapped to the general concept "grain weight" as defined, and given global identity (URI), in the CGIAR Crop Ontology[7]. The CO term 'Grain Weight' can, in turn, be mapped to 'Grain' in AGROVOC and GACS. More information can then be discovered through a query system using this mapping that will return, aside from datasets related to grain weight, references to published papers where grain weight was studied.

Acknowledgments

Special thanks to the GACS Working Group: Tom Baker, Caterina Caracciolo, Anton Doroszenko, Lori Finch, Sujata Suri, and Osma Suominen.

References

[1] Rosemary S., Matteis L., Skofic M., Portugal A., McLaren G., Hyman G., Arnaud E.: 2012. 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, vol. 3

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