Workshop on the future of GACS: Analytical summary

Tom Baker
Version: 2017-02-06
This document: https://github.com/agrisemantics/2016_11_goettingen/blob/master/gacs_workshop_summary.pdf

A workshop held in November 2016 in Goettingen, Germany introduced the Global Agricultural Concept Scheme (GACS), a set of high-level concepts about agriculture derived from three major thesauri. The workshop elicited requirements for GACS from current users and considered alternative scenarios for its future development.

GACS was created between 2014 and 2016 as a partial merger of three major agricultural thesauri -- AGROVOC, CAB Thesaurus (CABT), and NAL Thesaurus (NALT)-- which are maintained, respectively, by the Food and Agriculture Organization of the UN (FAO), Centre for Agriculture and Biosciences International (CABI), and the USDA National Agricultural Library (NAL). GACS Core Beta was rolled out in May 2016 with 15,000 concepts and over 350,000 labels in 28 languages.

GACS is seen as a foundational step towards Agrisemantics, a future community network of semantic assets relevant to agriculture and food security. As originally envisioned in July 2015, Agrisemantics would provide "a well-integrated clearinghouse of machine-readable semantic assets in agriculture and nutrition, such as vocabularies, code lists, ontologies, taxonomies, and statistical indicators". In this context, GACS would serve as a hub vocabulary to which other semantic assets would map. A workshop in May 2016 took this idea further by envisioning a chain of mappings from the high-level concepts of GACS, through concepts defined with more precision in domain-specific ontologies and beyond, to concepts used in a vast diversity of empirical datasets, such as sensor readings and crop yields, in a multitude of software applications and serialization formats.

By the time of the workshop in Goettingen, the original hope for using GACS as a replacement for some of its source thesauri had come to be seen as unrealistic except, potentially, for the case of AGROVOC. However, an enthusiastic reception for GACS in the wider community, and expressions of interest from numerous potential stakeholders, seemed to validate the idea of an interoperability hub for semantics in agriculture.

Principles on which there was consensus

The workshop reached no clear conclusions, but there was broad recognition that the current GACS Core Beta was but a starting point for something new and potentially quite significant. Beneath surface disagreements, the participants found themselves in solid agreement on some basic principles regarding the future of GACS:

Design principles

- **Commitment to persistent URIs**. Concepts may become obsolete, but their identifiers, once published, must never disappear. Credible guarantees of URI persistence are crucial for establishing trust in GACS, and these can only be credibly made by the owners of the Internet domain under which the URIs are coined. At the workshop, the owners of the domain http://agrisemantics.org/ reiterated their commitment to the persistence of URIs coined in the http://id.agrisemantics.org/gacs/ namespace.
- **Lightly specified semantics**. Any semantic relationships defined among the concepts in GACS are inherited wherever those concepts are reused. There was general agreement that these relationships should be specified as lightly as possible in order to maximize their reusability. The workshop solidly approved the decision to express GACS as a SKOS concept scheme, with broad-brush semantics, and not as a formally stricter OWL ontology.
- **Focus on concepts widely used and globally significant**. All content in semantic assets is in principle reusable, but some parts are actually reused more than others and thus more important for interoperability, while the parts that are the most business- or application-specific are less likely to be used widely. There was general agreement that GACS should focus on concepts more likely to be widely used.

Potential role in the data ecosystem

- **A diverse ecosystem of vocabularies**. There was agreement that a healthy data ecosystem needs a diversity of vocabularies, maintained by diverse communities of experts, with their own editorial boards, at various levels of generality and specialization and in namespaces under separate ownership and control. (As discussed below, however, there was less agreement about how much of this diversity should be encompassed within GACS itself as opposed to within the broader vocabulary ecosystem of Agrisemantics and of the Web as a whole.)
- **A hub for more specialized semantic assets**. The lightly specified, widely used, globally significant concepts of GACS can be linked to more specialized concepts in domain-specific ontologies. GACS can usefully function as a switchboard, glue, or interoperability layer loosely connecting clusters of quasi-equivalent concepts.
- **A hub for "local" semantics embedded in a diversity of data formats**. With its globally valid URIs, GACS can provide semantic authority control for "non-semantic" resources such as databases and spreadsheets, the tables and columns of which overwhelmingly use identifiers of strictly local scope.
- **A source of building blocks for other semantic assets**. The pool of shared concepts in GACS can provide building blocks for constructing more specialized semantic assets, such as ontologies, either by directly reusing GACS URIs, unchanged, or by creating more specific concepts mapped to GACS URIs.

- **Support for views, or subsets**. Subsets of GACS concepts on focused topics such as "crops" should be accessible, through Web services or selective views, to users with special requirements.

Structure of GACS maintenance

- **Distributed maintenance**. The maintenance of anything as comprehensive as GACS may need to be spread across multiple editorial boards.
- **Progressive decentralization of authority**. The scope of GACS currently includes specific types of concept, such as organisms and viruses, that continually evolve and are thus hard for centralized maintenance teams to keep up-to-date. Over time, responsibility for maintaining such concepts should be delegated to external communities of experts.
- **Use cases beyond bibliographic indexing**. GACS Core Beta was created as a merger of concepts from thesauri that are used primarily for indexing databases of bibliographic abstracts, but there was general agreement at the workshop that the scope of GACS could be broadened to support the interoperability of datasets.
- **Growth through community input**. GACS should draw new concepts from a diversity of sources. Users of GACS should be enabled to propose concepts of global significance to fill gaps in the common pool, making those concepts available, in turn, as hub concepts for interoperability and as building blocks for semantic applications downstream.

GACS as seen by the GACS Working Group

The majority view of the GACS Working Group (staff from the three partners plus consultants) sees the Global Agricultural Concept Scheme as a concept scheme according to the model of Simple Knowledge Organization System (SKOS), itself based on standard models for thesauri. The group feels that GACS should provide concepts that are widely used in the agricultural domain, though this scope need not be limited to the thesauri from which it was derived. With some dissension, the GACS Working Group subscribes to the following principles:

- **One namespace**. Using just one base URI, as GACS currently does, is a common practice followed by mainstream resources such as GeoNames, EuroVoc, and Library of Congress Subject Headings, as well as by its sources AGROVOC, CABT, and NALT.
- **Distributed maintenance**. To remain relevant, GACS needs to grow with new concepts while de-emphasizing concepts that become less widely used or obsolete. For reasons of maintainability and coherence, the set of terms maintained jointly by GACS partners should remain fairly stable, or ideally even shrink in size. In order to prune obsolete concepts without violating the principle of URI persistence, one part of the GACS namespace needs to be considered "core" (in the sense of jointly curated) with respect to parts of the GACS namespace not under joint curation to which infrequently used or obsolete concepts can be moved ("extensions"). The current GACS cannot be

pruned without having a destination, or status, within GACS, for pruned concepts.

- **Devolution of authority**. It is desirable, over time, to delegate the maintenance of distinct subsets of GACS concepts to external authorities, for example, to delegate the maintenance of countries to FAO Geopolitical Ontology. Doing this without violating the principle of URI persistence implies that the existing GACS concepts for countries would need to be mapped to, and considered henceforth to be dependent on, concepts maintained by an authority external to GACS (such as FAO Geopolitical Ontology). In this sense, a set of relevant GACS URIs would be considered a "reflection" of primary URIs maintained by the external authority. Reflected URIs serve as a backup against changes in external organizations, because if mapped URIs under DNS domains not controlled by GACS were to disappear, the GACS URIs would remain.
- **Support for views, or subsets**. The GACS Working Group sees a clear need to make subsets of GACS concepts on particular topics available as views or for download, whether these be called subvocabularies, concept groups, modules, profiles, or remixes. In principle, such subsets could draw on concepts from the core, extensions, or reflections, and could be maintained either as an integral part of GACS or in the form of externally maintained link sets layered over GACS.

The GACS Working Group has never considered GACS Core Beta to be fully finished in its currently published form. The beta release of May 2016 was meant to elicit feedback. With a manageable amount of effort and modest funding, the Working Group could decide on principles for hierarchical relationships, correct GACS accordingly, and complete various other tasks related to quality control, such as adding definitions. The Working Group feels this can and should be done without increasing the size of GACS and could, indeed, result in a smaller, tighter GACS. The development of modules, or of subvocabularies or concept groups or profiles or remixes, should proceed with proof-of-concept experiments that clarify principles and processes before attempting more ambitious expansions of mission or scope.

Discussion

With hindsight, GACS Core Beta appears to have been called *core* for two distinct reasons: 1) Since July 2015, GACS has been envisioned as a (or perhaps *the*) central hub, or *core*, within the broader semantic landscape of Agrisemantics; and 2) In order to prune concepts from the part of GACS under joint maintenance without deleting their URIs, the working group needed a name (*core*) for the jointly curated part of the GACS namespace as distinct from the parts curated in a more agile, decentralized way.

Much of the discussion at the workshop involved deconstructing such concepts and challenging their underlying assumptions, specifically:

- the notion of GACS as a SKOS concept scheme with a single namespace and central editorial board;

- the notion of GACS as a namespace partitioned into a jointly maintained *core*, with lesser-used, *long-tail* concepts held in non-jointly-maintained *extensions*;
- the notion of GACS, envisioned in the July 2015 workshop, as the central hub vocabulary for a larger universe of semantic assets (Agrisemantics).

Alternative, partially overlapping visions of GACS were proposed:

- GACS as a universe of vocabularies unto itself;
- GACS as but one star in a semantic constellation.

GACS as a universe of vocabularies unto itself

One view holds that "Global Agricultural Concept Scheme" should be understood as an inclusive umbrella for a broad collection of semantic assets and namespaces. At one extreme, GACS might be seen as a set of focused, highly curated lists exhaustively covering specific topics such as crops, pests, commodities, geographical places, agricultural practices, and livestock, each list governed by its own editorial board and free either to use the main GACS namespace, http://id.agrisemantics.org/gacs/, or to use an external namespace yet still be considered to be part of GACS. Through interfaces that leverage Web services, mappings, and hierarchical information, relevant lists or selections could be presented to users. GACS could be crowdsourced in Github, where any community could propose a vocabulary, with its own namespace, for approval by the GACS community. GACS would then consist of many namespaces, democratically, with no particular center.

On the other hand, the story "one concept scheme, one namespace" is arguably easier to explain and to sell. If GACS really did consist of seventeen namespaces, those namespaces would be replicated in any semantic assets derived from GACS, and webmasters dislike having to deal with multiple namespaces. In a decentralized landscape of more focused vocabularies, it would also be difficult to control the inevitable overlap, for example the coining of multiple URIs for *wheat* seen somewhere as a plant and elsewhere as a commodity. The idea of modules, moreover, in principle satisfies the requirement for focused subsets, as seen in the very instructive example of LandVoc (see Appendix).

GACS as but one star in a semantic constellation

Some participants saw GACS Core Beta as an upper-level concept scheme useful for generic description of agricultural publications but not specific enough to ensure compatibility between databases. For them, whether GACS should be called "GACS" or something else was not the main issue; rather, the issue was the word *core*, as it could imply in this context that bibliographic databases are the most important application, as opposed to database columns or reference lists more relevant for an audience of, say, startup companies.

Others asked whether the existing GACS Core Beta could be recast as but one example of the sort of artifact that the GACS (or Agrisemantics) community wants to

produce -- an *exemplar* in terms of process and scope. Might it retain its own namespace, governance, and identity, and even distinguish between more and less intensely curated parts? Inasmuch as GACS was created as a joint effort of three organizations, might GACS provide a model for other such collaborative efforts?

Space for compromise?

One take-away from the workshop is that names really do matter. The ambitious handle Global Agricultural Concept Scheme, and GACS Core Beta, create both high expectations and considerable difficulties. Beyond what merits terms such as *core* (and its countpart, *extension*) may have as technical terms related to curation and process, it cannot be denied that they are often understood to imply *first-class* versus *second-class*, *important* versus *unimportant*. And what is important, or central, is clearly a question of standpoint; terms seen by some as a long tail may be seen by others as core.

Many of the disagreements expressed at the workshop boil down to naming. Everyone agreed that GACS should be a springboard for building a more coherent universe of semantics for the sake of improving discoverability and interoperability across a diversity of communities and applications. But should "Global Agricultural Concept Scheme" be the name for the whole universe? Or should GACS be the sun at the center of an Agrisemantics solar system? Or just one star in an Agrisemantics constellation, itself just a part of a galaxy (the Web)? These distinctions are not absolute, because everyone agrees with the notion of a module, or vocabulary within a vocabulary. The nature of these disagreements shows that we are re-thinking this ecosystem in ways barely imaginable just a quarter century ago.

To use another metaphor, it is also clear that the GACS garden, whatever crops it holds, must change and evolve if it is to remain healthy and nourishing. We can only hope that this garden, throughout its transformations, will at least remain well-tended. If we assume that its plants, the URIs, cannot simply be weeded out and discarded, but merely transplanted, we must look not just to the well-tended plot but to the conservation of the somewhat wilder fields beyond.

Links

1. Tom Baker. Workshop on the future of GACS: Report, February 2017, https://github.com/agrisemantics/2016_11_goettingen/blob/master/gacs_workshop_report.md
2. Workshop on the future of GACS: Meeting minutes, https://github.com/agrisemantics/2016_11_goettingen/blob/master/minutes
3. Workshop on the future of GACS: Presentations, https://github.com/agrisemantics/2016_11_goettingen/blob/master/presentations
4. Tom Baker, Caterina Caracciolo, Anton Doroszenko, Lori Finch, Osma Suominen, Sujata Suri. GACS Core: Creation of a Global Agricultural Concept Scheme, November 2016, https://github.com/agrisemantics/2016_11_goettingen/blob/master/mtsr_paper
5. Tom Baker, Caterina Caracciolo, Yves Jaques. Report on the workshop "Improving Semantics in Agriculture", July 2015, http://aims.fao.org/sites/default/files/Report_workshop_Agrisemantics.pdf

Appendix: Use cases and best practice

AgroPortal is a searchable repository of ontologies in the agronomic domain. Its guiding use cases are: data integration related to rice; development of a framework for publishing wheat data using open standards; publication vocabularies of produced by INRA scientists in order to foster their reuse; support for the Crop Ontology project, which publishes ontologies for describing germplasms, traits, and evaluation trials; and support for VEST, the GODAN map of agri-food data standards. AgroPortal annotates its ontologies and provides mappings to ontologies in the NCBO Bioportal. Users can also add mappings by hand. These mappings are used by AgroPortal users to transform datasets and create indices. GACS can provide a target for these user-added mappings and a vocabulary for ontology annotations.

LandVoc is a set of 270 terms about land governance created and maintained by by the Land Portal organization as a distinct concept scheme within AGROVOC. All lower-level concepts, such as "land degradation", have AGROVOC URIs; these are grouped under overarching concepts, such as "Land use management and investment", with Land Portal URIs. The scope of LandVoc is land governance in the broadest sense, with terms related to land tenure, government policy, land management, and urban planning. LandVoc is is integrated into the content management system of Land Portal, where it is used to tag all types of data on the website, from geographic data to statistical indicators. However, Land Portal wants to make the product available for use by the wider land governance community. Because land is a controversial topic, any standard is likely to be contested unless it is seen as a truly collaborative product. LandVoc is being enriched by partners who add missing terms, synonyms, and translations. One third of LandVoc terms are already mapped to GACS (via AGROVOC), so moving LandVoc onto a GACS basis is within reach. GACS concepts could help other organizations discover land-related terms for annotating their own data.

Crop Ontology, a SKOS concept scheme, gives identity to the phenotypes (traits) measured by agronomists and breeders of plants, such as plant height and grain weight. Grain weight is specified differently for crops such as wheat, rice, and bananas, because breeders of specific types differ with respect to methods and scales by which traits are measured. These terms are intended to be used in metadata as keywords for tagging files held in data repositories. Terms in the Crop Ontology, such as Wheat Grain Weight, are then related to a Trait Ontology, maintained by a collaborating institution, which identifies traits in a species-neutral way. In principle, terms such as Fruit Weight (Trait Ontology) and Grain Weight (Crop Ontology) could be mapped to Grain (in GACS), taking care to distinguish between grain as product or plant. Ideally, Grain Weight would also be added to GACS. Currently, most of the keywords used in metadata about datasets lack URIs; ideally, GACS URIs would link to Crop and Trait Ontology URIs, which would in turn link to datasets. In other words, search could be improved by leveraging mappings to search across both datasets and publications.

The **French National Institute for Agricultural Research (INRA)** has several types of semantic asset that need to be made more interoperable among themselves and with external resources in order to enhance discovery and reuse. GACS is interesting to INRA less as an indexing language, because resources such as NALT and AGROVOC are more complete, than as a hub of unambiguous and easily reusable URIs to which the following semantic assets can be mapped:

- **Thesauri**. A French-language thesaurus, VocInra, is used for manual indexing of scientific publications in institutional archives. VocInra needs to be mapped more extensively to other thesauri, such as AGROVOC, and by extension to GACS, in order to improve the visibility of resources in French. To avoid having construct additional mappings, the thesaurus team would rather have one hub, such as GACS, through which they could connect to other thesauri and ontologies, one-to-one-to-many, in order to improve interoperability with databases, tools, and services.
- **Domain-specific ontologies**. Domain-specific ontologies are developed and maintained by INRA researchers for use in decision-support systems and for annotating datasets. These ontologies are expressed in a variety of formalisms, such as OBO. If mapped to GACS, these ontologies could be made more discoverable and interoperable.
- **Reference lists**. Research teams at INRA maintain reference lists for use in databases about entities such as crop pests. If these reference lists were made available in RDF, they could be mapped to GACS, or their concepts could be incorporated into GACS. This would solve a problem for researchers who constantly need to refer to the same animals, crops, and diseases, never know which reference to use, and thus frequently re-do the work of adapting and extending parts of various existing resources. Having reference lists mapped to GACS would help INRA share their resources with industry and civil society.
- **Lexicons for text mining tools**. GACS, which overlaps in scope with INRA's own VocInra, can be used to augment lexicons with word patterns for mining text in specific domains. It would also help link textual content with relevant datasets. GACS is attractive in this context because it was formed from multiple thesauri and thus represents the agricultural domain in a broad sense.

The **Financial Industry Business Ontology (FIBO)** was presented not as a use case specifically for GACS, but as an example of how one industry consortium manages the development and maintenance of domain-specific ontologies from the standpoint of process. FIBO was designed to improve the interoperability and transparency of regulatory data reported by banks. It currently covers twenty-six domains, such as derivatives, indices, and loans, with a total of circa 70-80 OWL files, each with its own namespace, and a total of 680 classes and 528 properties. Each ontology focuses on the structure of a specific topic in great detail, using only a handful of properties and classes. Each domain ontology is developed by a content team of subject experts from member organizations, assisted by staff ontologists.

There are some dependencies among ontologies, particularly on a foundation ontology of concepts shared in common. Editorial changes are subject to integrity checks, then approved by leadership team that checks for possible conflicts among teams. This process is evolving to become more democratic and participatory, with proposals posted for review by members and ontologies held in Github, where they can be forked, branched, edited, and pushed back as pull requests.