Global Agricultural Concept Scheme (GACS):

A multilingual thesaurus hub for Linked Data

Thomas Baker, Sungkyunkwan University, Korea
Osma Suominen, National Library of Finland, Finland

Version 1.0. August 12, 2014

Introduction

This paper proposes next steps towards the creation of a Global Agricultural Concept Scheme (GACS) as a hub for thesauri in the agricultural field, in multiple languages, for use in Linked Data. The idea for GACS emerged out of discussions at the World Congress of IAALD[1], the International Association of Agricultural Information Specialists, in July 2013. The Food and Agricultural Organization of the United Nations (FAO), CAB International (CABI), and the National Agricultural Library of the USA (NAL) agreed in October 2013[2] to explore the feasibility of developing a shared concept scheme by integrating their three thesauri: the AGROVOC Concept Scheme[3], the CAB Thesaurus (CABT)[4], and NAL Thesaurus (NALT)[5]. In the GACS vision, the integration of these three thesauri is but the first step towards the realization of a hub that links to and from the concept schemes beyond the initial three, and in multiple language areas, perhaps in the context of a global consortium.

The GACS project has three phases:

- **Phase One: Feasibility Study (June 2014)**. FAO, CABI, and NAL commissioned a report on the status quo of each thesaurus. The report, "GACS: Status quo of three partner thesauri"[6], is summarized in Section One below and forms the basis for this proposal.
- **Phase Two: GACS Beta (this proposal)**. The next step will be to create and iteratively refine a new concept scheme mapped to and from a selection of circa 10,000 frequently used concepts from AGROVOC, CABT, and NALT. Candidate mappings for this shared set of concepts will be generated automatically, then verified manually by experts. This beta version of GACS will have an identity and global identifiers (URIs) separate from those of its sources, and the three organizations will pledge to keep this resource available for the long term under the terms of a Creative Commons license. The project will implement a distributed, Web-based editorial environment to support the maintenance of concepts by multiple language communities. GACS Beta will be published on a Web platform that supports user-friendly browsing and easy access to machine-processable Linked Data.
- **Phase Three: GACS 1.0 and beyond (future)**. If resources are available for building GACS into more than the circa 10,000 concepts of GACS Beta, the GACS editorial team will develop an integrated semantic super-structure for the concept scheme, collaborate with concept providers on naming things like species, viruses, and chemicals with Linked Data URIs, and extend the GACS partnership to additional organizations and partner thesauri.

[1] http://iaald.library.cornell.edu/
[2] http://aims.fao.org/community/agrovoc/blogs/national-agricultural-library-usa-cabi-and-fao-agree-collaboration-developme
[3] http://aims.fao.org/standards/agrovoc
[4] http://www.cabi.org/cabthesaurus/
[5] http://agclass.nal.usda.gov
[6] https://github.com/tombaker/gacswg/blob/master/GACS_Status_Quo_0.99.pdf

1. Phase One: Feasibility Study

The extent of overlap between the three thesauri was roughly estimated using an ontology matching algorithm that aligned labels having closely matching strings. With the caveat that no additional manual checking was used to verify the accuracy of the mappings, it appears that some 13,000 concepts are shared by all three thesauri and an additional 30,000 concepts are shared by at least two of the three (see Figure 1).

The status quo analysis also showed that more than two thirds of the concepts in the three thesauri refer to species. Two other salient and clearly distinct categories of concepts in the thesauri are chemicals and places. An analysis of the concepts most frequently used in AGRIS, a database indexed using AGROVOC, showed a strong "long tail" distribution: 80% of the records were indexed using less than 2,000 of the AGROVOC concepts, 90% with less than 4,000, 95% with 6,000, 99% with 11,000 (see Figure 2). The results of the status quo analysis suggest a methodology for integration whereby candidate mappings are generated automatically and checked manually.

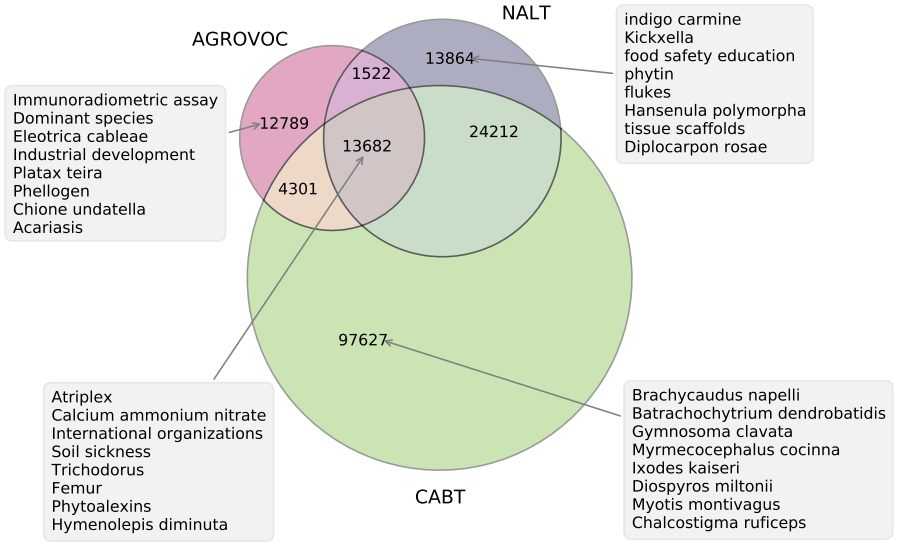

Figure 1: Overlap of participating thesauri.

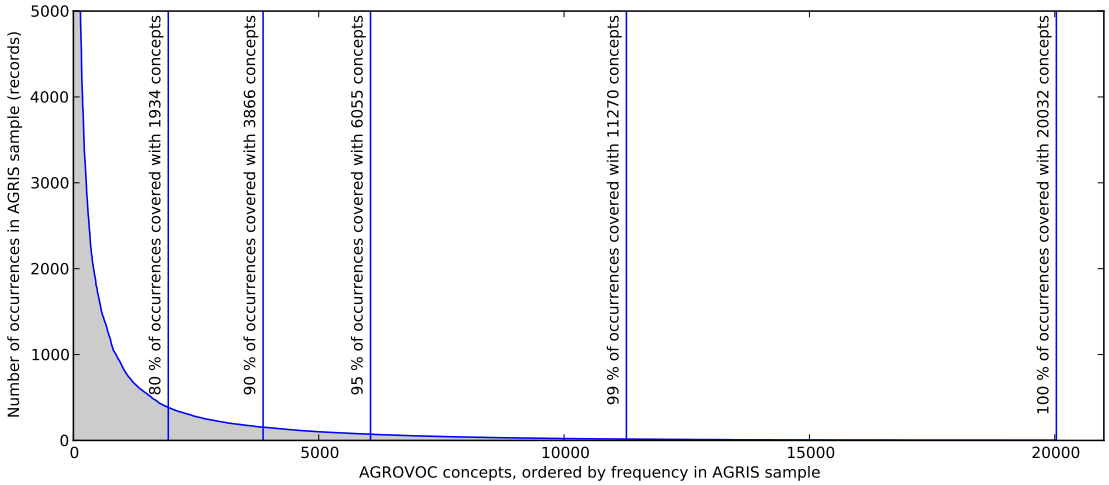

Figure 2: Long tail distribution of AGROVOC concepts in the AGRIS database.

2. Phase Two: GACS Beta

2.1. Objectives and scope

The three partners will integrate their thesauri through mapping to a new concept scheme, "GACS Beta", according to a hub-and-spoke model (Figure 3). GACS Beta will be expressed in SKOS as Linked Data, on the Web, with an open license that allows reuse. GACS Beta will not replace the three partner thesauri. Rather, the three thesauri will continue to co-exist with their different topical strengths targeted to different audiences. GACS Beta will constitute a real, usable, and durable product – something that demonstrates the added value to be achieved through the integration of three thesauri – with which to build an argument for an extension of the concept scheme.

GACS Beta will hold a set of circa 10,000 concepts selected primarily according to frequency of use. Some small but distinct sub-sets of concepts, such as countries, will be covered in their entirety.

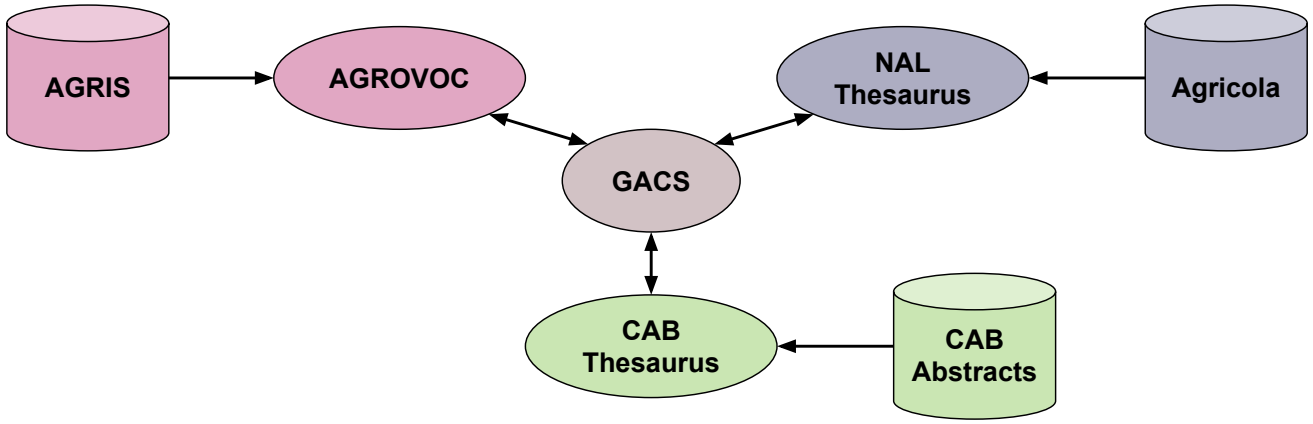

Figure 3: Overview of GACS Beta, its relationship to the participating thesauri via mappings that form a hub-and-spoke model, and the main publication databases currently referencing the thesauri.

2.2. Methodology and deliverables

The process of integration will start with an analysis of the major publication databases associated with the three thesauri (see Figure 3) to determine the pool of concepts from each thesaurus that are most frequently used for indexing. A label-matching algorithm will generate candidate pairwise mappings of these frequently used concepts among the three thesauri. The candidate mappings will be evaluated and approved manually. Concepts not automatically mapped will be manually evaluated for less obvious matches.

Concepts with confirmed mappings to and from the three thesauri will constitute the initial GACS Beta. GACS Beta will be partitioned into modules along the lines introduced in the Status Quo report: Species, Chemicals, Places, and Other. The integration of labels and semantic relations from the source thesauri will be handled automatically. This process is described in more detail in the Appendix.

2.3. Required resources and infrastructure

Phase Two will involve the manual and intellectual effort of partner organizations in the form of verifying and approving the correctness of candidate mappings.

Were funding not secured for deepening GACS, GACS Beta would be maintained on an ongoing basis by existing staff at the three contributing organizations as a function of their regular thesaurus maintenance tasks. Inasmuch as GACS Beta would initially hold the most commonly used and stable terms of the three thesauri, it is expected that once this set were created, the

maintenance burden would be limited to maintaining the mappings to and from GACS Beta and pulling updated labels, semantic relations, and information about the source of concepts automatically from the source thesauri.

The infrastructure required for GACS Beta consists of:

- a set of automated processes to generate candidate mappings for approval by thesaurus editors, algorithmically reconcile approved mappings by looking for suspicious mapping patterns, and flag potential errors.
- a distributed, Web-based editorial environment that will allow thesaurus editors at the three organizations to work in parallel on editorial corrections and enhancements to the shared concept scheme;
- a publishing platform that will allow users to explore the hierarchical and associative relations between concepts and view or download RDF representations of the concept scheme.

An initial investment will be required to set up the server and publication environments and to create and test the suite of automated tools for mapping and quality control. Inasmuch as the three organizations have committed to keeping the results of Phase Two available on the Web for the long term, the server infrastructure will be maintained.

3. Phase Three: GACS 1.0 and beyond

3.1. Objectives and scope

Moving beyond the initial set of mapped concepts would involve engaging more thesaurus maintenance organizations and concept providers as partners and by focusing on issues beyond those of a semi-automated mapping exercise.

3.2. Methodology and deliverables

Expansion of GACS Beta. The set of concepts in GACS could be expanded beyond the original set, either by taking on more concepts from the original source thesauri, or by pulling concepts from external providers into GACS. The addition of new concepts would also require that they be mapped to all the source thesauri. The methodology would be similar to that used to create GACS Beta.

Curated hierarchy. GACS Beta will initially be formed of concepts without much regard for the hierarchical context of the partner thesauri in which they are embedded. For GACS, the GACS Working Group could provide a common semantic superstructure for GACS in the form of a principled set of top concepts.

Towards ontology. Semantic relationships between concepts beyond the standard thesaurus relationships of broader, narrower, and related may be considered “ontological” when they specify how the real-world entities associated with concepts relate to each other. For example, a CABI compendium may say that Organism A “is a pest for” Plant B. AGROVOC uses several dozen “ontological” properties that could be evaluated for use in GACS. A judicious use of ontological classes and relationships could enhance the usefulness of GACS by making the real-world nature of its concepts more explicit, with the caveat that such enhancements increase the collective maintenance burden. Similarly, the nature of SKOS concepts could be more explicitly specified, whether by declaring them to be instances of more specific classes or by grouping concepts of a specific type under top concepts, sub-vocabularies, or separate concept schemes. Selecting a model for doing this will require further study of emerging practice.

3.3. Required resources and infrastructure

In addition to the resources required for the ongoing maintenance of infrastructure – the publication platform, editorial platform, and automated processes – the creation of a common semantic framework and addition of “ontological” semantics would require a deeper level of engagement among the participating partners, possibly requiring more sophisticated collaboration processes and workflows.

4. Policies and governance

RDF model. The SKOS model used for GACS will need to meet requirements such as the expression of provenance, the versioning of entities subject to historical change (such as countries and taxonomies), and advanced requirements for multilinguality.

URI policy. GACS will use a namespace URI different from those of the three participating thesauri. To achieve credibility for GACS, some organization, whether one of its partners or GACS itself (if constituted as an independent entity), would need to make a public commitment to the long-term persistence of its URIs and availability of related documentation.

Open access license. GACS will be made available on the Web under the terms of a Creative Commons license for open access. The license could be either CC0[7], which encourages reuse by not placing any restrictions on the data (similar to public domain works), or CC By[8], which requires attribution and is thus more restrictive. The licensing of translations used in GACS would need to be clarified.

Integrating other partners into GACS. GACS could partner with organizations or initiatives that following overlapping goals, such as the Chinese Academy of Agricultural Science[9] or the CGIAR Consortium[10]. GACS could also collaborate with specialized concept providers such as the Catalogue of Life[11], the Angiosperm Phylogeny group[12], or the International Committee on Taxonomy of Viruses (ICTV)[13].

Governance. Initially, GACS can be maintained by a standing committee of thesaurus experts that includes representatives of all constituent thesauri. Potentially, GACS could be constituted as a maintenance entity independent of the partner organizations. A consortium model would allow other partners, such as CGIAR and CAAS, to participate as project members. Decisions on mappings, associative relationships, and the hierarchical structure of the shared concept scheme could be delegated to a GACS editorial board. Responsibility for GACS as a sustainable endeavor could shift to a broader range of stakeholders.

5. GACS technical infrastructure

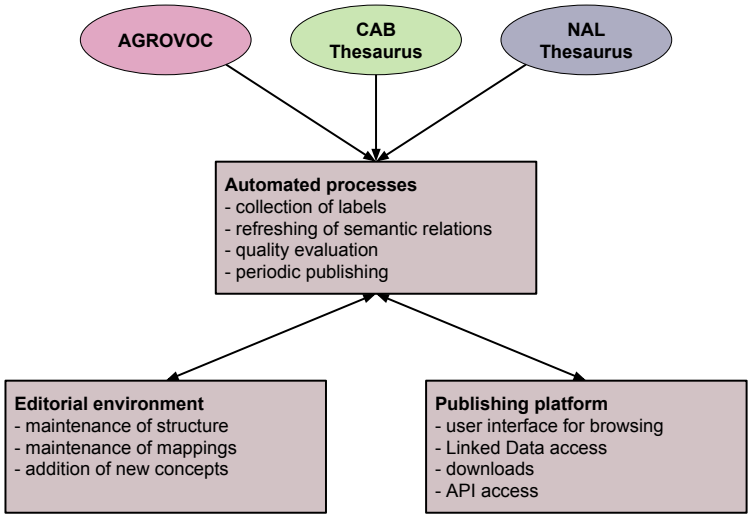

Figure 4: Overview of GACS technical infrastructure

The technical infrastructure required for maintaining GACS consists of three main components installed in a suitable server

[7] https://creativecommons.org/publicdomain/zero/1.0/
[8] http://creativecommons.org/licenses/by/4.0/
[9] http://www.caas.cn/en/
[10] http://www.cgiar.org
[11] http://www.catalogueoflife.org/
[12] http://www.mobot.org/MOBOT/research/APweb/
[13] http://www.ictvonline.org/

environment: a distributed editorial environment, a Linked Data publishing platform, and a set of automated processes for data management. An overview of these components is shown in Figure 4.

Distributed editorial environment. While GACS Beta can be created semi-automatically and refreshed automatically, any additional editorial work beyond such semi-automated maintenance, such as the assertion of GACS-specific hierarchical relations, the creation of top concepts, the coining of new concepts directly in GACS, or the additional or labels in multiple languages, requires that GACS have its own editorial platform. This platform needs to be Web-based, allowing multiple simultaneous users. The GACS editorial environment should support the distributed maintenance of a global concept scheme based not just on English and translations from English. Rather, it should allow concept growth to be driven by the needs of partners irrespective of language. In Phase Two, the GACS project will use VocBench[14], a multi-user platform for the maintenance of concept schemes.

Publishing platform. In Phase Two, the GACS project will use Skosmos[15], an open-source SKOS publication platform maintained by the National Library of Finland.

Automated processes. A set of GACS-specific processes would need to be created to update the GACS with the newest labels, translations and semantic relations from the source thesauri. The quality evaluation tools Skosify and qSKOS will be used to point out potential problems in GACS, which are likely to occur when information from many sources is automatically merged. The regular publication of new versions of GACS could also be automated.

6. Benefits of GACS

Strategic benefits of the global partnership

- **Provides a global public good**. A concept scheme for the agricultural domain, published in human-friendly as well as machine-readable form under the terms of an open license, will constitute a global public good.
- **Promotes re-use of domain knowledge**. Publishing the global concept scheme as Linked Data will make the resource available for re-use by anyone with a connection to the Web and generic tools for parsing SKOS data in RDF-compatible formats.
- **Enhances relevance and recognition**. By creating a prominent, harmonized concept scheme for the agricultural domain, the participating organizations position themselves as the leading experts in agricultural information management.
- **Builds Linked Data expertise in partner organizations**. GACS will expose the people working in each of the partner organizations to modern information management technologies such as SKOS, RDF and Linked Data. These skills can then also be used internally within those organizations for planning and implementing the next generation of information systems.
- **Avoids duplication of effort**. For thesaurus teams that currently work largely in parallel, GACS will provide a common goal in the push to open data.

Efficiency of information management

- **Provides a single target for mapping to and from thesauri beyond GACS**. GACS will provide a common target for mappings not just from the three thesauri, but from other, related resources both in the agricultural field and of more general scope, such as Eurovoc, DBpedia, or Geonames.
- **Provides common source for enriching partner thesauri.** Each thesaurus will contribute concepts, labels, and relations to the shared GACS concept scheme. In turn, partner organization can pull concepts, labels, and relations from GACS back into their own thesauri.
- **Serves as a channel for pooling translations.** Labels in all the 22 languages having substantial coverage in the source thesauri will be pooled to GACS, from which they can be pulled back to the source thesauri, significantly extending the linguistic coverage of all source thesauri for their most important concepts.

[14] http://aims.fao.org/tools/vocbench-2
[15] https://github.com/NatLibFi/Skosmos

- **Provides a basis for specialization among the maintainers**. If GACS provides a pool of concepts, labels, translations, and relations from which all draw, the GACS partners could divide maintenance tasks among themselves. For example, one organization could assume responsibility for maintaining part of a taxonomy.
- **Improves consistency and reduces variability**. Communication among partners about specific issues of semantic compatibility will improve the quality and consistency of the shared concept scheme.

Discoverability and interoperability of information

- **Promotes a common model of the agricultural domain**. The terms of discourse defined by GACS will provide researchers and practitioners with a common model and language for the agricultural domain that will likely be followed by information providers and improve the coherence of agricultural information globally.
- **Improves automatic indexing and information retrieval**. When adopted for describing and indexing resources, not only within the circle of databases and information services immediately associated with GACS partners but also beyond, GACS will improve the quality of indexing and retrieval.
- **Serves as a bridge for translating between indexing languages**. GACS will provide a bridge for translating queries formulated for searching a database indexed with one thesaurus into queries usable for searching databases indexed with another thesaurus, similarly to how the Unified Medical Language System[16] (UMLS) enables interoperability in the medical domain.
- **Serves as a spelling aid**. By pooling alternative labels used in practice for describing and querying information, along with hidden labels holding common misspellings and other erroneous information, GACS will help users in many languages find correct spellings or find the information they seek despite spelling mistakes.
- **Serves as a support for natural language processing**. GACS will hold rich data about terminology that can be exploited to improve the quality of natural language processing.

Cooperation within the global agricultural community

- **Provide a global platform for agricultural thesaurus maintainers.** GACS can provide a platform for involve new partners, such as the Chinese Academy of Agricultural Science[17] or the CGIAR Consortium[18], in the creation of a Global Agricultural Concept Scheme.
- **Provide an organizational basis for collaboration with concept providers.** Coordination among the three thesauri on the acquisition of new concepts can lead to agreement on the use of common sources. GACS, with its foundation in three of the key thesauri for the agricultural field, will provide an organizational platform for collaborating with concept providers such as the Catalogue of Life, the Angiosperm Phylogeny group, or the International Committee on Taxonomy of Viruses.

[16] http://www.nlm.nih.gov/research/umls/
[17] http://www.caas.cn/en/
[18] http://www.cgiar.org

Appendix: Methodology for Phase Two (GACS Beta)

The process:

1. **Select important concepts in each source thesaurus**. For each source thesaurus, select a subset with the K=10,000 concepts (the size was decided at the GACS meeting of 27 May) that are most interesting according to its maintenance organization. This could be done by examining concept frequencies in the main databases (AGRIS, CAB Abstracts, and AGRICOLA), but the process could also involve some human judgement and intervention. For example, it might make sense to include all countries, or the entire taxonomic tree could be included, up to its top levels, regardless of whether all levels occur frequently in metadata.

   Calculating scores. A score could be calculated for each concept based on the frequency it occurs in databases, with adjustments for other factors (e.g., each country concept could be given a large bonus score). To ensure that the taxonomy tree is complete, each taxonomic concept would get a score at least as big as the maximum score of all concepts below it in the hierarchy.

   The subset of concepts with the top 10,000 scores would be selected from each thesaurus. Let's call these subsets A (from AGROVOC), C (from CABT) and N (from NALT).

	A	B	C	D	E	F
1	**AGROVOC ID**	**Labels**	**Relations**	**CABT ID**	**Labels**	**Relations**
2	c_25189	Plant physiology es: Fisiología vegetal	BT Physiology	D91713	plant physiology es: fisiología vegetal	BT physiology
3	c_5956	Plant breeding es: Fitomejoramiento	BT breeding	D91647	plant breeding es: fitomejoramiento	BT breeding
4	c_49985	plant genetics es: fitogenética	BT genetics			
5	c_2981	Floor husbandry es: Crianza en el suelo	BT animal husbandry methods	D48617	floor husbandry es: crianza en corral	BT animal production
6	c_5962	Plant diseases es: Enfermedades de las plantas	BT diseases	D91659	plant diseases es: enfermedades de las plantas	BT diseases
	c_16196	pests of plants es: Plagas de plantas	BT Pests			
7	c_25187	Animal physiology es: Fisiología animal	BT Physiology	D10693	animal physiology es: fisiología animal	BT physiology
8	c_426	Animal diseases es: Enfermedades de los	BT diseases	D10668	animal diseases es: enfermedades de los	BT diseases

Figure 5: Example mapping table used to verify candidate mappings

2. **Generate mappings and mapping tables**. Generate pairwise mappings (e.g., using AgreementMakerLight) between the full thesauri, as was done for the Status Quo report. This could be done not just using subsets, but for the complete thesauri, which should yield more precise mapping scores. By relaxing certain criteria such as confidence threshold, the algorithm could suggest multiple mappings per concept. Let's call these mappings AC, CN, and NA.

   For each subset (A, C, N) generate a table, in the form of an Excel or Google Docs spreadsheet, with one or more rows for each concept in the source subset, together with all candidate mappings to concepts in the target subset (in the next thesaurus of the chain). For example, the A table would include candidate mappings from AGROVOC to CABT. The spreadsheet would include basic information about both source and target concepts, such as URI, prefLabel, broader concept, and definitions or scope notes. If the algorithm were to yield no candidate mappings, some rows may be left without target concepts.

3. **Evaluate mappings manually**. Pass the tables to each organization for evaluation. For example, the AGROVOC maintainers would evaluate the mappings to CABT, CABT maintainers to NALT, and NALT to AGROVOC. Good mappings would be approved and bad mappings removed. In the absence of candidate mappings, the thesaurus maintainers would seek suitable mappings in the target thesaurus manually, the URIs and prefLabels for new mappings would be added to the table.

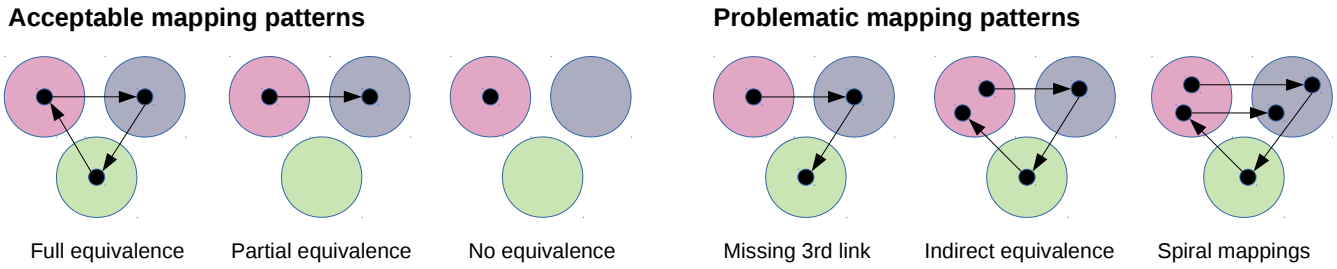

Figure 6: Possible mapping patterns.

4. **Reconcile mappings**. After the first evaluation round, an attempt would be made to algorithmically reconcile the mappings and create equivalence sets for the concepts, flagging potential errors. For example, a "spiral" mapping path would suggest that two concepts from the same vocabulary were equivalent (see Figure 5). If there were mappings from X to Y and Y to Z but not from Z to X, the maintainers could find out why. The mappings would be reworked until all were satisfied.

   Expand mapping tables. The mapping step will have touched on some concepts not in the original subsets A, C, and N, for example concepts from subset A mapped to CABT concepts not in subset C. Add information about these concepts, with their suggested mappings, to the mapping tables, as in step 4. For example, the mapping table from subset C to NALT would be completed with a few more CABT concepts than mapped to from subset A.

   Evaluate added mappings. Evaluate the added mappings, and when done, perform algorithmic reconciliation again. Rinse and repeat until no more new concepts are touched and no important errors are flagged. This should be achievable within two or three rounds, with diminishing amounts of work.

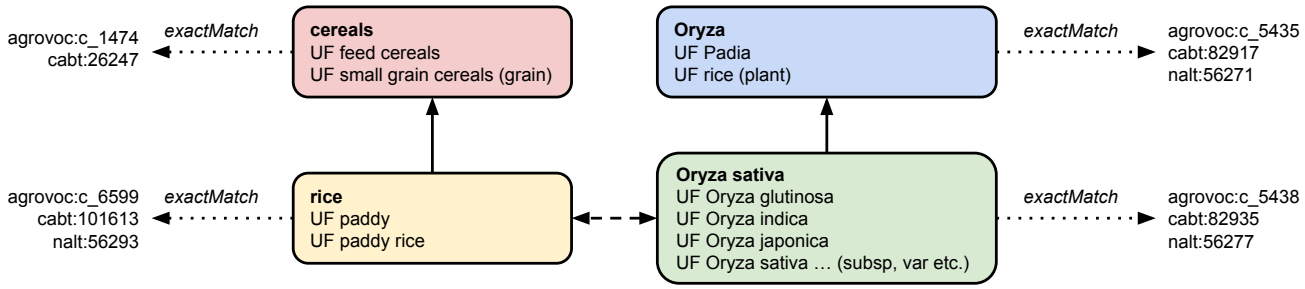

Figure 7: Modeling of rice-related concepts in GACS.

5. **Generate GACS concepts**. The reconciled equivalence sets would now form the set of concepts to be defined for the GACS Alpha, and each concept would map to all source thesauri (see Figure 6) – even for concepts not originally considered among the most important for a given source thesaurus. The final set could be perhaps 10-20% larger than the initial K value.

   Assign GACS URIs. Assign opaque URIs to the GACS concepts using a base GACS namespace URI plus an arbitrary number.

   Complete GACS concepts with additional information. Pull additional information about the concepts, apart from the mappings, from the source thesauri: for example, labels, some semantic information (hierarchy, related concepts, possibly typing information as in UMLS, possibly specialized relationships such as scientific name to common name). GACS, enriched with additional information from three thesauri, would become a source from which all three organizations could in turn enrich their own thesauri.

   Select preferred labels for display purposes. Labels would be expressed using SKOS XL in order to retain provenance information. For display purposes, prefLabels would need to be chosen, based perhaps on some deterministic algorithm (e.g., by majority vote if that gives a clear result, otherwise by some deterministic tie-breaking method).

6. **Edit GACS concepts**. Quality evaluation tools such as qSKOS and Skosify can be used to check for obvious problems. The structure could be reworked within GACS using the editorial platform VocBench.

7. **Evaluate process**. As part of the previous steps, collect additional information and notes about how the mapping process has worked, what kind of problems arise (e.g., invalid mappings, differences in granularity or point of view between the source thesauri, technical problems, amount of work required). These could then be used to provide an overall evaluation of the methodology, potential pitfalls, and an estimate of the amount of work required in subsequent phases.

Publish GACS Beta. As soon as these methods yield a usable result, GACS could be made available to the public. Were GACS to be subsequently expanded, similar processes would be followed. Mappings would be created for all new concepts and after checking for problems, they would be folded in to GACS and their semantic relations would be refined within the GACS editorial platform.