Future of the Global Agricultural Concept Scheme: a report

Tom Baker
Version: 2017-01-18 (draft)

Purpose of the workshop

This report describes a workshop on the future of GACS that was held from 22 to 24 November 2016 in the context of the 9th International Conference on Metadata and Semantics Research (MTSR'16) in Goettingen, Germany. The goals of the workshop were to introduce the Global Agricultural Concept Scheme (GACS), a set of high-level concepts about agriculture derived from three major thesauri, discuss ongoing issues of policy and future directions for GACS, and elicit requirements for GACS from users and potential new stakeholders.

GACS was first envisioned in 2013 as a partial merger of three major agricultural thesauri: AGROVOC, CAB Thesaurus, and NAL Thesaurus. Their maintenance organizations -- Food and Agriculture Organization of the UN (FAO), CAB International (CABI), and the USDA National Agricultural Library (NAL) -- initially saw this task as a way to improve the semantic reach of their respective bibliographic databases and to achieve economies of scale by maintaining a joint concept scheme in collaboration. More broadly, GACS was seen as a first step towards improving the integration and semantic interoperability of agricultural information for the benefit of researchers, policy-makers, and farmers, with the ultimate goal of enabling innovative responses to the challenges of food security under conditions of climate change.

The creation of GACS began in 2014 with the normalization of all three thesauri to a W3C standard data model, Simple Knowledge Organization System (SKOS). Starting with three sets of 10,000 concepts, the GACS Working Group iteratively mapped frequently used concepts from the three thesauri, pairwise, and verified the mappings by hand. Vetted mappings were algorithmically checked for inconsistencies, which were resolved through discussion on teleconferences and in occasional face-to-face meetings. Each time mappings were manually corrected, GACS was iteratively re-generated, until the set of concepts was considered stable enough for publication. GACS was then automatically re-generated from its sources one last time and loaded into an editorial platform for future maintenance.

GACS concepts were assigned to one of five sub-types of SKOS Concept: *chemical*, *geographical*, *organism*, *product*, and *topic*. In addition to the standard semantic relations (*broader*, *narrower*, *related*), and mappings from GACS back to its sources, two properties were coined to relate organisms with associated products. Thematic groups originally developed by the three partners in the 1990s were implemented in GACS as ISO 15964 Concept Groups, a sub-type of SKOS Collection, to tag concepts for structured browsing.

GACS Core Beta 3.1, or "GACS Core", was released at the Open Harvest workshop in Chania, May 2016, with 15,000 concepts and over 350,000 labels in 28 languages. GACS Core was seen as a foundational step towards Agrisemantics, a future community network of semantic assets relevant to agriculture and food security originally outlined at a July 2015 workshop in Rome. At that workshop, Agrisemantics was pictured as "a well-integrated clearinghouse of machine-readable semantic assets in agriculture and nutrition, such as vocabularies, code lists, ontologies, taxonomies, and statistical indicators". In the context of Agrisemantics, GACS was to serve as a hub of high-level semantics, mapped 1-to-N to more specialized semantic assets.

The workshop in May 2016 took this idea further by envisioning that the high-level concepts of GACS would map to concepts defined, with more precision, in domain-specific ontologies which, in turn, would provide global identity and authority control for 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 described in this report, the original goal of developing GACS as a replacement for its three 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, suggested new directions for the idea of an interoperability hub for semantics in agriculture. At six months from the publication of GACS beta, the workshop in Goettingen was convened to take stock of developments to date, examine ongoing issues such as the role of hierarchical relations in GACS, and gather requirements from potential new stakeholders.

Discussion at the workshop examined the assumptions underlying GACS and considered alternative scenarios for its future. Much attention was paid to the nature of GACS as a *core* to be extended, and the handle "GACS Core" may change as a result. To avoid ambiguity, this report consistently refers to the currently published version of GACS, from May 2016, as "GACS Core Beta". Other types of semantic resource, such as code lists, reference lists, and ontologies are referred to collectively as semantic assets or knowledge organization systems (KOS).

The workshop in Goettingen aimed at eliciting feedback, requirements, and ideas about the future of GACS. Contributions to that discussion are freely paraphrased here, with attribution. For comparison with what participants actually said, a near-verbatim transcript of the 12-hour meeting may be found in Github. The positions presented at the workshop evolved over the course of the three days as discussion circled around a few key themes and participants revised their views. This account is meant to inform decisions on next steps in the development of GACS.

1. Opening remarks by Johannes Keizer

"Data" is not just statistics, but also text, images, videos, maps -- anything that can be digitized. The G20 and European Commission follow the principle that scientific data needs to be "FAIR": Findable, Accessible, Interoperable, and Reusable. Making data FAIR requires an infrastructure, or data ecosystem. The EU Project AgInfra, for example, was not just about semantics, but also about the tools, backend storage, and APIs required for such an infrastructure.

"Semantics" serve many purposes: indexing texts and information retrieval (thesauri), information organization (taxonomies), the interoperability of datasets (entity and code lists), machine learning, and knowledge-driven applications (ontologies), as well as text and datamining (lexicons). If semantics remain isolated in stand-alone silos, they cannot contribute to interoperability. However, simply mapping between ontologies, N-to-N, does not scale.

In order to link multiple ontologies, it is more practical to refer, with rough equivalence, to concepts in a hub vocabulary such as GACS. Such a shared vocabulary can serve as an authority for building new ontologies and knowledge organization systems. Any new concepts of global significance created in specialized KOS can, in turn, be contributed back to the common pool.

Any semantic relationships defined among these shared concepts are also inherited by systems that refer to GACS. Semantic relations intended to be of common use should be specified as lightly as possible, because overspecialization and overcommitment make concepts less commonly useful. This is also a reason why GACS should remain in SKOS, with its broad-brush semantics, and not be defined in a formally stricter ontology.

The pool of shared concepts in GACS can provide building blocks for constructing semantic applications. If those applications have additional concepts that are important for interoperability, then ideally, they would contribute these back into GACS, to fill gaps in the common pool, making those concepts available, in turn, as building blocks for semantic applications downstream.

2. GACS as seen by the GACS Working Group

As seen by the GACS Working Group, the strengths of GACS Core Beta are its manageable size; its availability with persistent URIs, in RDF and SKOS, as Linked Data; and its weak, non-overcommitted semantics. Its foundational use case -- indexing non- or semi-structured data -- makes it well-suited for use in information retrieval, but also for translation between languages, spell checking, user interfaces, data mining, and query expansion. The working group looks with interest towards new use cases, such as the semantic authority control of data elements in spreadsheets. In the meantime, the group would like to bring GACS Core Beta into conformance with thesaurus standards and best practice, through clean-up and quality control, and publish it as Version 1.0 under whatever name the GACS partners may decide.

In preparing the workshop, two contrasting views emerged in the working group about how GACS should be structured and governed in the future, though discussion in Goettingen showed the views to be more compatible than they first appeared. These views were:

GACS as a source of focused lists, each with its own namespace. A minority view held that the future GACS should emphasize a set of focused, highly curated lists exhaustively covering specific topics such as crops, pests, commodities, geographical places, agricultural practices, and livestock. Each list would be governed by its own editorial board of credible specialists, and each list would be free either to use the main GACS namespace, http://id.agrisemantics.org/gacs/, or to use its own namespace yet still be considered to be part of GACS. Relevant lists could be presented to users through interfaces that leverage Web services, mappings, and hierarchical information.

GACS as a distinct concept scheme with a single namespace. The majority view felt that the easiest story for GACS would be to follow the example of AGROVOC, NAL Thesaurus, CAB Thesaurus, Schema.org, and other semantic assets, which consist of distinct sets of terms in single namespaces. Accordingly, GACS would have a distinct identity as a concept scheme (as implied by the name Global Agricultural Concept Scheme). Anything and everything in the http://id.agrisemantics.org/gacs/ namespace would be considered "GACS", while everything else would be either in Agrisemantics or elsewhere on the Web. The GACS namespace would be maintained according to governance model first presented at the May 2016 workshop in Chania:

- **Core.** *Core* was meant as the name for the part of the GACS namespace under joint governance. With hindsight, GACS Core Beta appears to have been called *core* for two distinct reasons: 1) After July 2015, GACS Core Beta was envisioned as the central hub (*core*) for 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. Discussion at the workshop, summarized below, focused more on the former and questioned whether a concept scheme based on indexing terms was properly scoped to serve as the center of a semantic universe for agriculture. Pre-workshop discussion in the GACS Working Group, however, emphasized the latter, using *core* as a name for the part under joint governance, to which the partners might credibly make a long-term maintenance commitment.
- **Extension.** *Extension* was meant as the name for any parts of the GACS namespace outside of the *core*. At the workshop, *extension* was understood two distinct ways: 1) as a name for any semantic asset within the orbit around the *core*; and 2) as a name for any part of the GACS namespace not under joint curation, for example as a destination for concepts pruned from the core, or as

a place to put the 21,000 concepts of AGROVOC not included in GACS Core Beta (its "long tail"). As with *core*, discussion of *extensions* at the workshop focused more on whether it was appropriate to designate everything beyond the existing scope of GACS Core Beta as an extension. As discussed in the GACS Working Group, however, *extension* was simply a name for any part of the GACS namespace not under joint governance. *Core* and *extension*, in other words, were meant to be understood as parts of the GACS namespace under different governance policies.

- **Subvocabulary (Module).** *Subvocabulary*, or *module* -- also known as *concept group*, *remix*, *view*, or *profile* -- was meant as the name for any vocabulary on a specific topic, such as crops or livestock, based on a selection of concepts from GACS (both *core* and *extension*), potentially with concepts from other vocabularies. Such a module could be hard-wired into the definition of GACS, maintained separately from GACS, or layered over GACS as a link set. This notion of *module* is compatible with the vision of GACS as a source of focused lists.

- **Reflection.** *Reflection* was meant as the name for any set of concepts in the GACS namespace considered to be dependent on concepts outside of GACS to which they are mapped. This notion was introduced to support the longer-term goal of delegating maintenance responsibility for entire classes of GACS concepts, such as organisms, viruses, chemicals, or geographic names, to outside specialists. Assuming this would be done without deleting the mapped GACS URIs, then the question from a governance point of view is how dependent GACS URIs can be held in synch with "primary" URIs maintained outside of GACS -- a question of governance protocol between their respective editorial boards. Assuming the "secondary" (or "dependent") GACS URIs were not simply ignored, they would remain valid as aliases for the external URIs and could again be considered "primary" if the external authority were to disappear. A *reflection* would thus provide a way for the GACS maintainers to guarantee the persistence of concepts defined under DNS domains controlled by others.

2.1. Discussion: Core and extension

What we now call "core" is tied too tightly to the bibliographic use case. The concepts "GACS Core", "long tail", and "extensions" relate to a very specific and limited view of what has been done to date. We should not assume that GACS Core Beta meets 95% of the needs in agriculture just because it represents 95% of the terms used for indexing CAB Abstracts and Agris. There are applications for which the long tail might actually be more important than the core. GACS Core Beta may be a core for bibliographic indexing but not for agriculture as a whole. This is an intrinsic problem with the notions of core, long tail, and extension, not a criticism of specifically GACS Core Beta. I see GACS Core Beta as an upper-level concept scheme useful for generic description of agricultural publications but not specific enough for database compatibility. Whether we call this "GACS" or something else is not the main issue. The danger that I see in calling GACS Core Beta a *core* is that this implies that bibliographic databases are the most important application, as opposed to database columns or reference lists for an audience of, say, startups. If you treat a database as a publication, you can use GACS Core Beta to say that it is about "cows", and this fine for publications, especially non-scientific publications. But if you want datasets to be interoperable, in an operational sense, you may need a list of 27 cow breeds, or of different types of milking machine. Alot of databases are not interoperable simply for lack of an ontology to contextualize the column headings. What is the core that will support data interoperability for the analysis of, say, weather, air pollution, or precision agriculture? [Johannes]

The GACS Core Beta message should de-emphasize bibliographic indexing. If GACS Core Beta continues to be presented as the most frequently terms for indexing bibliographic references, that is how we will continue to think of it. In that case, there might be criticism that the core is not really reflecting science, but was simply a convenient way to create GACS. We must change the way we present GACS and envisage that the core may end up different from what it is today. [Elizabeth]

The notion of *core* is about governance and curation of a namespace. I see no reason that GACS Core Beta must remain tied to the bibliographic use case. Our analysis, based admittedly on automated mappings, suggested that the three thesauri shared about 13,000 concepts. This indicates that the concepts are in some sense global. I see no problem with adding new things to the core if they are deemed to be global, shared, and required by a use case, such as database interoperability. The notion of *core* was intended to be mainly about governance -- the shared part that should be governed in collaboration. [Osma]

A large namespace can (must?) have parts that are more highly curated. GACS can be a big, expanding pool of concepts, but there is value in being able to designate one part of that as curated with tighter quality control -- a manageable set to which people can turn for high-level vocabulary. To me, this is "core" in the sense "curated" or "guaranteed". If we were to open up the GACS namespace to input from many sources, it would be hard to scale the quality control, and it is quality that will make GACS attractive to use. [Tom]

GACS should be the part that is highly used and governed as authoritative. All content in semantic assets is in principle reusable, but some bits are actually reused more than others. The bits that are business-specific will be used the least. The highly used bits that are the ones that should be governed and given the stamp of authority. That part is GACS. [Derek]

Extensions help avoid redundancy and overlap within a namespace. Given a shared core, it is helpful to know that a given set of concepts (the *extension*) does not overlap with the core. In a decentralized landscape of more focused vocabularies, there would inevitably be overlap, and one might have to choose among multiple URIs for "wheat" (e.g., as a crop and as a commodity). With coordination, and with rules against introducing anything in an extension that is already in the core, such overlap can be avoided. [Osma]

GACS must be able to evolve. Science is always evolving as new concepts emerge and become frequently used, for example to annotate data. [Elizabeth] If GACS is to follow the natural evolution of science, it must be possible to promote some concepts and de-emphasize or deprecate others. If GACS consists of a large, even amorphous URI space with some concepts in a highly curated core (skos:inScheme), it must be possible to move concepts in and out of the more curated part without changing their URIs. [Tom]

GACS Core Beta cannot be pruned without a place to put pruned concepts. GACS Core Beta could be tightened, for example by pruning concepts that were brought in from the source thesauri simply to fill gaps in the hierarchy. Given the existing commitment to the persistence of GACS URIs, however, this could not be done without having a place to which the pruned concepts could be moved. We have been calling that place an *extension* in order to emphasize that the concepts there would not fall under the same curation guarantees as the *core*. [Tom]

2.2. Discussion: Reflections

By assigning URIs to non-semantic resources, GACS serves as a proxy. Most GACS concepts are based on authorities that are not available in RDF or SKOS, but only in "non-semantic" databases or printed books. By giving them URIs in the GACS namespace, GACS is serving as their proxy in the Linked Data space. As these external authorities themselves become semanticized, or published in RDF with persistent URIs, GACS should reach cooperative agreements with their maintainers, defer to the external URIs as "primary", and designate the GACS URIs as "secondary". This could involve granting access to the GACS editorial platform and letting the external maintainers create or modify concepts in the GACS namespace.

Large external sources should not be reflected. If large sources, such as databases on plant taxonomy or chemicals, are taken as reference authorities, they should not be entirely reflected with GACS URIs. Some people would prefer to use the source over the reflection. [Johannes]

Reflected URIs insulate GACS against changes in external namespaces. Let's say GACS were to reference alot of terms from an external namespace, for example with BT, NT, and RT relations. If that namespace were to change, alot of triples in GACS would also need to change. But if the external namespace were reflected with GACS URIs, GACS would not need to change. Indirection comes at a cost, but it buys you some insulation against changes. [Dean]

External namespaces routinely get replicated within corporate firewalls. For reasons of policy, control, or trust, external namespaces may get replicated within corporate firewalls (e.g., Bank of America and Syngenta) so that production systems need not load them from external sources. [Dean, Derek]

Even if external namespaces or domains disappear, reflected URIs remain. URI persistence guarantees, both real and perceived, are crucial for establishing trust in a vocabulary. One can credibly assert a persistence policy for URIs in a namespace that one owns, but not for URIs in namespaces owned by others. Without owning a domain, there is no way to ensure that it will not resolve, five years from now, to a click-bait site. If GACS URIs were considered "secondary" to "primary" URIs maintained by an external authority, and that external authority were to disappear, the GACS URIs would still remain. [Tom]

2.3. Discussion: Namespace

GACS could have multiple namespaces. If GACS Core Beta were considered just another extension, among others, with its own editorial board, and if GACS were crowdsourced in Github, then any community could propose its own vocabulary, with its own namespace, for approval by the GACS community. GACS would then consist of many namespaces, democratically, with no particular center. Each namespace would show its original provenance. This is doable in the Semantic Web. [Dean]

But multiple namespaces would be inherited by any derived vocabularies. Let's suppose GACS really did consist of seventeen namespaces. To create something like LandVoc, one might use half a dozen of those namespaces. The multiple namespaces of GACS would be replicated in any semantic assets derived from GACS. [Tom]

Webmasters hate multiple namespaces. Schema.org has one big namespace because, they say, webmasters hate to deal with the additional complexity of multiple namespaces. (Note: Dean subsequently found this argument compelling.) The story "one concept scheme, one namespace" is easy to explain and to sell, and URI persistence can more credibly guaranteed by the actual owners of a domain. Providing a space for other concepts schemes and namespaces was the original idea behind Agrisemantics, not GACS. [Tom]

3. Future directions for GACS

GACS should be the glue, or switchboard, between specialized ontologies. All concepts used to build classes in ontologies should be in GACS, and it is through GACS that these ontologies should be linked. GACS should be the glue between other, more specialized KOS -- a generic switchboard for concepts. GACS should not primarily be used directly, but as a source of building blocks for constructing application-specific semantics. A cow is a cow. It may be seen in various ontologies as a meat production thing, a milk production thing, or a disease-bearing thing, but it's still a cow. If those ontologies were all to refer to the cow URI in GACS, it would be easy to link them, even if they are not exactly the same. To me, the question is: What can we deliver as a common basis, to increase interoperability, with light semantics that leave people the freedom to adapt terms to their own purposes? [Johannes]

In contrast to ontologies, GACS can provide "identities". Ontologies describe observable individuals, while vocabularies such as AGROVOC and GACS provide identities, or universals in the Platonic sense -- things that are not observable directly but serve as tags that can be attached to observables. For example, "cow" is an identity attached to an individual of a particular population, used in a particular way. The individual is in my ontology, and the cow is in GACS. Individuals should not be in GACS. This dichotomy is crucial to understanding what vocabularies can do for ontologies and for interoperability. [Ferdinando]

GACS should recommend and endorse authorities to which one should map. To me, GACS is an idea: if you coin your own URIs, at least map them to existing URIs. The role of GACS should be to recommend authorities to which one should map, say, gene sequences or country names. These recommendations should be based on reliability checks that favor stable organizations such as CGIAR, FAO, and CABI. [Johannes] Ideally, people should not have to choose between two sources of URIs for the same domain, because having two URIs for the same plant breaks interoperability, even if they both refer back to GACS. Instead of reflection, a form of aliasing, GACS should provide endorsements. [Ferdinando] Endorsement should not be about persistence alone. Vocabularies should be publicly available at a stable location and reusable in accordance with the FAIR principles. [Johannes]

GACS can link to more specialized vocabularies. If there are multiple communities with good soil classifications, GACS could simply stop at "soil" and point off to the specialized soil vocabularies. [Brandon and Johannes]

Resources other than GACS can serve as sources of reusable concepts. I came to this workshop seeing a clear distinction between, on one side, GACS as a reference file for reusable concepts and relationships, and more specialized ontologies on the other. Most specialized ontologies are not reusable because they were designed for a specific purpose and do not apply for other purposes. However, their core concepts should come from GACS, which should consist of semantics that are global and reusable. But this may have been too simplistic, because resources other than GACS, such as Plant Ontology, can serve as reference files too. I see GACS not as an application to use, but as a reference base where people go if they want to create applications. So GACS does not substitute for any existing applications, but creates an interoperability layer for existing applications. In this sense, GACS is more a governance process than a physical thing. Either way, the message does not change: if you want to interoperate, then map your concepts to a reference resource. Over time, the body of semantic material available for reuse should grow. [Johannes]

GACS could serve as a glossary. Only one in four GACS concepts has a definition. [Osma] No concept can be accepted into FIBO without a definition; it is considered a syntactic failure. [Dean] Definitions should be required in GACS too. GACS could be more like a glossary, providing multiple natural-language definitions. [Brandon] On the other hand, with definitions from several domains, concepts can become unusable. [Caterina]

OWL ontologies can be made available in SKOS for simpler uses. FIBO has a process for converting a normative OWL ontology into a SKOS concept scheme, as an informative "shadow" of the OWL ontology, in a fairly automatic way. This is very useful, because data managers may just want a machine-readable data dictionary, or something to feed into their entity extraction tool -- and suddenly their data dictionary is aligned with FIBO. It is not always necessary to know how the structure of a swap is modeled in OWL, if all one wants to know is that there are three types of swap, with definitions that can be looked up. [Dean] At Bioversity we would also like to have an easy way to put OWL ontologies into a simple format for use as keywords. [Elizabeth]

GACS could be crowdsourced in Github. GACS could be extended with Github-style crowdsourcing: somebody (or even some other editorial board) could make a fork of GACS, edit it, then submit a pull request to have their changes folded into the master version of GACS. In this scenario, the owner of the GACS repo would have complete prerogative to impose quality criteria for pull requests, then to accept the requests or not. This approach raises questions of policy: whether forks should use the GACS namespace from the start, or use their own namespaces. If a pull request were accepted, should this trigger the creation of new GACS URIs, with exactMatch mappings to the source URIs? GACS would then turn into the sum of its pull requests, submitted by multiple editorial boards using multiple namespaces, though the GACS message with regard to trust and provenance would remain unclear. [Dean]

GACS Core Beta could just as one exemplar among others. GACS Core Beta could be recast as one example of the sort of artifact that GACS, as a group, wants to produce. It could have its own namespace, process, and a name like GACS Biblio (avoiding "core"). One could enumerate the requirements for other exemplars to become part of GACS -- for example, that it resource reconcile two or three existing ontologies. GACS Biblio could be consolidated as a core of concepts for that domain and a good exemplar of what GACS wants to provide. [Brandon]

GACS Core Beta should be cleaned up and finished as a high-level module. Even if it does not meet all requirements, the current GACS Core Beta does have some very nice use cases. It could be cleaned up, improved, and take its place as an important contribution to GACS, alongside others: a high-level subject vocabulary for agriculture with a maximum of, say, 15,000 to 20,000 concepts. The work plan would be to make this module (not *core*) more usable. Another module could focus, for example, on traits, because getting meaningful insights out of traits and genes is a global problem. [Johannes]

4. Structure and hierarchy in GACS

Osma Suominen presented the results of a survey about alternative scenarios for structuring hierarchical relations in GACS.

The twenty-six respondents use GACS primarily for "linking to GACS from my own concept scheme" and "supporting an information retrieval system". The respondents included some maintainers of GACS. Others expressed interest in helping to deepen GACS in the areas of earth science, forestry, geology, hydrology, land-related terms, plant traits, soils, and Portuguese terminology.

The respondents perceived the strengths of GACS to be: as a hub of important agricultural concepts, backed by solid organizations, collaboratively maintained, published according to good Linked Data practice, international and multilingual, and well governed, with a basis in reference thesauri. Perceived weaknesses included: its difficulty in reconciling different worldviews; gaps in coverage; organism taxonomies that will be difficult to maintain; inconsistent hierarchies; and that it tries to cover too much or is larger than it needs to be. Responses were in part contradictory, with a strength for one listed as a weakness by another. This prompted Johannes to remark that, with AGROVOC too, nobody was ever happy: it was either too broad and detailed, or not enough.

4.1. Polyhierarchy and top concepts

GACS inherited top concepts and hierarchical relations from three sources, with somewhat incoherent results. Clarifying the role of top concepts and cleaning up the hierarchy accordingly is seen by the GACS Working Group as the biggest task remaining before GACS Core Beta (however it will be called) can be published as Version 1.0.

With 578 top concepts, GACS Core Beta falls between AGROVOC and NAL Thesaurus (with 25 and 17 top concepts) and CAB Thesaurus (with 3,000). One quarter of the concepts in GACS have more than one broader term, and some have more than two broader terms (polyhierarchy). In the most extreme case, "casein" has seventeen different hierarchical paths to six top concepts: *substance*, *biology*, *product*, *science*, *characteristic*, and *property*. This means that if the hierarchy were seen as layers of is-a relationships, a common interpretation, one would in effect infer, nonsensically, that casein is a science.

The survey proposed three alternative scenarios for the GACS hierarchy and illustrated how a set of thirty GACS concepts would fit into three different scenarios:

- **Scenario A** is based on the General Finnish ontology (YSO), which in turn is based on DOLCE. On the very top level there are just three top concepts -- *objects*, *properties*, and *events* and *actions* -- that correspond roughly to nouns, adjectives, and berbs.

- **Scenario B** is heavily based on AGROVOC, which has twenty-five top concepts that function as types. What comes below can be said to be a "kind of" whatever comes above. This hierarchy is more shallow than in Scenario A because the concepts are more domain-specific.

- **Scenario C** is based largely on the 1999 CAB Classified Thesaurus and therefore on the thematic groups in GACS. The idea here is that most of today's messy hierarchy would be discarded, leaving only a very small hierarchy that is probably shared by the source thesauri for things against which nobody can argue. The result would be many top concepts; indeed, many of the thirty example concepts would find themselves at the top of the hierarchy. This hierarchy would be very flat except for things that can reasonably be organized in a hierarchy, such as elements, organic compounds, and animals. The structure of GACS would be provided more by thematic groups, which put concepts into clusters, but not in a way that is usable for indexing. This scenario involves some repetition, with "diseases", "breeding", and "aquaculture" present, with two separate names, as both concepts and thematic groups. Scenario C has no high-level concepts of an abstract nature; the concepts here are more concrete.

Voting results. Scenarios A and B both got relatively high marks, the most common criticism of Scenario B, by far, being that certain key distinctions are unintuitive, for example between *phenomena* and *entities*. Opinions about Scenario C were split between those who hated it and those who liked it alot. This split was due perhaps in part to the combination of two quite different ideas: that of discarding most of the hierarchy, and that of using a thematic approach. When asked "If I had to design a hierarchy...", people responded that they preferred A; A and B; A with thematic classes like the top concepts of B; just B; B but incorporating levels of C; just C; or no hierarchy at all -- pretty much every combination. The thematic groups are considered "more useful than useless", with many positive comments but also some criticism of artificially limiting concepts to just one group.

4.2. Fixing the structure of GACS: a proposal

Respondents suggest that complementary approaches be combined, though GACS does this already, with its hierarchy, concept types, and thematic groups. Not all purposes are best served with hierarchy. Thematic groups can be more useful for providing an overview, while hierarchy can be more useful for disambiguation. In light of the survey, Osma and Tom suggest:

1. GACS should keep the thematic groups as an additional view, possibly with some corrections and tweaks, such as: allowing concepts to belong to multiple groups, classifying the 20% of concepts that are not currently classified, and perhaps dissolving the category *general*.

2. The current GACS concept types -- *organism*, *chemical*, *geographical*, *product*, and *topic* -- should perhaps be converted into top concepts. Some of these are

top concepts already. This would mean dropping the notion of concept types in order to avoid encoding the same information two different ways. A new type, *property*, could be split off from *topic*. An additional layer of organization could be added below these top concepts, especially for differentiating among the large number of topics.

3. The hierarchy should be cleaned up by reducing unwarranted polyhierarchy. This would involve flagging and correcting situations where broader and narrower terms have different concept types (circa 1,300) and moving some leaf concepts out of GACS Core Beta. The option of eliminating the hierarchy entirely, leaving developers to design their own, is not attractive because some obvious hierarchical relations are helpful for disambiguation. For example, it seems unlikely that anyone would disagree that *nitrogen* is narrower than *non-metallic element*.

4.3. Discussion

Semantic relations help define concepts. Less than 3,000 of the 15,000 concepts in GACS Core Beta have definitions. If one were to remove too many semantic relations, that would leave only the labels. With just a label, and no definition or hierarchical context, it could be difficult for an Arabic or Russian speaker to know what a concept means. [Tom, Osma]

Is polyhierarchy really a problem? I'm not sure I see polyhierarchy as a problem. The multiple hierarchy chains for "casein" help me understand the contexts in which it can be used. [Dean] These are just association paths, which are fine. [Ferdinando]

Modeling types with sub-categories of SKOS Concept is good. I approve of modeling types with sub-categories of SKOS Concept. The five types very efficiently convey what GACS is about, and simply knowing that organisms are associated with products also tells me alot. The hierarchy can be a facet that is orthogonal to that. This allows you to say, for example, that an organism is related to a product. Doing this with rdf:type strikes me as the clean way to do it. [Dean]

Policies on hierarchy must be simple enough to apply consistently. The point is not whether existing polyhierarchical relations are helpful, but whether they can realistically be applied and maintained, as a matter of policy. Hierarchical relations are most useful when one can trust that they have been applied consistently and exhaustively. To achieve this, the policies for hierarchy would need to be simple and clear. [Tom]

Allow hidden concepts to machine-learn themselves into hierarchies. Do we really need to commit to a hierarchy? Whatever hierarchy is chosen, it will leave out alot of use cases today and will definitely need to change over time. Or is the responsibility of this group simply to make many other hierarchies possible? Perhaps there could be a "dark core", not exposed to an API, with minimal redundancy and maximum connectedness. The concepts there could be free to associate with other, to machine-learn themselves, and cluster themselves into hierarchies. On top of this, one could expose hierarchical views that fit particular types of usage. [Ferdinando]

To be reusable, and FAIR, hierarchy must not be too specific. I would like to apply FAIR principles to the discussion of hierarchy. If there is absolutely no hierarchy, only a pool of concepts, then the concepts are hard to find and access. As for interoperability: the more you commit to specific hierarchies, the less reusable the concepts become. One should commit to just enough hierarchy to make it accessible. The hierarchy should not be so specific that it cannot be easily reused. [Johannes] But hierarchy can enhance reuse, and disambiguation enhances reuse, so if hierarchy helps us disambiguate, it enhances reusability. [Dean]

Enhancing hierarchy with richer relationships has limits. We could perhaps address the limits of hierarchy by introducing richer relationships. [Brandon] AGROVOC uses more specific relationships, though not very systematically, and many are used very few times. If you want to do it, then I say: do it properly. And make sure to use the right relationship every time, and not just occasionally. That is the challenge if you go for this more complex model. [Osma] The "ontological" hierarchy defined for AGROVOC was not constructed systematically; we never had the intellectual capacity and the financial capacity to do this properly. It grew organically. [Johannes] The NCI cancer ontology goes even further by adding a dozen or so specific relationships, for example "occurs in", as in: "this gene occurs in this organism". NCI encodes alot of scientific knowledge this way, though it is alot of work, and it is unclear whether this way of doing science yields sufficient value. [Dean]

Speak not of hierarchy, in the singular, but of hierarchies. The hierarchy of Geopolitical Ontology is strictly political, but one could have hierarchical views based on continents, or GDP levels, or type of climate. As long as such hierarchies are useful, and can be reused, it could be useful to put them into GACS. [Johannes]

Distinguish community views from views hard-coded into GACS. We should differentiate between relationships we try to apply exhaustively, according to maintenance policy, and user-generated views, which could be layered over GACS as optional link sets. User-generated views may be very interesting, but if the core maintainers of GACS are not in a position to commit to their maintenance, they could be abandoned and become obsolete. [Tom]

FIBO, converted from OWL into SKOS, has a flat hierarchy, with definitions. Why not just list concepts alphabetically, with definitions, and with relationships that provide a machine-readable way to understand what they mean. That's basically what we wind up having when we take the OWL version of FIBO and turn it into SKOS. The SKOS version has a strict is-a backbone, because it was converted from OWL, but it is seriously flat. Everything has a full definition, and every definition has a source, such as Barron's Law Dictionary or InvestiPedia. We are fanatical about citations. It is not just the underlying logic that helps disambiguate, but also the prose and the provenance. [Dean]

References

[1] https://github.com/agrisemantics/2016_11_goettingen/blob/master/minutes

[2] https://github.com/agrisemantics/2016_11_goettingen/blob/master/presentations

[3] https://github.com/agrisemantics/2016_11_goettingen/blob/master/mtsr_paper

Appendix: Use cases and best practice

AgroPortal (Anne Toulet, LIRMM)

AgroPortal is a searchable repository of ontologies in the agronomic domain, based on the same underlying technology as the NCBO Bioportal. 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 (Lisette Meij, Land Portal)

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. One third of the AGROVOC concepts in LandVoc are already mapped to GACS. The scope of LandVoc is land governance in the broadest sense, with terms related to land tenure, government policy, land management, and urban planning.

Land Portal is a global platform. Its partners serve as intermediaries at a country or regional level, producing information on very local issues and sharing it on the Web.

LandVoc 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. A Gap Analysis carried out in the context of a GODAN Action partnership found little use of standards in the land sector; systems often use uncontrolled tagging or even lack any sort of metadata. LandVoc will be published as part of the next release of AGROVOC with an open-source license.

Land Portal may establish a task force with terminology experts to help maintain the vocabulary and ensure consistency. This could involve providing more than one definition per term. For example, "land ownership" can mean many different things depending on the legal system in which it is used. They are currently looking for terminology experts that have written about land and worked on land extensively, from different sectors, who speak different languages, and who understand the controversies around this subject. Land Portal would coordinate the process and provide technical expertise for managing the vocabulary.

Land Portal also wants to do capacity building around LandVoc in the form of tutorials and webinars, starting with awareness of standards and how they can be used. Land Portal is targeting regional policy makers to advocate for open data. The next step would be to follow up with webmasters and librarians who are processing information directly.

Land Portal wants to remain as independent as possible because land is very controversial. To get a widely used tool or standard is difficult because any proposal is likely to be contested. In order for the vocabulary to be widely used, Land Portal believes it will have to make compromises. It will also have to avoid claiming that LandVoc is Land Portal's vocabulary. Rather, it needs to be the product of a collaborative process.

Land Portal sees two scenarios for future development. In one scenario, LandVoc would remain within AGROVOC, build up the hierarchy, and top-level concepts would be managed in their own content management system. In the other scenario, LandVoc would become an independent vocabulary, managed completely by Land Portal, with its own namespace. This would be ideal from the standpoint of Land Portal, but not necessarily ideal for the land sector, because Land Portal does not have the same perceived stability as AGROVOC (and its URIs).

Currently, the LandVoc concept scheme draws its concepts from locations scattered throughout the AGROVOC hierarchy. Where a parent and child concept are both in AGROVOC, LandVoc also includes their BT/NT relationships. For the remaining terms, Land Portal wants to fill out their hierarchical context with LandVoc relationships -- a task that seems doable given LandVoc's small size. Top-level concepts are also being added.

All lower-level concepts, such as "land degradation", have AGROVOC URIs. Through a collaborative process, they have created "Land use management and investment" as an overarching category with a Land Portal URI. **If another organization, such as Bioversity, were to need a Land Portal term for annotating data, the term could perhaps be discovered by way of GACS.**

LandVoc is being enriched with help of partners, who add missing terms, synonyms, and translations. For example, a Global Land Indicator Initiative has created a list of of terms with definitions. Land Portal has added these to AGROVOC as synonyms. Collaborating partners in the Mekhong region have translated all 270 terms and definitions into Khmer, Thai, Vietnamese, and these have been folded back into AGROVOC via the LandVoc concept scheme. Land Portal is adding value to such initiatives by making them available in RDF vocabularies.

The collaboration between LandVoc and AGROVOC follows AGROVOC editorial guidelines. AGROVOC began as something edited by thesauri experts, but now the editing has been distributed to a wider circle of domain experts. A proposal is submitted, the AGROVOC team checks whether the proposal appropriate. If so, the accepted proposal goes into the VocBench editorial environment. Once the terms are in VocBench, they can be linked into the AGROVOC hierarchy. Translations are submitted as Excel spreadsheets. The value proposition, from the standpoint of LandVoc, involves contributing expertise on land and and benefitting from the continuous updating and enrichment of AGROVOC terms with links to other agricultural vocabularies.

One third of LandVoc terms are already mapped to GACS (via AGROVOC), so putting LandVoc onto a GACS basis is within reach.

Crop Ontology (Elizabeth Arnaud, Bioversity, CGIAR)

Phenotype (trait) data from plant breeders describes germplasm varieties according to variables such as plant height and grain weight. The abbreviations used for these variables in the Excel tables are quite cryptic, such as "GW" and "PH". Bioversity has collected the traits measured by breeders and agronomists, along with their abbreviations, and given each trait a unique identifier in the context of a Crop Ontology, a simple concept scheme in SKOS that is available online. These terms distinguish between various crops, with Grain Weights specifically for crops such as wheat, rice, and bananas, because breeders of specific types differ with respect to methods by which traits are measured. The definitions for these terms in Crop Ontology specify recommended measurement methods and scales. Such terms are intended to be used in metadata about the data files held in repositories, typically as keywords.

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. Crop-specific terms from the Crop Ontology are added to the Trait Ontology, and the terms of Plant Ontology are mapped to the terms in Trait Ontology as automatically and as precisely as possible (preferably with exactMatch). Terms from the Crop Ontology and Trait Ontology are merged on the Planteome portal, which displays the crop-specific terms under the species-neutral terms and provides access to data files annotated with those terms.

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. As there is currently no term for grain weight in AGROVOC, one can currently search in Agris for publications about wheat that refer to "grain weight" in their title or abstract. This search could be improved by extending the result set to include, for example, wheat yield data in the CIMMYT repository that has been annotated with Wheat Grain Weight. AgTrials, a global repository of evaluation trials, has annotated their data files with variables from the Crop Ontology, and adding Wheat Grain Weight as a variable to a query will retrieve datasets from many institutions. *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.*

From the standpoint of CGIAR, it would be helpful if Agrisemantics could help ensure that multiple authorities on specific topics, such as soil, work together to produce a reference ontologies that could then be incorporated into GACS.

French National Institute for Agricultural Research (Sophie Aubin, INRA)

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. In the context of Analysis and Experimentation on Ecosystems (ANaEE), a European project, INRA constructed a thesaurus for improving access to analytics, observations, and datasets, and mapped this thesaurus to sources such as AGROVOC and GEneral Multilingual Environmental Thesaurus (GEMET). 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 biological 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. Such processes 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.

From the INRA point of view, GACS should avoid a strong commitment to is-a relationships, which can be objects of contention. Rather, the hierarchy should be as light as possible: enough to help disambiguate, or perhaps to help navigation, but not more. Facets, thematic groups, and other views should be used for exracting sets of needed concepts.

Financial Industry Business Ontology (Dean Allemang, Working Ontologist)

The Financial Industry Business Ontology, or FIBO, is owned by a consortium of circa 150 members, primarily banks. FIBO was built in response to the Basel Commission on Banking Supervision document 239 (BCBS 239), which determined that the 2008 financial crisis was caused in part by problems of data visibility. FIBO was designed to improve the interoperability and transparency of regulatory data reported by banks. The process of creating FIBO began in 2010, when an enterprise architect interviewed experts in the banking industry and distilled their knowledge into UML models of twenty-six domains, such as derivatives, indices, and loans. OWL ontologies were developed within each domain for a total of circa 70-80 OWL files today, each with its own namespace, with 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.

The development of domain ontologies is carried out by a content team of subject experts from member organizations. This development is coordinated by a governing council, the Enterprise Data Management Council (EDMC), with a full-time project manager, full-time director, and three part-time consulting ontologists, who provide technical support and OWL modeling expertise. One key domain, Foundations, provides general concepts that cross-cut the other domains, such as "contract parties", "money", and "commitment". Domains often depend on other domains: Derivatives relies on Indices, Loans relies on Business Entities, Debt relies on Loans, and all domains rely on Foundations. Note that although the Foundations module has a central role, it has never been seen as a "core" in relation to "extensions"; they are more like different chapters in a book. The work has been split up into these pieces more for reasons of practical organization and governance.

Editorial changes are currently approved according to extensively documented processes. Changes to ontologies are subject to automatic tests related to OWL logic, integrity ("every class must have a definition"), and the correctness of URIs, then rubber-stamped by a leadership team which checks for possible conflicts with other 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. Versioning is provided by Github and by an in-house ontology history tool. Only four or five of the twenty-six domains actively work on their ontologies at any given time. Planning for the next twenty domains is outlined in a roadmap.

The ontologists attached to content teams spend most of their time on infrastructure, for example to implement sanity checks. Agile proof-of-concept teams are formed if working systems need to be deployed quickly. Much energy goes into perfecting OWL restrictions and cardinalities in the ontologies; in hindsight, SKOS may have sufficed, and having a simpler model might have made it easier to broaden the scope of the ontologies.

The FIBO model resembles how data description is handled at Syngenta, where work is divided into domains that are aligned with areas of R&D science and tightly focused on business deliverables such as chemical invention and trait discovery. Syngenta has no in-house equivalent to the Foundation ontology, which cross-cutting domains such as environment, in part because no one group wants the extra work.