GACS: Status quo of three partner thesauri

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

Version 1.0. August 12, 2014

Introduction

In October 2013, the Food and Agricultural Organization of the United Nations (FAO), CAB International (CABI), and the National Agricultural Library of the USA (NAL) agreed to collaborate[1] in the development of their respective thesauri: the NAL Thesaurus, CAB Thesaurus, and the AGROVOC Concept Scheme. As part of this collaboration agreement, the three organizations are exploring the feasibility of developing a Global Agricultural Concept Scheme (GACS). As the first step in this exploration, this report provides a detailed analysis of the three thesauri and assesses their respective strengths and weaknesses.

The three organizations share a long history of exchange and cooperation. In a precursor to the GACS Project, the three organizations joined forces in 1989 to work on a Unified Agricultural Thesaurus[2] (UAT). The UAT project ended on completion of a UAT classification scheme in 1995 and the near-simultaneous retirement of its principal collaborators from FAO, CABI, and NAL. In the context of the UAT project, the three organizations jointly vetted improvements to AGROVOC and CABT and developed an upper classification structure to which AGROVOC and CABT were mapped. A UAT-classified version of CAB Thesaurus was made available to users until 1999.

AGROVOC (created in 1982), CAB Thesaurus (1983), and NAL Thesaurus (2002) were designed in conformance with the thesaurus practice of their day as laid down in the ISO 2788[3] standard for monolingual thesauri (1974) and the ISO 5964[4] standard for multilingual thesauri (1985). In ISO 2788 and ISO 5964, the primary entity of interest is the *term* – a word or phrase with a specified semantic relationship to other *terms*. The inherent ambiguity between index *terms* and the *concepts* underlying those *terms* was acknowledged in the 1986 revision of ISO 2788. However, the notions of *broader term*, *narrower term*, and *related term* were so deeply embedded in thesaurus practice (as the tags *BT*, *NT*, and *RT*), and the use of words as index keys was so deeply embedded in contemporary database design, that thesauri based on ISO 2788 and ISO 5964 may be characterized as "term-based".

ISO 25964[5], the successor standard to ISO 2788 and ISO 5964 published in 2011 (Part 1) and 2013 (Part 2), explicitly characterizes a thesaurus as a list of *concepts*, each of which is labelled with a preferred term (in each language) and relevant synonyms and equivalents. Thesauri based on ISO 25964 may be characterized as "concept-based". Simple Knowledge Organization System[6] (SKOS), published as a W3C Recommendation in 2009, provides a vocabulary for expressing a concept-based thesaurus for use in Semantic Web and Linked Data applications.

Inasmuch as the GACS Project is taking SKOS as its point of departure, it is worth calling out the sources of potential confusion between the terminology used in term-based thesaurus practice and the terminology used in SKOS:

- *Term* refers to the word or phrase used to label a *concept* in a thesaurus. In an information-technology sense, a *term* is a literal, or string. From a SKOS perspective, a "term-based" thesaurus is an indexing language consisting of what SKOS calls *labels* – words or phrases encoded as strings.

[1] http://aims.fao.org/community/agrovoc/blogs/national-agricultural-library-usa-cabi-and-fao-agree-collaboration-developme
[2] http://www.nal.usda.gov/pubs_dbs/ann_rpts/1994/94arint.htm
[3] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=7776
[4] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=12159
[5] http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53657
[6] http://www.w3.org/TR/skos-reference/

- A *concept* is a unit of thought that has semantics, or meaning. As Leonard Will puts it in his glossary of terms related to thesauri[7]: "Concepts exist in the mind as abstract entities which are independent of the terms used to label them". In a concept-based thesaurus, *terms* (words and phrases) correspond to the *labels* of *SKOS concepts*.
- A *concept scheme* is a set of concepts, optionally specified with semantic relations between concepts. AGROVOC, a *SKOS concept scheme*, is still referred to either as the AGROVOC Thesaurus or as the AGROVOC Concept Scheme. Technically, the category *concept scheme* is broader than the category *thesaurus* and includes subject heading lists, taxonomies, glossaries, classifications, and other types of controlled vocabulary.

An *ontology*, as defined by ISO 25964, is "a formal, explicit specification of a shared conceptualization". The notion of *ontology* appears only at the margins of this report but will become more relevant as the GACS Project moves towards implementation. The word *ontology* has a wide range of meanings in current usage. For the purposes of this report an ontology is a set of statements about things, groups of things, and relations between things in a model of the world, expressed in a language that can be used to verify the logical consistency of that knowledge or to make implicit knowledge explicit. While there are gray areas between the two, an *ontology* is a construct engineered to enable logical reasoning and inference, while a *concept scheme* is a more flexible conceptualization optimized to support the sorts of indexing and structured browsing for which the three thesauri considered here were originally designed.

1. AGROVOC Concept Scheme

History. AGROVOC was created in the early 1980s by FAO and the Commission of the European Communities as a printed thesaurus of agricultural terms for use in indexing the Current Agricultural Research Information System (CARIS), a database of agricultural research projects, and AGRIS[8], a database of bibliographic records. In 2000, work began on expressing AGROVOC as a vocabulary for the Semantic Web. The first iteration, released in circa 2004, was expressed as an OWL ontology. After the publication of Simple Knowledge Organization System[9] as a W3C Recommendation in 2009, with a new SKOS eXtension for Labels (SKOS-XL), the OWL ontology for AGROVOC was converted into SKOS[10]. From circa 9,000 concepts in 1982, AGROVOC grew to 16,000 in 2000 and 32,000 today.

Linguistic coverage. Initially available in English, French, and Spanish, AGROVOC is available today in twenty languages, with four new translations in the works. The translations generally reflect the hierarchical structure used for the English original. In principle, however, AGROVOC is not intended to be based on English alone. It is anticipated that concepts will increasingly be added to AGROVOC by maintainers in other language areas, with labels that do not necessarily have English equivalents. AGROVOC already has 136 concepts with no preferred label in English.

Maintenance platform. AGROVOC is maintained using VocBench[11], a Web-based, multilingual editing and workflow tool for managing thesauri, authority lists, and glossaries in SKOS. VocBench, previously known as the AGROVOC Concept Server Workbench, has been developed by FAO and its partners since 2005[12] and is currently available as open source. The development of VocBench was motivated by a desire to enable authorized participants in various maintenance roles to edit parts of the central AGROVOC ontology simultaneously, for example to prepare translations of terms in other languages or to add relationships between terms. The move to a distributed architecture was seen as a way to loosen the dependence of AGROVOC on terms entered canonically in English, then "translated" into other languages, towards an environment in which users could add new locally-specific terms in any language. The user community of VocBench currently includes FAO's Fisheries and Aquaculture Department and data.fao.org Project, the European Commission Publications Office, the European Environment Agency, and the Italian Senate. The VocBench server for AGROVOC is hosted by FAO Centre of Excellence MIMOS Berhad in Kuala Lumpur.

Editorial workflow. All formal communication among editors of AGROVOC is channeled through VocBench. Roles within VocBench include Term Editor, Ontology Editor, Validator, Administrator, and Publisher. A term can have the status of Draft, Revised, Validated, Published, Proposed Deprecated, or Deprecated. The addition of new concepts is largely a manual process by which proposals are vetted by authorized editors. The AGROVOC Team provides editorial guidelines in the form of a user manual

[7] http://www.willpowerinfo.co.uk/glossary.htm
[8] http://agris.fao.org
[9] http://www.w3.org/TR/skos-reference/
[10] http://www.fao.org/docrep/article/am324e.pdf
[11] http://aims.fao.org/tools/vocbench-2
[12] http://aims.fao.org/interviews/vocbench

and video tutorials for VocBench. Changes made in VocBench are staged in VocBench, then published in periodic releases of AGROVOC. VocBench holds change information at a higher degree of granularity than the version of AGROVOC published on the Web.

Translations. In the 1990s, AGROVOC was maintained in a central database, and further languages were added by sending database dumps to FAO partners for translation. Until recently, translators would get a dump of the database, make their translation, then send the data back to FAO for merging into the master copy. Information on who actually made or revised a translation was internal to the partner institution and was not reflected in the published data. In the VocBench environment, translation follows the same formalized editorial workflow that governs other maintenance activities. Permissions to edit specific languages can be assigned to editors. For some languages, a commercial translation tool is used to generate rough drafts for manual review and correction by translators.

Mappings. In a process tested extensively in 2011 and 2012, candidate mappings are generated automatically using both publicly available and in-house algorithms, then evaluated manually by a thesaurus manager. This process has proven to be quite manageable in terms of editorial workload. As of 2014, AGROVOC has been mapped to the Chinese Agricultural Thesaurus (20,702 concepts), the US National Agricultural Library Thesaurus (13,195), DBPedia (11,015), the Gemeinsame Normdatei of the German National Library (6,212), Aquatic Sciences and Fisheries Abstracts (1,784), Eurovoc (1,269), the General Multilingual Environmental Thesaurus of the European Environment Information and Observation Network (1,185), STW Thesaurus for Economics (1,125), Library of Congress Subject Headings (1,086), TheSoz Thesaurus for the Social Sciences (827), the FAO Biotechnology glossary (793), RAMEAU of the National Library of France (671), Dewey Decimal Classification (401), FAO Geopolitical Ontology (253), and Geonames (206).

Copyright and license policies. For the FAO official languages – English, French, Spanish, Arabic, Russian, and Chinese – copyright stays with FAO. For other languages, copyright stays with the institution that created the entries or prepared the translation. In a situation where it is anticipated that many editors will contribute content in multiple languages and domains, the long-term implications for ownership and copyright are unclear. In practice, at present, AGROVOC may be freely used by anyone. The copyright license is variously asserted to be Creative Commons 3.0 Attribution[13] (in a VoiD file[14]) and Attribution-NonCommercial-ShareAlike 3.0 Unported[15] (on the AIMS website[16]).

AGROVOC users and uses. AGROVOC is used primarily by libraries and other information services for information management and indexing, though a recent survey has shown that it is also used by software developers, translators, and researchers. The AGRIS database is the main user of AGROVOC within FAO. While the AGROVOC team encourages the use of concept URIs in indexing, it is worth noting that some users, such as the data.fao.org project, use the AGROVOC labels, for example as content tags[17]. Users can communicate with the AGROVOC team by email through a form, or by direct mail to `agrovoc@fao.org`, or they can exchange experience among themselves on a recently started Google Groups mailing list. News and announcements about AGROVOC are published on the AIMS website.

2. CAB Thesaurus

History. After CAB Abstracts[18] was created as an online bibliographic database of the Commonwealth Agricultural Bureaux in 1972, the CAB Thesaurus[19] (CABT) began as a controlled list of keywords for indexing the abstracts. The first printed edition of CABT was published in 1983 with 48,000 preferred and non-preferred terms. Since 1999 the thesaurus has been available in electronic form only. The total number of terms has grown from 63,000 in 1999 to 81,000 in 2008, to 245,000 today.

Linguistic coverage. Initially available in English only, Spanish and Portuguese were added to CABT in 1999 and Dutch was added in 2012. The coverage of descriptors in these languages is substantially complete. Lesser content is available in a further seven Western European languages. Before 2012 the translations were tied to the English language original. However, currently there are approximately 19,000 terms with no English equivalents.

[13] http://creativecommons.org/licenses/by/3.0/
[14] http://aims.fao.org/aos/agrovoc/void.ttl
[15] http://creativecommons.org/licenses/by-nc-sa/3.0/
[16] http://aims.fao.org/standards/agrovoc/functionalities/download
[17] http://data.fao.org/tagsearch
[18] http://www.cabi.org/publishing-products/online-information-resources/cab-abstracts/
[19] http://www.cabi.org/cabthesaurus/

Maintenance platform and editorial workflow. CABI uses the commercial software program MultiTes Pro[20] to maintain the working copy of CABT. The CABT working data file is kept on a networked drive potentially accessible to all CABI staff worldwide with installation of MultiTes client software. Currently there are two thesaurus maintainers, and three other users in the Plantwise team, but this could be increased to any number of users with appropriate licencing. The thesaurus team appreciates MultiTes for its ease of use, the clarity of its interface, the ease with which imports can be made in plain text, and the simplicity with which customized extracts can be exported in various formats, including plain text, comma-limited text, XML and SKOS, using a wizard interface. The MultiTes Web Deployment Kit[21] makes it easy to generate the presentation of CABT that is posted on the Web. Proposals for new descriptors come from many sources, mostly in-house, for approval by the thesaurus managers.

The use of CABT for indexing CAB Abstracts. CAB abstracts are indexed using CABT descriptors. Put another way, the abstracts are indexed with the literal terms used to label CABT concepts. The string "cows" may be replicated in many thousands of records. When the preferred label for the concept *climate change* was changed from "climatic change" to "climate change", many thousands of records needed to be updated with the new string value and re-indexed. Any such change in the index affects all CABI products that are derived from the abstracts. The subset of abstracts published in the journal Plant Breeding Abstracts, for example, is extracted from CAB Abstracts by running a complex search profile. Any change in a preferred label may affect thousands of records, numerous search profiles, and any number of derivative products downstream. Proposals for changes to CABT descriptors are vetted by the CABT maintenance team, and the changes are published in major updates of the entire thesaurus every one to two years. The CAB Abstracts database indexing fields are then updated to reflect the thesaurus changes.

CABI Compendia. Parallel to CAB Abstracts, the CABI Compendium program produces online datasheets in the areas of Aquaculture, Animal Health and Animal Production, Crop Protection, Forestry, and Invasive Species. Compendia are indexed using a species hierarchy that parallels corresponding sections of the CAB Thesaurus. In creating datasheets, the Compendium editorial team follows CABT where possible, but CABT may lack the needed terms, or its terms may differ from those supplied by the expert authors of the datasheets, for example in cases where scientific consensus is in flux. All of the 10,000 "full" and 120,000 "basic" Compendium datasheets produced over the past twenty years are maintained in a central database. The scientific and common names of organisms and diseases, in English, French, and Spanish, are held in the fields of individual datasheets. References in datasheets are linked to abstracts in CAB Direct using a reference manager that pulls authors out of correctly formatted references and puts them into specific fields, then performs a string match with author names in CAB Abstracts. The quality of linkage between a Compendium datasheet and a CAB abstract likewise depends on how well the name string used in the datasheet matches a preferred label in the thesaurus. The Compendium and CABT editorial teams cooperate to ensure that updates to organism names in particular are coordinated. For any given organism, state-of-the-art taxonomic information may be recorded first either in the compendia or in CABT. Efforts are then made to synchronize the latest names across both products as soon as possible. However, while compendia datasheets are published to the Web shortly after an update, thesaurus changes are not made publicly available until the annual or biennial release.

Plantwise factsheets. CABI leads a global program, Plantwise[22], which works with partners in about 35 developing countries to improve food security and the lives of rural poor by reducing crop losses. Plantwise supports the operation of community-level clinics where farmers can bring diseased plants for diagnosis by "plant doctors" and obtain up-to-date, validated advice about fertilizers, pesticides, and farming methods. The program also promotes the integration of local clinics into country-level plant health networks of government agencies, input suppliers, extensionists, and researchers. Up-to-date information and advice is provided in the form of one-page factsheets about crops, pests, weeds, diseases, and invasive species. Plantwise factsheets are produced using a back-end Knowledge Bank that draws, in part, on CABI's own research-oriented Compendia. Factsheets are tagged with concept IDs. For example, the concept ID of an insect is associated in the back-end database with the insect's common and scientific names. If a name changes, or species merge or split, concept IDs will always point to the most current names. A concept ID is also associated, via a lookup table, with the MultiTes ID of the preferred scientific name in CABT.

Distribution database. The location of things described in the CABI databases or observed by extensionists for Plantwise – species, pests, but also buildings, clinics, and partner organizations – are logged in a central "distribution database" that associates a concept ID with a time and a place (as defined by geographic coordinates). The database, currently maintained specifically for use by Plantwise, is separate from the database underlying the Compendia. As with the Plantwise Knowledge Bank, the link to controlled common and scientific names in the CAB thesaurus is established by matching a distribution database concept ID with the MultiTes ID for the CABT concept.

[20] http://www.multites.com/productsPRO.htm
[21] http://www.multites.com/productsWDK.htm
[22] http://www.plantwise.org

Copyright. The CAB Thesaurus is available for browsing via an online search form, and sample excerpts from the thesaurus can be downloaded in several formats (plain text, comma-delimited text, XML, and SKOS/RDF), but machine-readable copies of the entire thesaurus (usually requested in XML or comma-delimited formats) must currently be purchased. Purchasers are typically organizations that host CABI databases and which use the thesaurus in their own search products for expanding queries or displaying term hierarchies. For the purposes of the GACS Project, CABI has committed to the idea of creating a common core thesaurus as open data, as a proof of concept, to be published jointly with FAO and NAL under the terms of a Creative Commons license.

3. NAL Thesaurus

History. The NAL Agricultural Thesaurus[23] (NALT) was developed in the late 1990s as an in-house resource for the Agricultural Network Information Center[24] (AgNIC). A first version of NALT, with 2,000 terms, was made accessible online in 1999 to scientists of the Agricultural Research Service[25] (ARS) of the United States Department of Agriculture (USDA), who helped extend the coverage to a wide range of fields. NALT was posted on the Web in January 2002. From the beginning, NALT has been offered exclusively as an online service.

Linguistic coverage. NALT was originally available in English, with American English as preferred terms (and British English as non-preferred terms). In cooperation with the American Distance Education Consortium, a complete translation of the 2006 edition of NALT provided a Spanish version known as “Tesauro Agrícola”. All NALT terms are available in both English and Latin American Spanish. Since 2007, the Spanish translation has been maintained in collaboration with the Inter-American Institute for Cooperation on Agriculture[26] (IICA).

Maintenance platform and editorial workflow. Like CABI, NAL uses MultiTes Pro for maintaining its thesaurus and the MultiTes Web Deployment kit for generating Web pages. The standalone nature of MultiTes Pro is seen as a limitation, especially for a workplace that practices teleworking. The main working file for NALT is maintained in MultiTes by a thesaurus coordinator. Every four to six weeks, data integrity checks are performed and the data is scrubbed. Every six to eight weeks, the working file for NALT is exported from MultiTes to XML, converted into SKOS, then loaded into Luxid[27], an editorial platform used for the annotation and indexing of AGRICOLA[28], NAL’s online bibliographic database of citations to agricultural literature. The AGRICOLA indexers, in turn, send proposals for new terms and other changes to the thesaurus coordinator. Selection criteria and style guidelines are documented internally. Once per year, a new release of NALT is generated, along with representations in SKOS, XML, PDF, and Word.

Use of NALT for indexing AGRICOLA. The AGRICOLA (AGRICultural OnLine Access) is organized into two bibliographic datasets:

- The NAL Online Public Access Catalog contains citations for books, audiovisual materials, serials, and other materials. The Catalog uses Library of Congress Subject Headings as its controlled vocabulary.
- The Article Citation Database contains citations to journal articles, book chapters, reports, and reprints selected primarily from the NAL Catalog. In 1984, before developing its own thesaurus, NAL began using the CAB Thesaurus as the controlled indexing vocabulary for the citation database. In 2005, the NAL Thesaurus was adopted as the controlled vocabulary for indexing AGRICOLA, and for reasons of consistency, the CABT descriptors used in the database were converted to NALT descriptors. Each year, the Article Citation Database undergoes bibliographic conversion so that all citations are aligned to the latest version of NALT.

NALT is not currently integrated into AGRICOLA search, such that users could take advantage of synonym rings and hierarchical relationships. A new search engine, currently under development, will emulate the ability of PubMed to automatically include synonymous terms in queries and will support the “automatic explosion” of a query to include narrower terms.

[23] http://agclass.nal.usda.gov
[24] http://www.agnic.org/
[25] http://www.ars.usda.gov/main/main.htm
[26] http://www.iica.int
[27] http://www.temis.com/home
[28] http://agricola.nal.usda.gov

NALT and automated indexing. In 2010, NAL ceased manual indexing and transformed its indexing operation into an automated process using Luxid. NAL uses a combination of machine-aided indexing and automated indexing with limited review. NALT is central to "annotation plans" – the scripts, rules, and workflows used for the automatic assignment of NALT descriptors to text.

NALT users and uses. NALT serves a variety of users from scientific researchers to policy makers, small producers, company CEOs, information professionals, and US citizens. To meet the needs of diverse users, NALT includes common terminology alongside the scientific and technical. Definitions are published in a separate bilingual (English and Spanish) glossary that is used by educators, students, and translators. Within USDA, the AGRICOLA database is the main user of NALT. NALT is also used by other public and private providers of agricultural information, such as the Food Safety Research Information Office[29], AgNIC (Agriculture Network Information Center), JIFSAN[30] (Joint Institute for Food Safety and Applied Nutrition), Pecan *ipm*PIPE[31], and REMBA[32] (The Mexican Network of Agricultural Libraries).

Translations. Every four to six weeks, new English descriptors, definitions, and scope notes are sent to to IICA for translation into Spanish. Translations are provided by IICA expert translators and are only reviewed and edited for format prior to input into the working file by NAL staff. Translations for new terms and definitions are made available to users at the annual release of the thesaurus.

Mappings. Some experimental mappings between AGROVOC and NALT, and between AGROVOC, NALT, and the GEMET thesaurus, were prepared in 2006 and 2007 in the context of the Ontology Alignment Evaluation Initiative[33] (van Hage et al. 2010). At present, however, no mappings are published in the SKOS representation of NALT as Linked Open Data.

Copyright. The Terms and Conditions of Use[34] for NALT specify that no license is required to obtain NALT data. If use of the thesaurus is not personal, NAL must be identified as its maintainer, the version used must be clearly stated by year, and any modifications made in its content must be described and documented. The publication of NALT as Linked Open Data in 2011 was undertaken in the spirit an Open Government Directive[35] to make government data publicly and freely available. In May 2014, the White House released a US Open Data Action Plan[36] that supports the use of the Creative Commons license CC0 Public Domain Dedication[37].

4. Comparison of thesauri

This section presents the results of a comparison of the content of the three thesauri undertaken with various tools and by manual inspection. The comparison was based on the following snapshots:

- For **AGROVOC**: the SKOS version of AGROVOC Core in RDF/XML, dated 2013-12-17, was used for all comparisons.
- For **CABT**: the SKOS version (in RDF/XML syntax) provided by CABI on 2014-04-01 was the basis for the RDF vocabulary analysis in Section 4.2. However, this version presented modeling difficulties and was not complete enough as a representation of CABT for some comparisons. XML dumps of CABT provided by CABI on 2014-04-16 were converted into an ad-hoc SKOS version for the linguistic coverage and overlap analyses.
- For **NALT**: the SKOS version (in RDF/XML syntax) of the NALT 2014 Edition, dated 2013-12-14 was used for most comparisons. XML dumps of the English and Spanish versions of NALT, provided by NAL on 2014-04-17, were used for the analysis of provenance information.

[29] http://fsrio.nal.usda.gov/
[30] http://jifsan.umd.edu/
[31] http://pecan.ipmpipe.org/
[32] http://remba.uaa.mx
[33] http://www.few.vu.nl/~wrvhage/oaei2007/
[34] http://www.nal.usda.gov/web-policies-and-important-links
[35] http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf
[36] http://www.whitehouse.gov/sites/default/files/microsites/ostp/us_open_data_action_plan.pdf
[37] https://creativecommons.org/publicdomain/zero/1.0/

4.1. Linguistic coverage

Figure 4 in Appendix A provides an overview of the linguistic coverage of all three thesauri. A simple count of the numbers of terms in each language would have been skewed by the common practice in all thesauri of including scientific (Latinate) names of organisms in every language as if they were terms in that language. This analysis tries to separate out the scientific names and count them separately. The figures are not exact due to the variety of ways in which scientific names are (or are not) indicated in each thesaurus and because of occasional inconsistencies (e.g., with the tagging of Turkish terms in AGROVOC). Separating scientific names from regular terms brings the differences in language coverage into sharper relief.

- **AGROVOC** includes the most languages, with preferred terms for at least 50% of all concepts (excluding scientific names) available for 20 languages.
- **CABT** contains by far the most scientific names and also has the largest number of English, Spanish, Portuguese and Dutch preferred terms. Terms are available in 11 languages, but only the English, Spanish, Portuguese and Dutch terms are actively maintained. Terms in other languages were often added when importing databases of species names without actively undertaking their translation. In addition to natural-language terms, CABT holds 4,493 CAS registry codes of chemicals and 358 enzyme codes.
- **NALT** contains only English and Spanish terms, with both covering 100% of concepts. It contains the largest number of English terms due to the high number of English non-preferred terms. The number of Spanish terms is the highest of the three thesauri. NALT contains 43,641 hidden labels in English (not shown in the graph) that range from variants in spelling (e.g., "rice-flower" with hyphen), in parts of speech, singular versus plural, and the like. These hidden labels are not expressed using `skos:hiddenLabel` or distributed as Linked Open Data. The maintainers of NALT have also followed a policy of liberally accepting words used by non-scientists as lead-in terms (non-preferred labels).

British and American English. English is the language best covered in the three thesauri. Both CABT and NALT have English preferred terms for all concepts. AGROVOC and CABT prefer British English forms, while NALT uses American English. CABT includes 604 terms in American English.

Letter case. In CABT and NALT, the convention of writing terms in lowercase is followed for most languages except in cases where the type of term requires an initial uppercase letter (e.g., for scientific names, proper names, and acronyms). In AGROVOC, however, terms are consistently written with an initial uppercase letter in several of its languages (e.g., English, Spanish, Portuguese, Turkish, Italian, Polish, and French), and a large part of the German terms are written entirely in uppercase.

4.2. Semantic structure and RDF vocabularies used

Use of SKOS vocabulary. All three thesauri are represented in RDF using basic SKOS constructs such as `skos:Concept`, `skos:prefLabel`, `skos:altLabel`, `skos:definition`, `skos:broader`, and `skos:related` (see table in Appendix B). AGROVOC contains 32,295 concepts, CABT 139,822 concepts and NALT 53,280 concepts. AGROVOC does not assert `skos:narrower` relationships in either its Core or LOD versions; they must be inferred by the user when necessary. Of the three, only AGROVOC provides a `skos:ConceptScheme` instance representing the thesaurus itself. AGROVOC systematically uses the SKOS eXtension for Labels vocabulary (SKOS-XL) to represent thesaurus terms as resources identified with their own URIs. For the benefit of data consumers and applications that may not understand SKOS-XL, the LOD dataset for AGROVOC supplements SKOS-XL labels (e.g., `skosxl:prefLabel`) with plain-literal SKOS equivalents (e.g., `skos:prefLabel`) automatically generated by inference.

Completeness of SKOS version. Since AGROVOC uses SKOS and SKOS-XL natively, all information contained in the thesaurus (except for some editorial metadata maintained only within VocBench) is represented in the SKOS version. AGROVOC and CABT use Dublin Core properties for timestamps. AGROVOC also uses FOAF for describing images. CABI and NAL have used basic SKOS, with no extensions, to publish subsets of their thesauri. Since SKOS lacks properties for expressing provenance information about terms, such as timestamps, none have been provided, though the SKOS output from MultiTes may be modified to provide such information in the future.

Custom properties for relationships. Both AGROVOC and CABT make some use of custom properties for relationships between concepts and terms that are more specific than the traditional thesaurus notion of a related term (see Tables 5 and 6

in Appendix B). Both thesauri also use custom properties to classify concepts and terms (see section 4.3). They link common names to scientific names using a similar mechanism.

- **AGROVOC** relates concept labels among themselves with custom properties such as `hasSynonym`, `hasAbbreviation`, and `hasOldName`. The Agrontology vocabulary[38] defines properties relating concepts as sub-properties of `skos:related` and properties relating labels as sub-properties of `skosxl:labelRelation`. Most of the 179 object properties defined in Agrontology are relationship refinements that date from a joint initiative of FAO and ICRISAT[39] in the mid-2000s. In practice, only 22 Agrontology properties are used more than 500 times in AGROVOC. Most of the Agrontology properties are shown as plain RT relationships in the AGROVOC browsing interface. Their use in AGROVOC has more recently been de-emphasized, and a reversion of some relationships to the standard `skos:related` is being considered. AGROVOC also contains VocBench-specific fields for image metadata as well as provenance and status information for concepts and terms (see section 4.5).
- **CABT.** The use of custom relationships in CABT has been deliberately limited to a few basic relationships such as *Crop Plant* and *Disease Agent* in order to avoid the maintenance burden. CABT also uses some custom fields, such as nameAuthor and termSource (discussed in section 4.5) as well as custom relationships for chemical and enzyme codes. These are erroneously asserted to be part of the SKOS namespace in the current RDF representation. Non-SKOS URIs should be coined for such properties.
- **NALT** uses no such custom extensions in its SKOS representation, which uses a limited subset of the basic SKOS vocabulary.

Significantly, the CABT and AGROVOC teams have reached similar conclusions about the usefulness of some custom relationships, such as *produces* (though in this case, one property for *produces* is broader than the other).

Hierarchical relationships. Polyhierarchy – where a concept has more than one broader concept – appears in all three thesauri. In AGROVOC this is the case for 1,200 (3.7%) concepts, in CABT 3,512 (2.5%) concepts and in NALT 2,476 (4.6%) concepts. CABT also contains the custom relationships *Related Term Broader (RTB)* and its inverse *Related Term Narrower (RTN)*. This relationship is used to link distinct parts of the hierarchy, as in the example *forest trees RTN Abies alba*. *Abies alba* is an example of a *forest tree*, but the terms reside in different hierarchies: *Forest trees* is in a functional-use, descriptive hierarchy while *Abies alba* is in a hierarchy for scientific names. This relationship has been represented using `skos:broader` and `skos:narrower` in the original SKOS version, but later versions represent it as `skos:related`. Custom subproperties of `skos:related` could potentially be used to represent this relationship. The hierarchical relationships used give each thesaurus a particular shape, depicted with "icicle" visualizations in Appendix C.

Modelling example. Figure 1 shows how concepts related to *rice* have been modelled in the three thesauri. All thesauri represent the product *rice* and the plant species *Oryza sativa* as separate concepts, with a relationship connecting the two. AGROVOC uses the *produces* relationship from Agrontology (a refinement of `skos:related`), while CABT uses the custom relationship *crop plant* and its inverse, *harvested product*. NALT has no special relationship types, so a standard RT (`skos:related`) relationship is used. All three thesauri have a different subdivision of concepts below *rice* in the hierarchy. Both AGROVOC and CABT place *rice* under *cereals*, while NALT has a more complex grain product hierarchy. The taxonomic hierarchies of the *Oryza* genus are all somewhat different. CABT includes many associative relationships (not shown) from *rice* and *Oryza sativa* to various diseases and plant viruses.

Types of thesaurus concepts. Some of the custom relationships implicitly classify the source or target concepts into distinct types in addition to the structural organization discussed in section 4.3. For example, the *Harvested Product* relationship in CABT implies that the source concept is a plant and that the target concept is a plant product. Similarly, the AGROVOC *hasTaxonomicLevel* property classifies the source concept into a specific taxonomic level. Only the standard `skos:Concept` type is used for these concepts. If a more ontological structure were desired, it would be possible to use more specific RDF types (subclasses of `skos:Concept`) for these concepts.

Extended term relationships. Both CABT and NALT contain USE-AND relationships, where compound terms are represented using a combination of simpler terms. The USE-AND relationship could be expressed using the ISO 25694-1 extensions to SKOS[40] (SplitNonPreferredTerm, USE+ and UF+). CABT also contains USE-OR relationships, where ambiguous or deprecated terms are

[38] http://aims.fao.org/aos/agrontology
[39] http://www.icrisat.org/
[40] http://purl.org/iso25964/skos-thes#

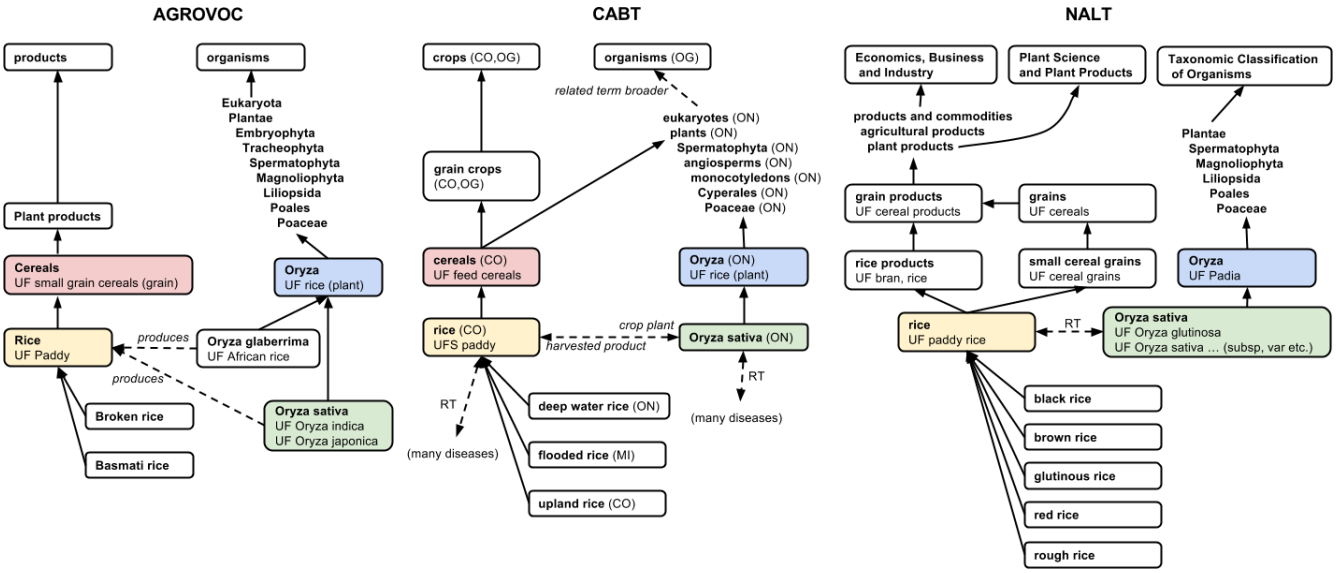

Figure 1: Modelling of rice-related concepts in the three thesauri. The plain arrows represent hierarchical relationships, while dashed arrows indicate non-hierarchical relationships. Some parts of the hierarchy are shown as abbreviated lists. Non-preferred terms are included, indicated by the UF tag, for the most important concepts. For CABT, subject category codes (e.g. ON for organism names) are included. The concepts are colored according to the automatically generated mappings; the same color is used for concepts that were mapped to each other.

redirected to several possible valid terms. There is no standard RDF representation for USE-OR, but a recent discussion[41] on the public-esw-thes mailing list ended up suggesting that the USE-OR relationship be expressed by adding the ambiguous term as a non-preferred term (altLabel) to all valid concepts.

Thesaurus quality issues. The SKOS representations of the three thesauri were examined for quality issues using the qSKOS[42] vocabulary quality analysis tool (Suominen and Mader 2014). The process uncovered some minor and easily fixable errors, such as missing, empty, and overlapping labels, and a few redundancies or clashes in the hierarchy, but no serious problems of a logical nature.

4.3. Structural organization

All three thesauri have defined structures that can be used to select subsets of the vocabulary.

Top level concepts

- **AGROVOC** has 25 top-level concepts that may be considered facets – distinct types of concepts that are generally exclusive. The top concepts are *activities*, *entities*, *events*, *factors*, *features*, *groups*, *location*, *measure*, *methods*, *objects*, *organisms*, *phenomena*, *processes*, *products*, *properties*, *resources*, *site*, *stages*, *state*, *strategies*, *subjects*, *substances*, *systems*, *technology*, and *time*. The distribution of concepts below the top-level categories is very uneven, with *organisms* containing by far the largest number of concepts. The concepts *substances* and *entities* follow far behind, and the remaining concepts form a long tail.
- **CABT** has no systematic top-level structure. It contains 3,298 top level concepts.
- **NALT** has 17 top-level concepts, or subject categories[43], that are thematic in nature. The top concepts are *Animal Science and Animal Products*, *Biological Sciences*, *Breeding and Genetic Improvement*, *Economics*, *Business and Industry*, *Farms*

[41] http://lists.w3.org/Archives/Public/public-esw-thes/2014Apr/0032.html
[42] https://github.com/cmader/qSKOS/
[43] http://agclass.nal.usda.gov/dne/search_sc.shtml

and Farming Systems, *Food and Human Nutrition*, *Forest Science* and *Forest Products*, *Geographical Locations*, *Government*, *Law and Regulations*, *Health and Pathology*, *Insects and Entomology*, *Natural Resources*, *Earth and Environmental Sciences*, *Physical and Chemical Sciences*, *Plant Science and Plant Products*, *Research*, *Technology and Engineering*, *Rural and Agricultural Sociology*, and *Taxonomic Classification of Organisms*. Of these, *Taxonomic Classification of Organisms* contains the most concepts, followed far behind by *Biological Sciences* and *Physical and Chemical Sciences*. The seventeen categories resulted from a compromise between people who wanted to keep the number of categories small enough to fit on one screen (with circa ten concepts) and those who wanted to ensure that specialized communities, such as soil scientists, would be sufficiently represented (with more than thirty).

Concept categories or groups

In some thesauri, concepts have been classified into categories separately from the hierarchical structure. For example, when the hierarchy reflects a generic is-a relationship, a separate classification may classify concepts according to theme or domain of use. The ISO 25964-1 standard defines *concept group* as "A group of concepts selected by some specified criterion, such as relevance to a particular subject area".

- **AGROVOC** contains five **sub-vocabularies**: *Chemicals*, *Fishery related term*, *Geographical above country level*, *Geographical country level*, and *Geographical below country level*. The sub-vocabularies together account for only 6% of concepts; the remaining concepts have not been placed in any sub-vocabulary.
- **CABT** contains 22 **subject categories** with associated category codes: *AB Animal Breeds*, *AM Anatomical and Morphological Structures*, *AT Activities Terms*, *BG Biogeographic Units*, *CH Chemicals and Chemical Groups*, *CL Climate Terms*, *CO Commodities*, *DS Disciplines*, *Occupations and Industries*, *DT Disease Terms*, *GE Geographic Entities*, *HT Habitat Terms*, *IN Infrastructure Terms*, *IO Institutions and Organisations Terms*, *MI Miscellaneous Terms*, *OG Organism Groups*, *ON Organism Names*, *PG People Groups*, *SO Soil Types*, *TF Topographic Features*, *TM Techniques, Methodologies and Equipment*, *TP Time Periods*, and *VT Vegetation Types*. By far the biggest subject category is *ON Organism Names*, followed far behind by *CH Chemicals and Chemical Groups* and *MI Miscellaneous Terms*. All concepts have at least one subject category. A concept may occur in multiple categories, though this is rare.
- **NALT** has no concept categories apart from the 17 top-level concepts.

Technical categories and term types

Similar to concept categories or groups, the terms themselves may be classified independently from the concepts they label. For example, singular and plural forms, or scientific versus common names for the same species, may be described using term types or technical categories.

- **AGROVOC** terms can have a **term type**. There are eleven term types: *Acronym*, *Common name for animals*, *Common name for bacteria*, *Common name for fungi*, *Common name for plants*, *Common name for viruses*, *Taxonomic terms for animals*, *Taxonomic terms for bacteria*, *Taxonomic terms for fungi*, *Taxonomic terms for plants*, and *Taxonomic terms for viruses*. A term may not have multiple term types. About 44% of AGROVOC terms have been assigned a term type.
- **CABT** terms are similarly organized in nine **technical categories**: *ABB Abbreviation*, *COM Common Name (Organisms)*, *HOM Homograph*, *POP Popular Name*, *P Plural Form*, *RN Registry or Code Number*, *R Registered Name*, *SCI Scientific Name (Organisms)*, and *S Singular Form*. 88% of CABT terms have a technical category.
- **NALT** has no term types or technical categories.

4.4. Thematic coverage and overlap

All thesauri contain a large number of **species**. Species constitute approximately 63% of AGROVOC, 81% of CABT, and 66% of NALT concepts. This category of concepts includes scientific names and classifications, common names, and in some cases non-scientific classifications such as *livestock* and *pests*. Another distinct category of concepts is **chemicals** and substances, and a third, not very large but easily distinguishable subset is that of geographical **places**. The remaining **other** concepts range

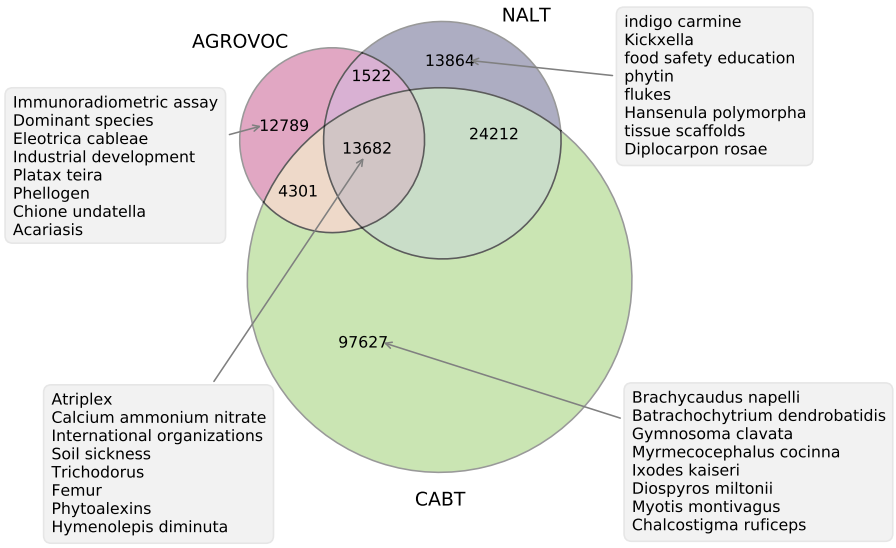

Figure 2: Estimated overlap of the complete thesauri.

from scientific disciplines to agricultural methods and the like. This division of the concepts into four groups is very similar to the categorization of concepts used in the OAEI 2008 mapping task (van Hage et al. 2010).

The three thesauri were examined for overlap, and specifically for the subsets of **Species**, **Chemicals**, **Places**, and **Other** (see Table 1). The overlap estimate was created by first generating pairwise mappings between all three pairs of thesauri using the AgreementMakerLight ontology matching tool (Faria et al. 2008), and then merging the mappings. The figures are not exact because the mappings have not been manually validated, but they provide a ballpark estimate. The general overlap is visualized in Figure 2 as a Venn diagram showing the relative sizes of the thesauri along with examples of concepts that overlap or are unique to each. The overlap within each subset is presented in Appendix D.

The diagrams show that the common core shared between all three thesauri is circa 13,500 concepts. Circa 44,000 concepts are shared between at least two thesauri, while a union of all three thesauri would contain approximately 168,000 concepts.

	AGROVOC	CABT	NALT
Species	Everything below *organisms*	Everything in categories *ON Organism Names*, *OG Organism Groups*	Everything below *Taxonomic Classification of Organisms*
Chemicals	Everything in sub-vocabulary *Chemicals*	Everything in category *CH Chemicals and Chemical Groups*	Everything below *chemical substances*
Places	Everything in sub-vocabularies *Geographical below country level*, *Geographical country level*, *Geographical above country level*	Everything in category *GE Geographic Entities*	Everything below *Geographical Locations*
Other	Everything else	Everything else	Everything else

Table 1: Definitions of subsets used in analysis.

4.5. Provenance

	AGROVOC	CABT	NALT
Vocabulary level	**Modified**, VoiD metadata	-	-
Concept level	**Created**, **Modified**, **Status**	(not maintained explicitly, but all term-level information available for descriptors)	(same as CABT)
Term level	**Created**, **Modified** (11%), **Status**	Created, Modified, Approved, **History note** (73%), **Organism name author** (18%), **Source of term** (6%)	Created, Modified, Status, Source (22%)
Definitions	**Created**, **Modified**, **Link** (89%), **Source** (89%)		

Table 2: Provenance information available in each thesaurus. Fields set in **bold** are available in the SKOS version of the corresponding thesaurus. Percentages indicate the share of entities (concepts, terms, and definitions) having the specified provenance information where not available for all entities.

An overview of available provenance information is given in Table 2.

- **AGROVOC** maintains provenance metadata using VocBench, which keeps track of concept scheme-, concept-, and term-level timestamps. A separate VoiD file[44] contains metadata about the AGROVOC dataset, such as title, publisher, creation and modification dates, license, number of triples, and location of SPARQL endpoint. AGROVOC models concept definitions as separate resources, and VocBench keeps track of provenance metadata for definitions. Most definitions have a source and a link (URL). All of this provenance information is available in the published SKOS version, with term-level provenance information attached to SKOS-XL labels. The Agrontology[45] vocabulary is used for AGROVOC-specific properties such as status.
- **CABT.** In CABT, all provenance information is maintained on the term level for both descriptors and non-descriptors. Most terms have a history note such as “From 2011” in addition to the timestamps automatically maintained by MultiTes. Some organism names are described with the author of the name, and the source of the term is given for 6% of terms. The most common sources are “DSMZ”, “Universal Virus Database, ICTVdB”, and “Taxonomic Outline of the Bacteria and Archaea”.
- **NALT.** In NALT, timestamp fields are likewise maintained by MultiTes. In addition, the source is given for 22% of descriptors. The most common sources are “Bergey’s Manual of Systematic Bacteriology”, “International Committee on Taxonomy of Viruses” and “Germplasm Resources Information Network”.

[44] http://aims.fao.org/aos/agrovoc/void.ttl
[45] http://aims.fao.org/aos/agrontology

4.6. Publication in traditional formats

	AGROVOC	CABT	NALT
Printed book	Last edition in 1995	Last edition in 1999	Never published
Browse online	Yes	Yes	Yes
Bulk download	Yes, SKOS only	By request	Yes, several formats
APIs	Yes, web services	No	No
Linked Data	Yes	No	Yes
SPARQL endpoint	Yes, public	No	No
Customized excerpts	Yes, by request	Yes, by request	Yes, by request

Table 3: Publication of the thesauri in traditional formats and Linked Data.

The publication of each thesaurus in traditional formats and Linked Data is summarized in Table 3. All thesauri are freely browsable and searchable on the Web. AGROVOC is available for bulk download[46] as SKOS in two variants. "AGROVOC Core", with just concepts and labels, is available in either RDF/XML or N-Triples. "AGROVOC LOD" – the Core, plus provenance information and mappings – is available in TriX syntax. NALT provides PDF, XML, Word, and MARC authority record formats in addition to SKOS (RDF/XML only). CABT is available for purchase in plain text, comma-delimited, XML, HTML, and SKOS/RDF formats.

AGROVOC is the only thesaurus providing API access via a Web Service API[47]. However, no thesaurus provides a REST-style API, which could be more useful[48] than a Web Service API for web developers. AGROVOC was previously made available as a relational database dump and in XML, but these formats are no longer offered or requested.

4.7. Publication as Linked Data

Only AGROVOC and NALT are currently available as Linked Data, though a SKOS version of CABT is available on request.

AGROVOC URIs have the form

```
http://aims.fao.org/aos/agrovoc/c_nnnn       (for concepts)
http://aims.fao.org/aos/agrovoc/xl_lang_nnnn (for terms)
```

where `nnnn` is a number and `lang` is a language tag. Concept numbers have between one and five digits for older concepts, and thirteen digits for newer concepts created using VocBench. Term numbers always have thirteen digits and are assigned by VocBench. The `fao.org` domain used in the URIs is owned by FAO.

CABT URI policy is still evolving. The initial SKOS version provided by CABI used concept URIs of the form

```
http://www.site.edu/#adzuki%20beans
```

where the hostname is clearly a placeholder value and the local part is based on the English preferred term. The current plan is to move to an opaque, language-independent URI scheme in the `cabi.org` domain, which is owned by CABI.

NALT concept URIs have the form

```
http://lod.nal.usda.gov/nalt/nnnn
```

[46] https://aims-fao.atlassian.net/wiki/display/AGV/Releases
[47] http://aims.fao.org/standards/agrovoc/webservices
[48] http://aims.fao.org/interviews/make-vocabularies-easily-accessible-both-regular-users-and-developers

where the number `nnnn` is the number of the preferred term assigned by MultiTes. The `usda.gov` domain is owned by the U.S. Department of Agriculture, of which NAL is part. NAL operates the `nal.usda.org` sub-domain.

Both AGROVOC and NALT URIs have been set up to serve Linked Data. A normal Web browser accessing the concept URI will be given a HTML page with information about the concept, while Linked Data agents will be given machine-readable RDF/XML data if they request it using HTTP content negotiation. Alternative RDF syntaxes such as N-Triples and Turtle seem not to be supported by either for the Linked Data access. The AGROVOC Linked Data access has been set up by FAO partner MIMOS Berhad using the Pubby software package, while NALT has integrated Linked Data functionality into their vocabulary browser. The AGROVOC public SPARQL endpoint[49] is also operated by MIMOS Berhad and is based on AllegroGraph software. Both the AGROVOC Linked Data access and the SPARQL endpoint use the data from the latest AGROVOC LOD version published by FAO.

A recent survey of AGROVOC users shows that the SPARQL endpoint is used more than expected, while web services are used far less than expected (only by big institutions and libraries).

5. Strengths and weaknesses

5.1. AGROVOC

Strengths. Multilinguality, including move away from dependence on English in principle. Number of institutional users. Collaborative editing environment (VocBench). Experience and work done towards use in (Linked) Open Data environment. SPARQL access. API access potentially useful, though not used much in current setting. Rich provenance information separately for both concept and term levels. Many mappings available.

Weaknesses. Complicated structure (especially the Agrontology relations). Lowest number of concepts (especially species) of the three thesauri. No thematic categorization that would cover all concepts.

5.2. CAB Thesaurus

Strengths. Multilinguality. Used by much of the established publishing and library business. Very large amount of concepts and terms, especially species and chemicals. Chemical and enzyme codes are included. Large number of lead-in terms. RT relationships between species and their diseases.

Weaknesses. Not available as (Linked) Open Data, so of limited usefulness in open environments. SKOS version is incomplete and has modeling issues. URI policy still under development. No systematic top-level structure.

5.3. NAL Thesaurus

Strengths. Full coverage of terms in both English and Spanish. Many non-preferred terms and hidden labels. Large amount of chemicals. Available as Linked Data.

Weaknesses. SKOS version is not very rich. Only two languages. Only organized according to the 17 top level categories.

References

Faria et al. 2013 Daniel Faria, Catia Pesquita, Emanuel Santos, Matteo Palmonari, Isabel F. Cruz, Francisco M. Couto. The AgreementMakerLight Ontology Matching System. On the Move to Meaningful Internet Systems: OTM 2013 Conferences. Lecture Notes in Computer Science Volume 8185, 2013, pp 527-541.

van Hage et al. 2010 Willem Robert van Hage, Margherita Sini, Lori Finch, Hap Kolb, and Guus Schreiber. 2010. The OAEI food task: An analysis of a thesaurus alignment task. Applied Ontology 5, 1 (January 2010), pp 1-28.

[49] http://202.45.139.84:10035/catalogs/fao/repositories/agrovoc

Suominen and Mader 2014 Osma Suominen and Christian Mader: Assessing and Improving the Quality of SKOS Vocabularies. Journal on Data Semantics, Volume 3, Issue 1 (March 2014), pp 47-73. Preprint freely available[50]

[50] http://www.seco.tkk.fi/publications/2013/suominen-mader-skosquality.pdf

Appendix A. Linguistic coverage diagram

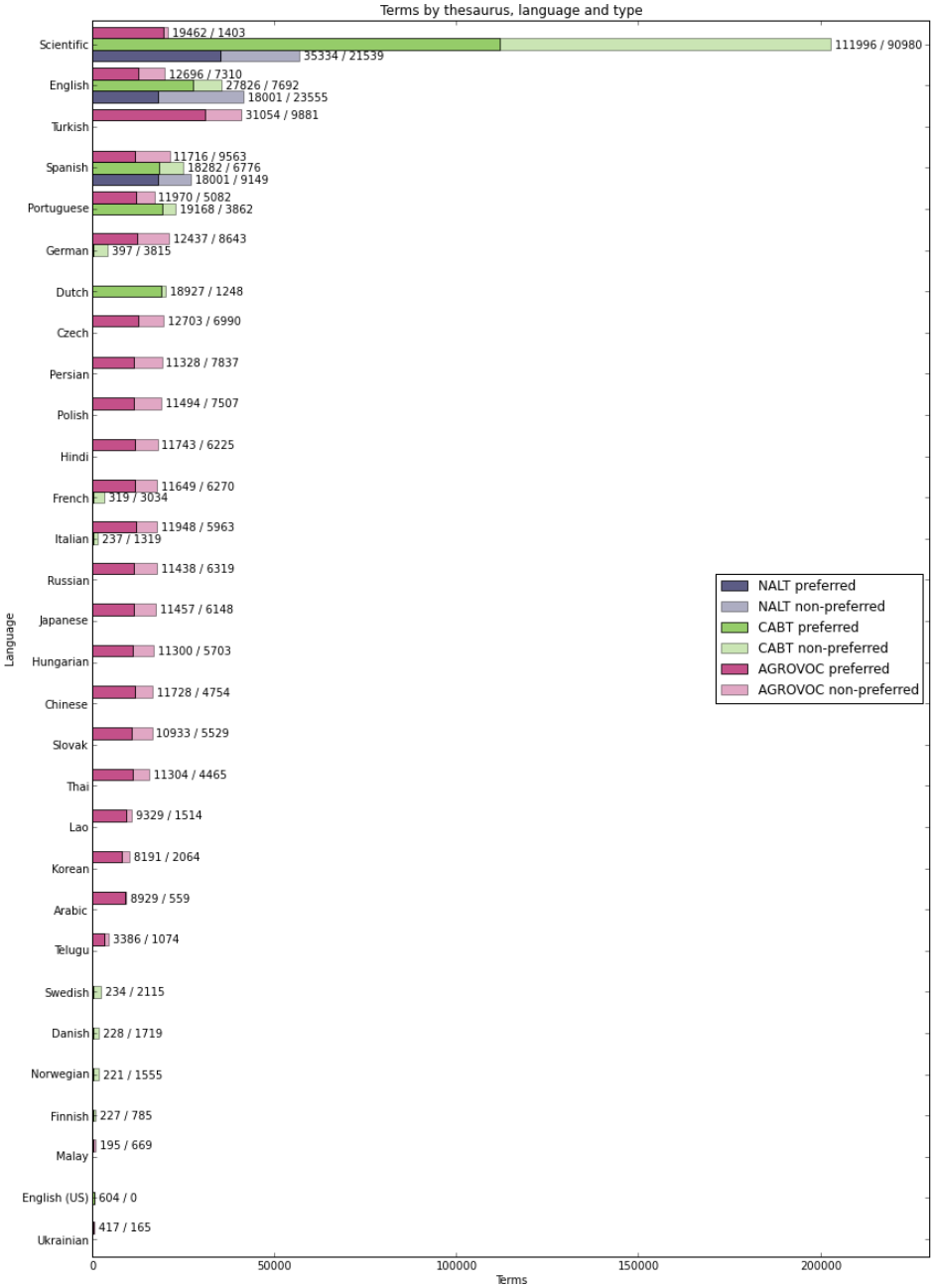

Figure 3: Linguistic coverage overview

Appendix B. RDF classes, properties and relationships

RDF classes and properties used in the SKOS versions

	AGROVOC	CABT	NALT
skos:Concept	X	X	X
skos:ConceptScheme	X	-	-
foaf:Image, foaf:depiction	X	-	-
skosxl:Label	X	-	-
dcterms:created, dcterms:modified	X	X	-
rdfs:comment	X	-	-
skos:prefLabel, skos:altLabel	LOD only	X	X
skos:notation	X	-	-
skosxl:prefLabel, skosxl:altLabel, skosxl:literalForm	X	-	-
skos:definition	X	X	X
skos:scopeNote	X	X	X
skos:editorialNote	X	X	-
skos:historyNote	-	X	-
skos:broader	X	X	X
skos:narrower	-	X	X
skos:related	X	X	X
skos:inScheme, skos:topConceptOf	X	-	-
Custom	Agrontology, VocBench	Various custom fields	-

Table 4: RDF classes and properties used in the SKOS versions of each thesaurus.

Custom concept-level properties and relationships

AGROVOC	CABT	Purpose	Example
isPartOfSubvocabulary	Subject Category	Classifies concepts by subject/theme.	See section 4.3.
	Related Term Broader / Related Term Narrower	Hierarchical relationship that bridges two distinct hierarchies.	*1-propanol* RTB *solvents*
hasTaxonomicLevel		Links an organism to a concept that represents a taxonomic level.	*Boletales* hasTaxonomicLevel *order (taxa)*
	Disease Name / Disease Agent	Links between a disease and an organism causing the disease.	*Claviceps purpurea* DSN *ergot*
isUsedAs		Links a concept to another concept describing its usage.	*Opium* isUsedAs *Analgesics*
influences			*Freshness* influences *Quality*
includes			*males* includes *sons*
hasMember		Connects a concept representing a group to a member concept.	*Tropical fruits* hasMember *Figs*
makeUseOf			*trends* makeUseOf *Forecasting*
hasComponent			*Cola* hasComponent *caffeine*
spatiallyIncludes			*Boreal forests* spatiallyIncludes *Arctic tundra*
causes			*Runoff* causes *Water erosion*
produces	Harvested Product / Crop Plant	Links producer to product. Narrower usage in CABT between plants and their products.	*honey bees* produces *Beeswax*

Table 5: Custom concept-level properties and relationships in AGROVOC and CABT. Similar properties shown on the same row. The table includes all Agrontology custom concept properties that have been used more than 500 times in AGROVOC.

Custom term-level properties and relationships

AGROVOC	CABT	Purpose	Example
hasTermType	Technical Category	Classifies terms into types.	See section 4.3.
hasSpellingVariant		Links a term to a spelling variant, represented as a literal string.	*"haemophilia"* hasSpellingVariant *"hemophilia"*
hasSynonym		Links a term to its synonym term.	*"Herbicides"* hasSynonym *"Weed killers"*
hasNearSynonym		Links a term to a near synonym term.	*"Agribusiness"* hasNearSynonym *"Agroindustrial sector"*
hasBroaderSynonym		Links a term to a synonym term with broader meaning.	*"Siltation"* hasBroaderSynonym *"sedimentation"*
hasAcronym		Links a term to its acronym.	*"Value added tax"* hasAcronym *"VAT (tax)"*
hasOldName		Links a term to an older term.	*"Commonwealth of Nations"* hasOldName *"British Commonwealth"*
hasAbbreviation	Technical Category	Links a term to an abbreviation.	*"mitochondrial DNA"* hasAbbreviation *"mtDNA"*
hasSymbol		Links a term to a symbol.	*"Oxygen"* hasSymbol *"O (symbol)"*
hasRelatedTerm		Links a term to a related term.	*"Belgium"* hasRelatedTerm *"West Flanders"*
hasScientificName	Scientific Name / Common Name	Links a common name term to the scientific name (and vice versa).	*"spruce"* hasScientificName *"Picea"*

Table 6: Custom term-level properties and relationships in AGROVOC and CABT. Similar properties shown on the same row. The table includes all Agrontology custom term properties that have been used more than 500 times in AGROVOC.

Appendix C. Icicle diagrams

These "icicle" visualizations, created using the TreeViz[51] tool, show the overall shape of the three thesauri. The top level concepts are shown in the leftmost column, and lowel levels of the hierarchy are shown progressively towards the right edge. The concepts are colored according to the subsets: **green** = **species**, **red** = **chemicals**, **blue** = **places** and **grey** = **other**.

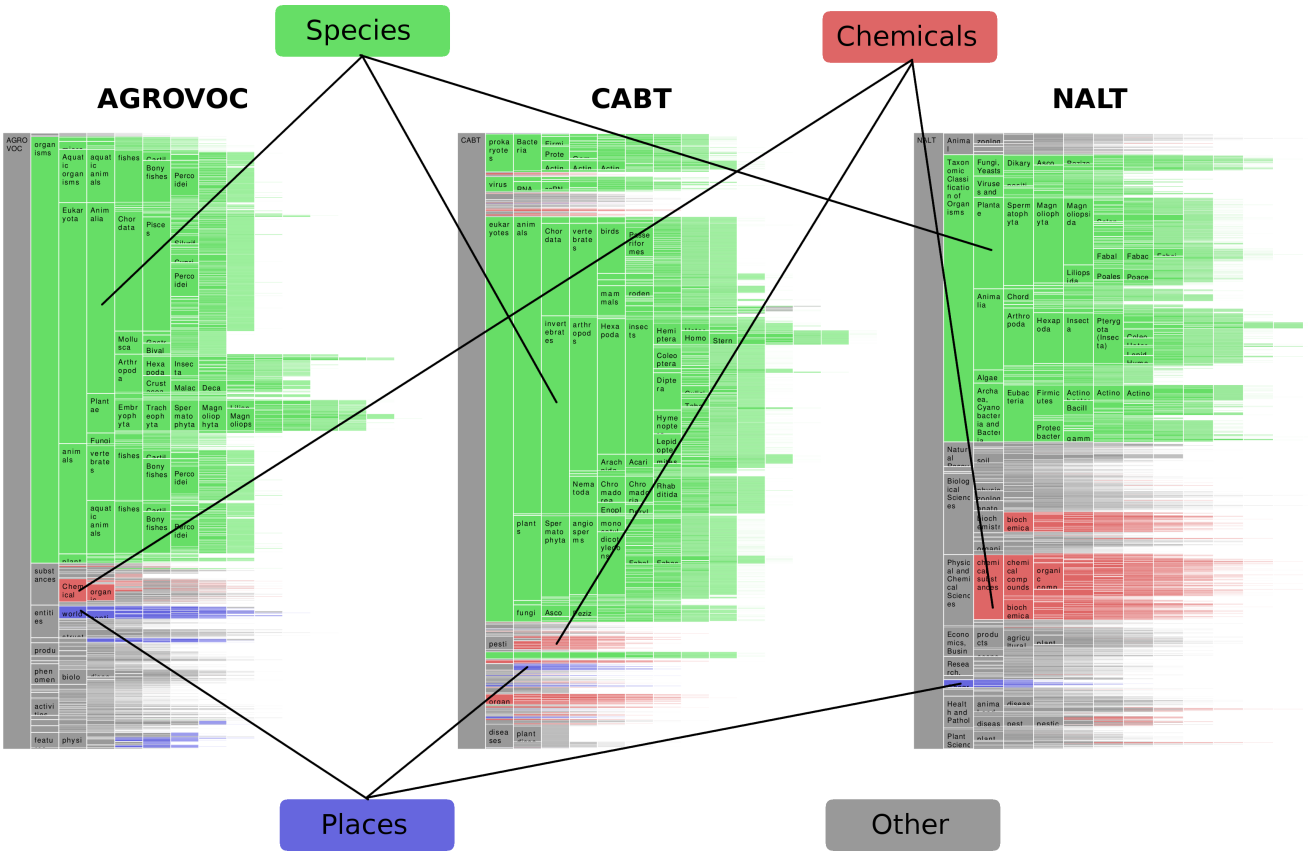

Figure 4: Icicle visualization of the three thesauri.

[51] http://www.randelshofer.ch/treeviz/

Appendix D. Subset overlap

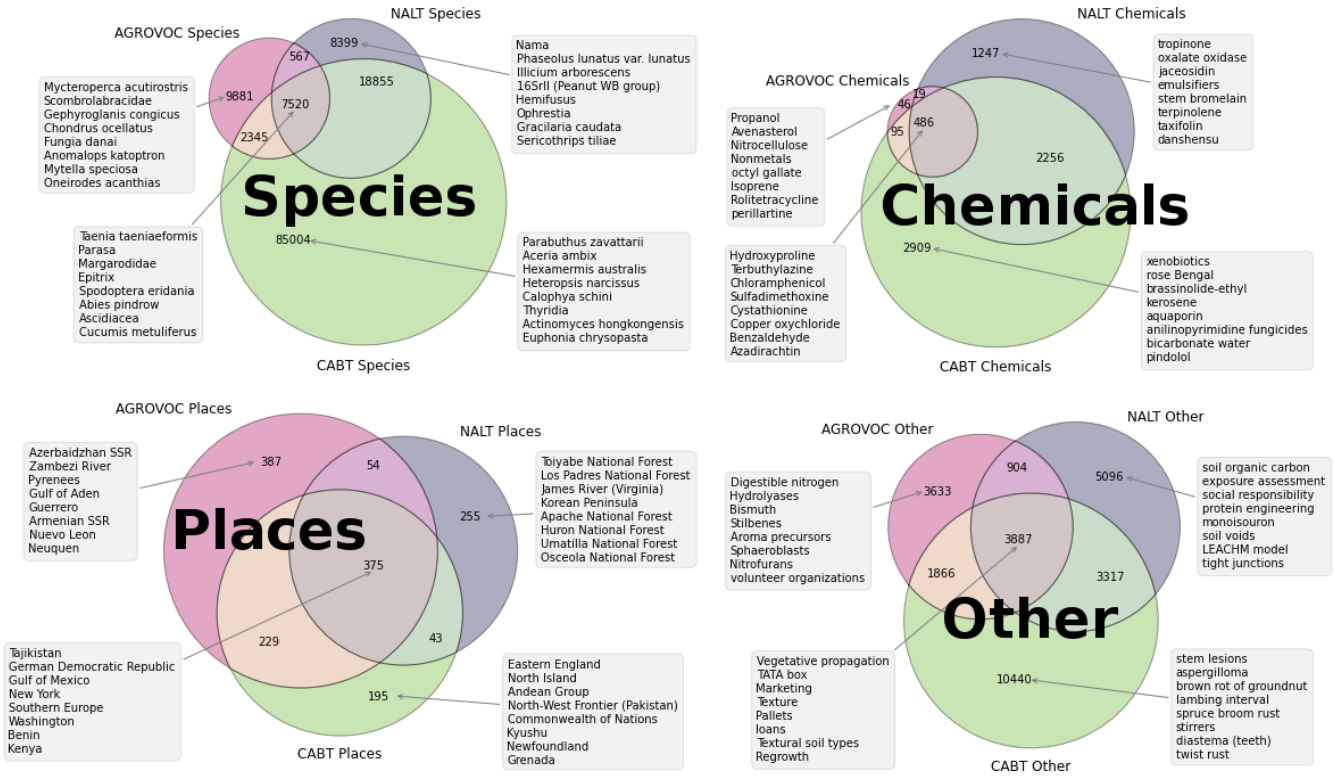

Figure 5: Estimated overlap of the four subsets.