Global Agricultural Concept Scheme

The collaborative integration of three thesauri

Prof Dr Thomas Baker [1]
Dr Osma Suominen [2]

Dini Jahrestagung
“Linked Data – Vision und Wirklichkeit”
Frankfurt, 28. Oktober 2015

[1] Sungkyunkwan University (Korea) and Dublin Core Metadata Initiative
[2] National Library of Finland

Three big thesauri in agriculture

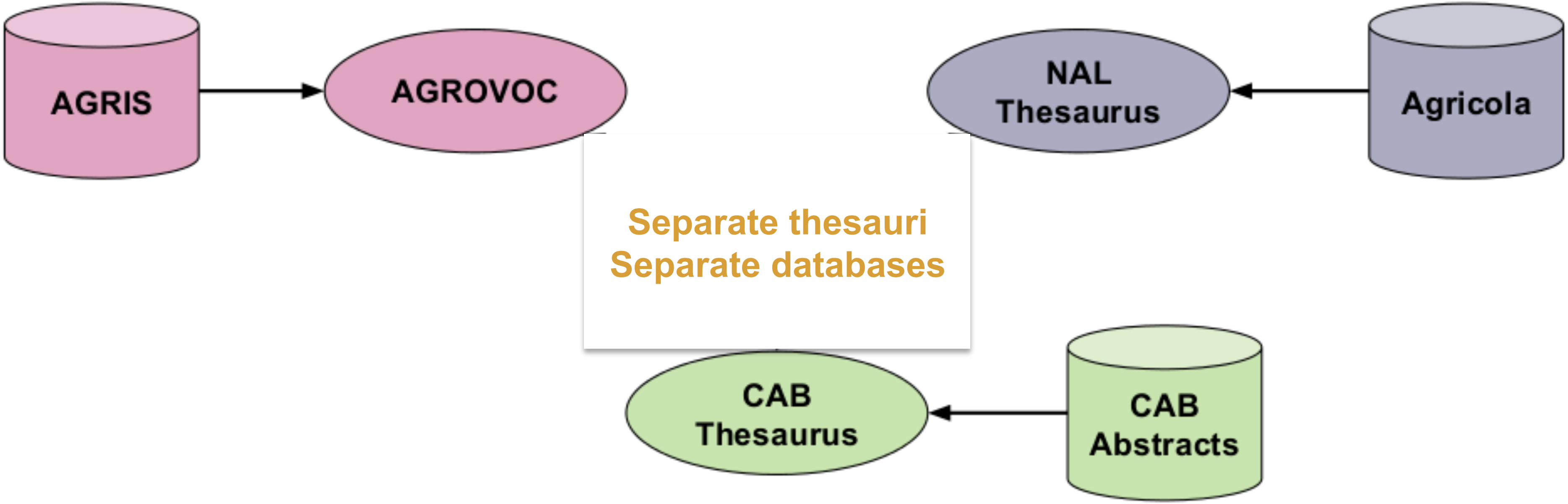
Three thesauri of terms and concepts related to agriculture -- concepts like ***rice***, ***ricefield aquaculture***, and ***plant pests***.

- FAO – Food and Agriculture Organization of the United Nations
- CABI – Centre for Biosciences and Agriculture International (UK)
- NAL – National Agricultural Library (US)

FAO FIAT PANIS

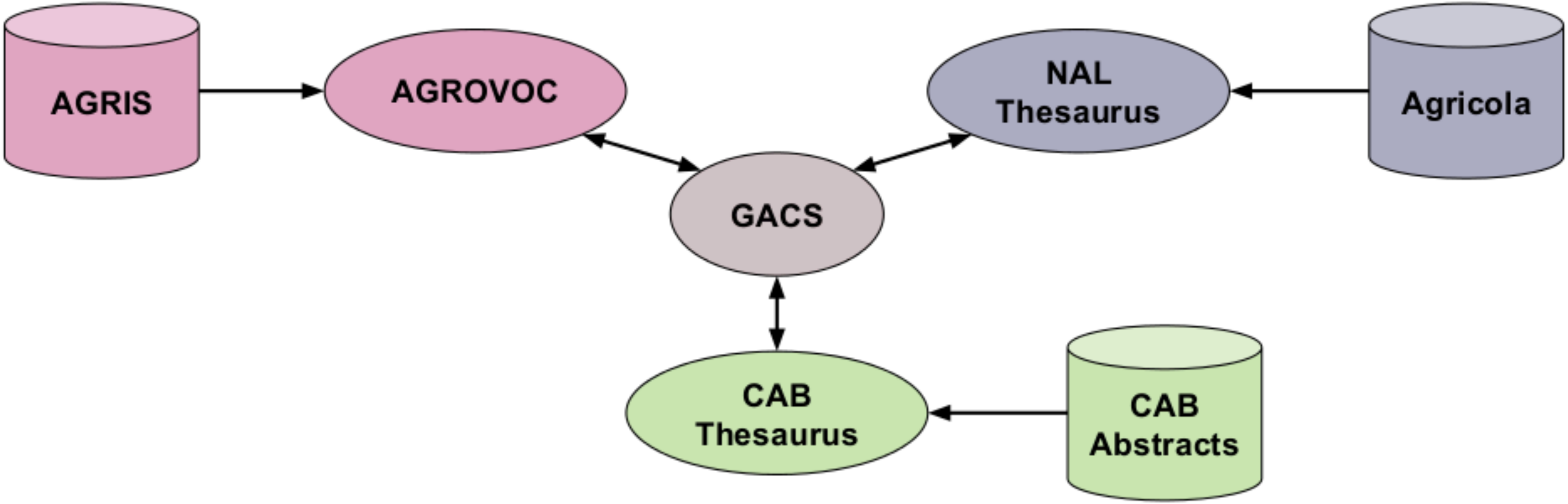
www.cabi.org

National
Agricultural
Library



Separate thesauri
Separate databases

Create GACS as glue linking them together



Global Agricultural Concept Scheme (GACS)

1. Improve semantic interoperability of the thesauri
2. Provide core concepts.
3. Achieve efficiencies through cooperative maintenance.

Requirements

1. Integrated view
2. Reuse of work, such as translations
3. Compatibility with existing databases
4. Based on RDF technologies: URIs, SKOS...
5. Available as Linked Open Data

Based on, mapped to, but independent of, its three source thesauri.

AGROVOC

32,000 concepts, >1.2M labels

English, Spanish, Portuguese, German, Czech, Persian, Polish, Hindi, French, Italian, Russian, Japanese, Hungarian, Chinese, Slovak, Thai, Lao, Turkish, Korean, Arabic, Telugu ...

CAB Thesaurus

140,000 concepts, >1.4M labels

English, Spanish, Portuguese, Dutch

+ many languages with lower coverage

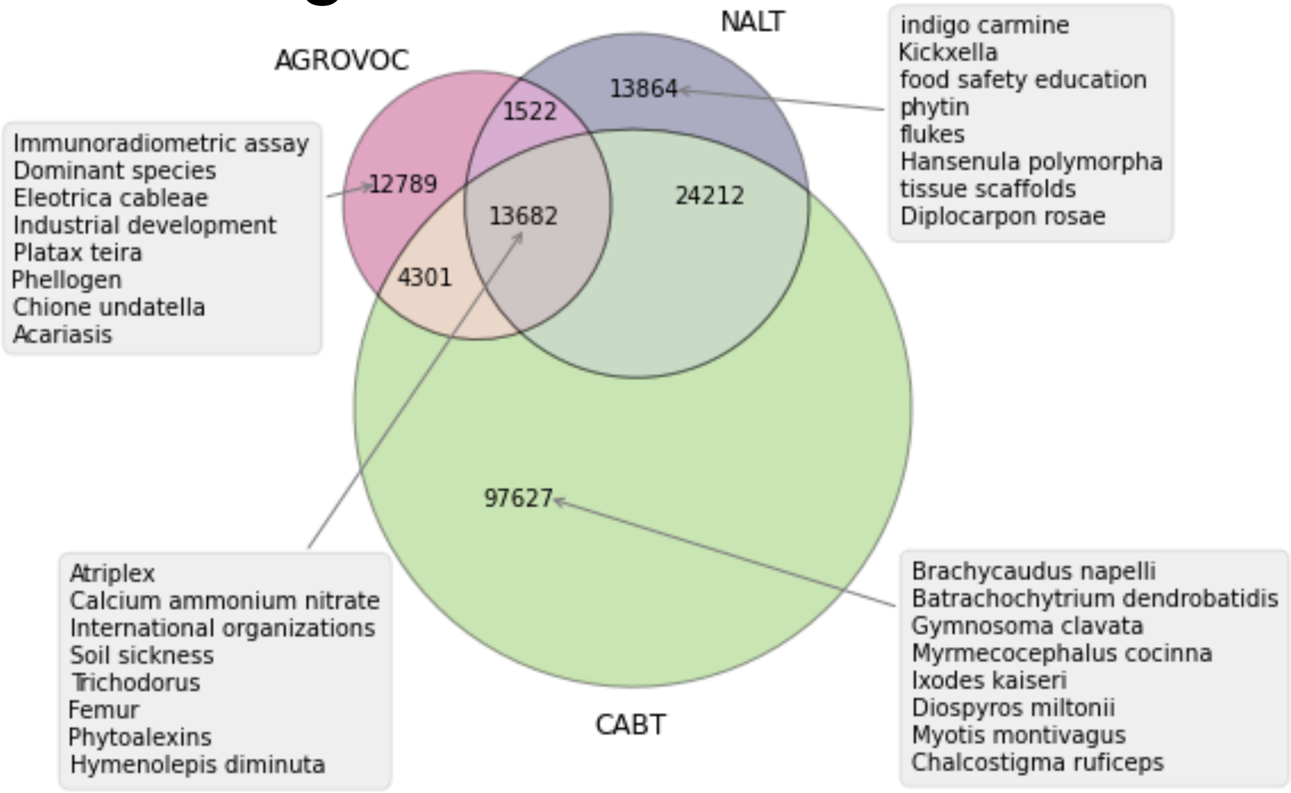
NAL Thesaurus

53,000 concepts, >200k labels

English, Spanish

First step: represent all three in SKOS

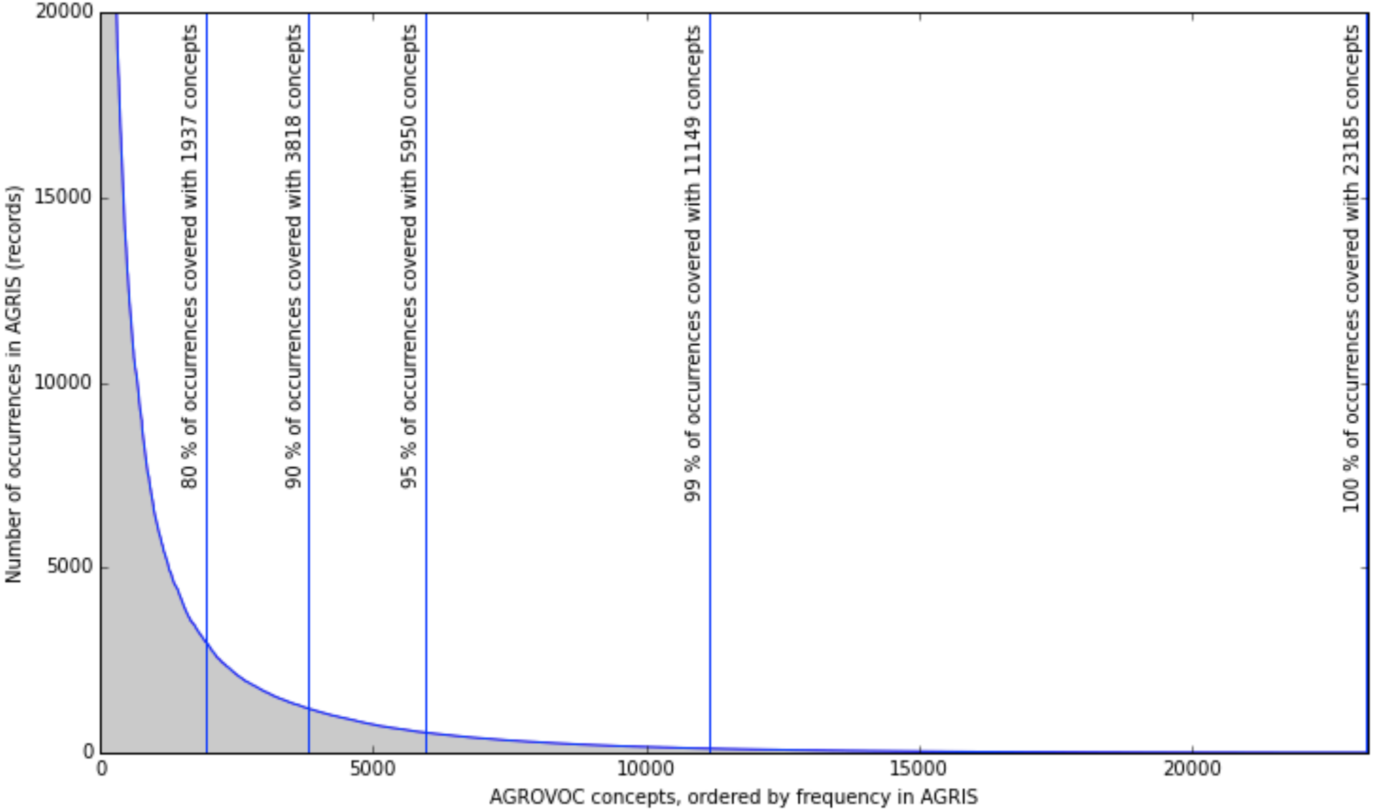
First rough estimate



Obtained via automatic mappings created using AgreementMakerLight

Long tail distribution (in AGRIS)

10,000 concepts cover nearly 99% of occurrences in metadata

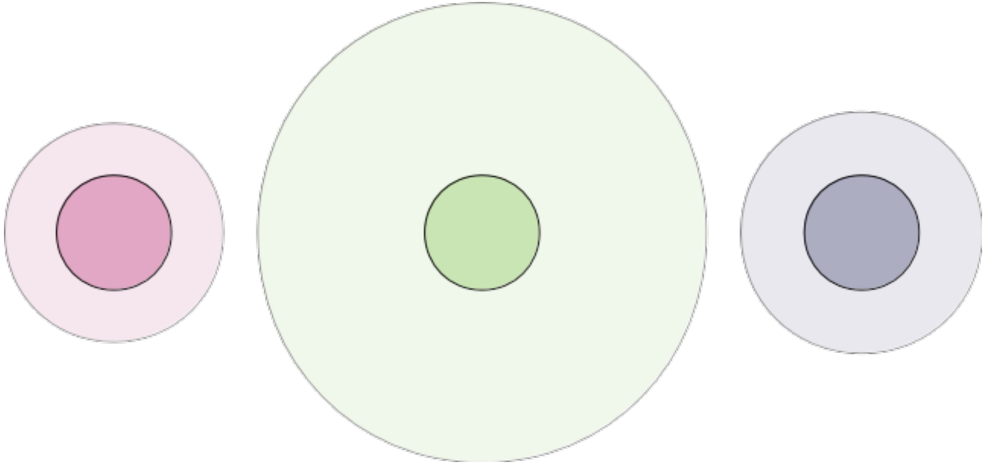


Top 10,000 concepts from each

Each partner organization provided the 10,000 concepts most frequently used in their respective databases.

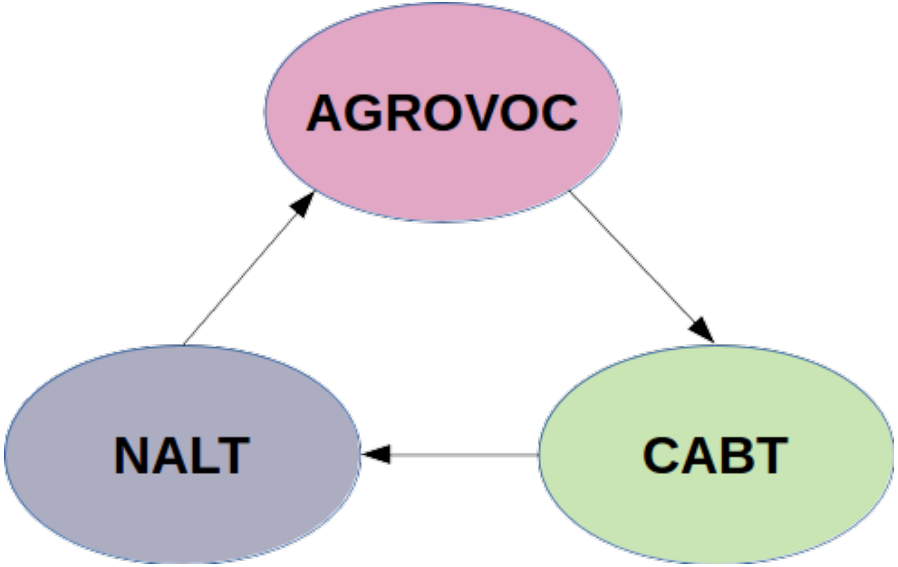
Added:

- all countries
- all higher-level organisms



Automated mappings

Created using AgreementMakerLight software between the full thesauri, for completeness

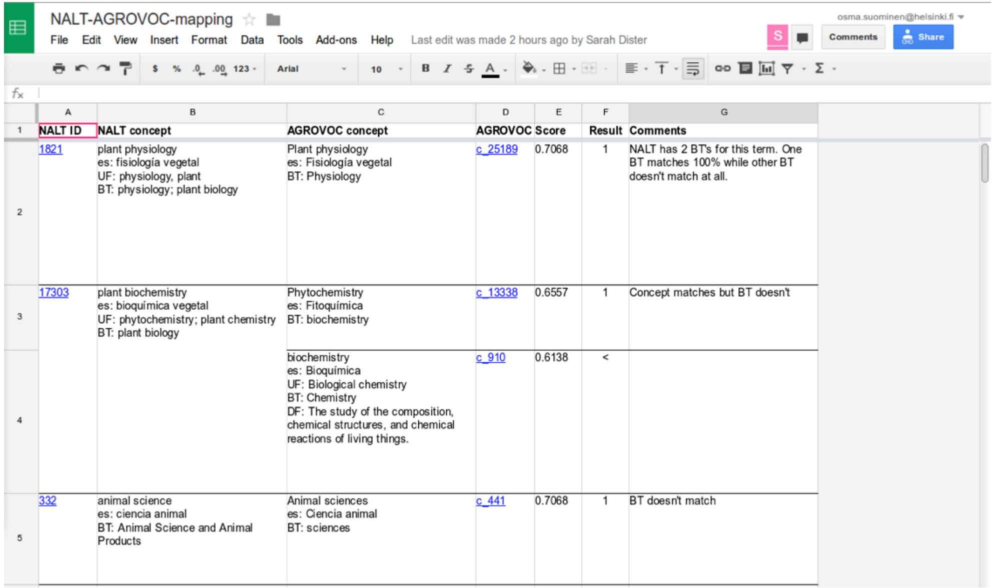


Human evaluation of mappings

Created Google Docs spreadsheets using the lists of selected concepts and the auto-generated mappings. Three sheets with circa 10,700 rows each.

Mappings manually evaluated by staff of partner organizations.

Evaluated 60 to 150 rows/hour. Evaluation took 500 to 600 hours for GACS Beta.



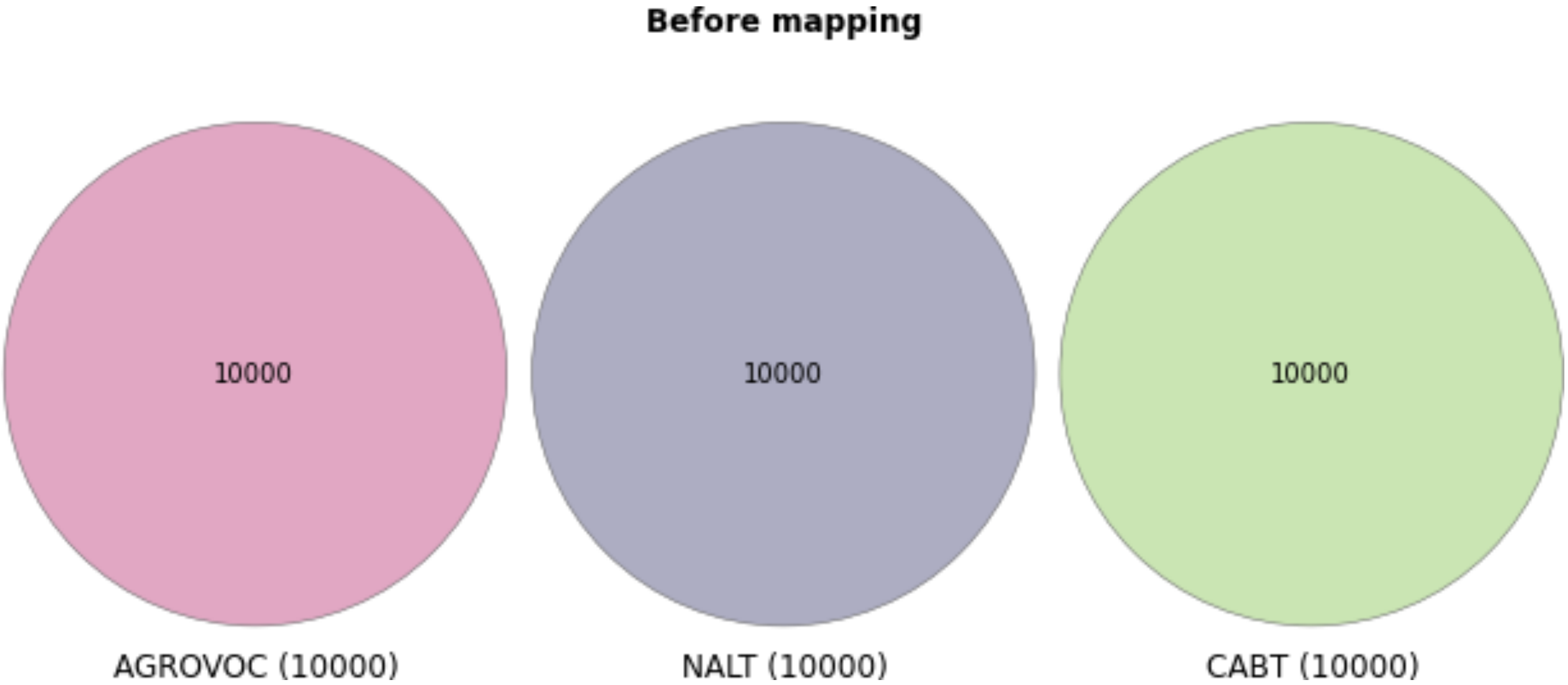
NALT-AGROVOC-mapping

Last edit was made 2 hours ago by Sarah Dister

	A	B	C	D	E	F	G
1	NALT ID	NALT concept	AGROVOC concept	AGROVOC	Score	Result	Comments
2	1821	plant physiology es: fisiología vegetal UF: physiology, plant BT: physiology; plant biology	Plant physiology es: Fisiología vegetal BT: Physiology	c_25189	0.7068	1	NALT has 2 BTs for this term. One BT matches 100% while other BT doesn't match at all.
3	17303	plant biochemistry es: bioquímica vegetal UF: phytochemistry; plant chemistry BT: plant biology	Phytochemistry es: Fitoquímica BT: biochemistry	c_13338	0.6557	1	Concept matches but BT doesn't
4			biochemistry es: Bioquímica UF: Biological chemistry BT: Chemistry DF: The study of the composition, chemical structures, and chemical reactions of living things.	c_910	0.6138	<	
5	332	animal science es: ciencia animal BT: Animal Science and Animal Products	Animal sciences es: Ciencia animal BT: sciences	c_441	0.7068	1	BT doesn't match

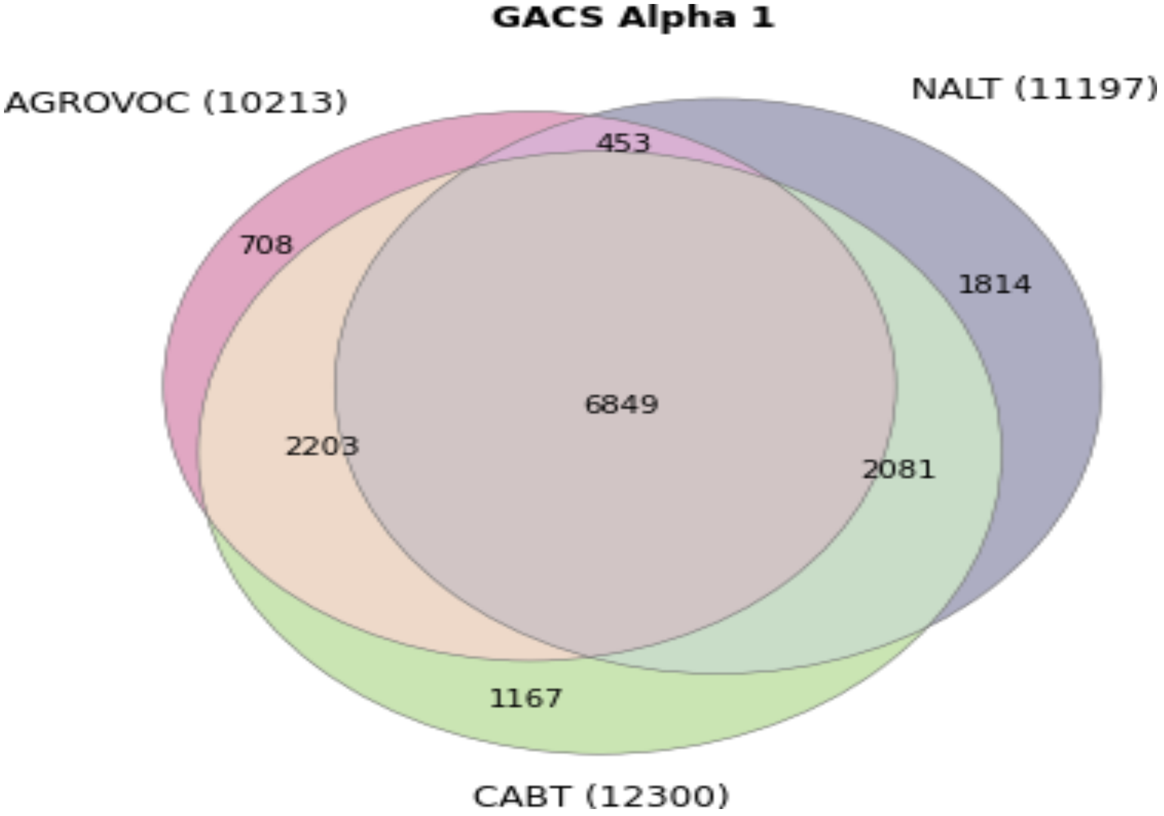
Starting point

October 2014



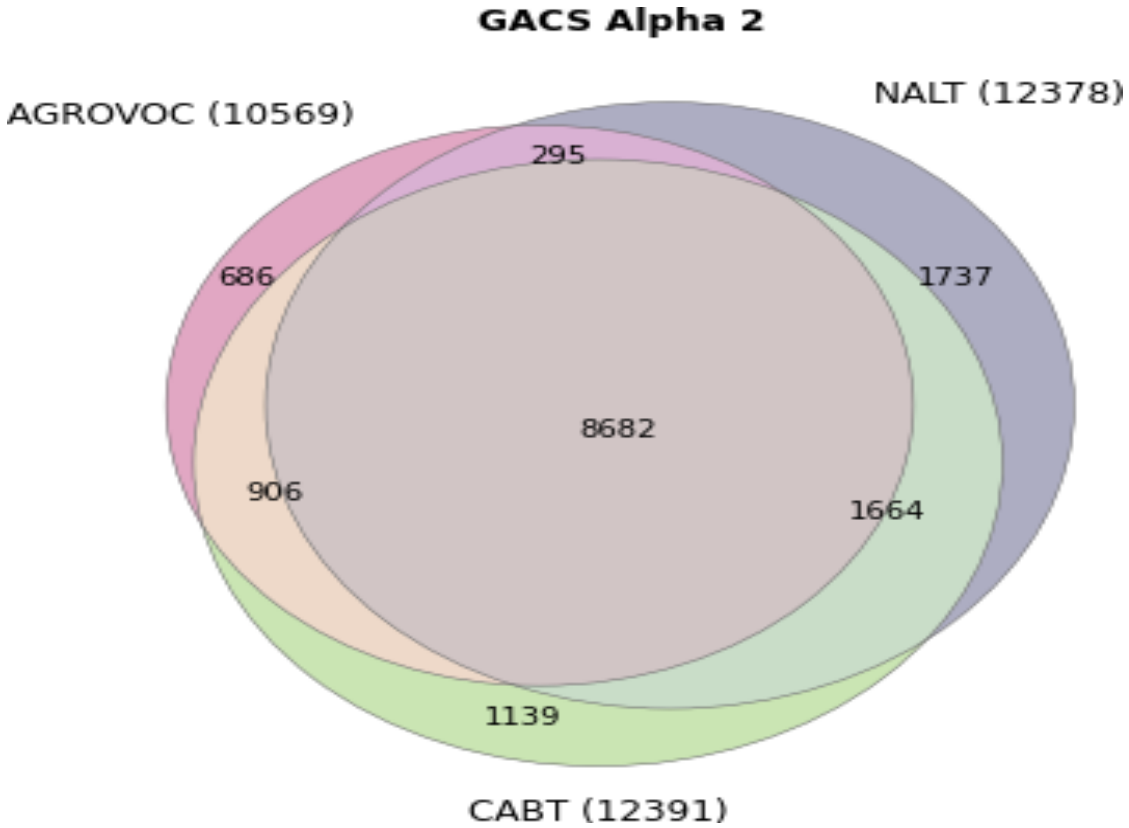
30,000 mappings later...

January 2015



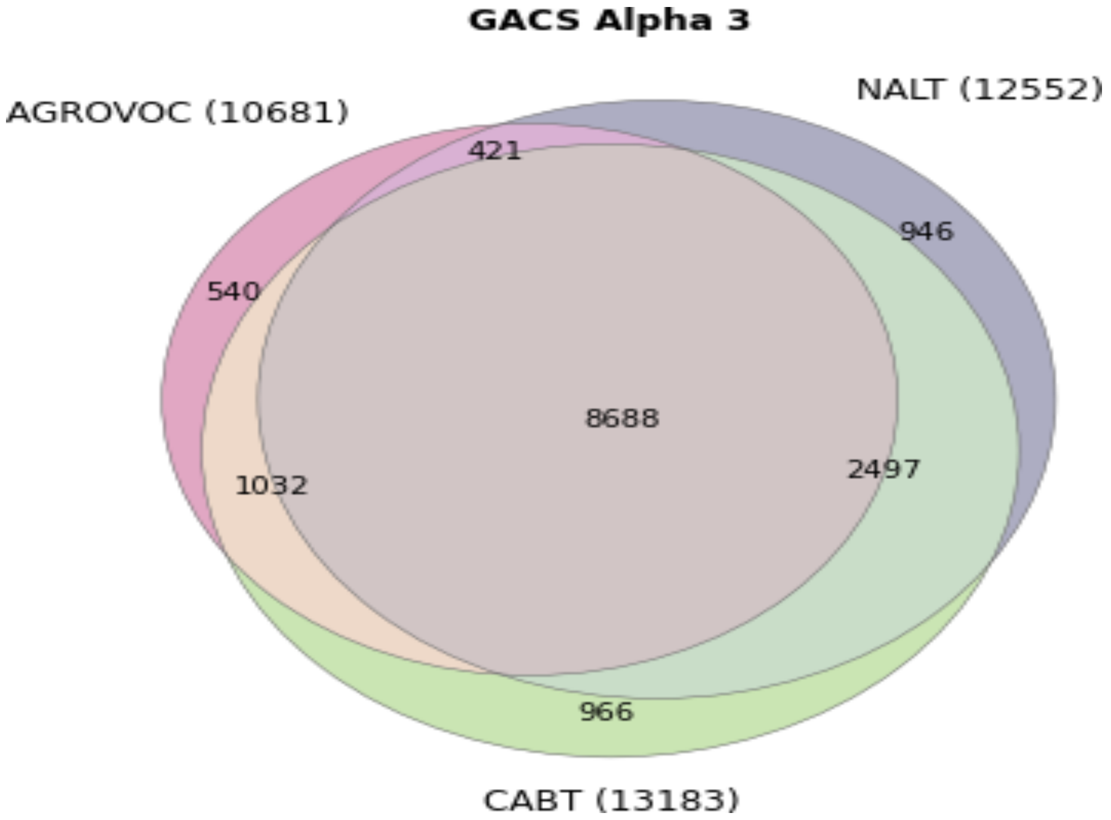
4,689 mappings later...

February 2015



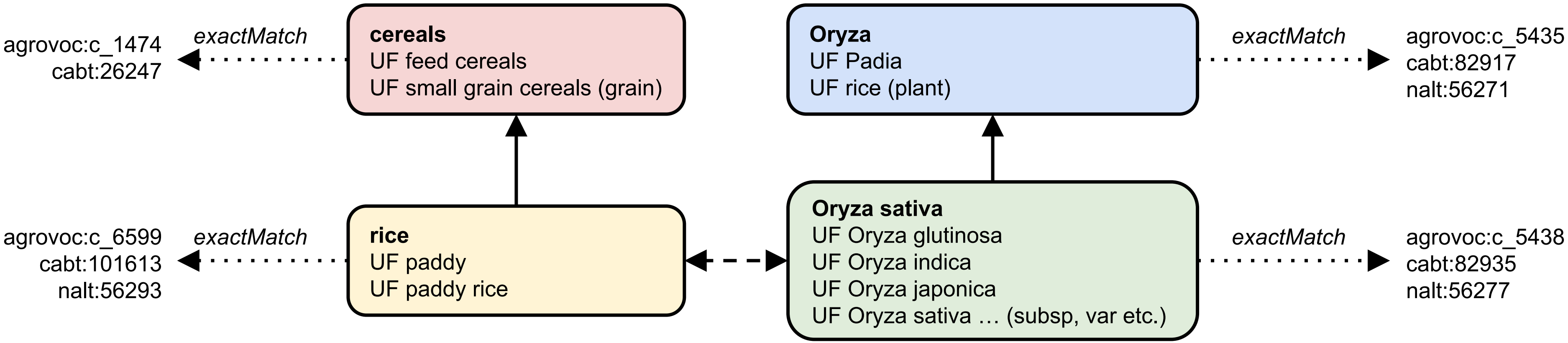
5,522 mappings later...

March 2015



Forming GACS concepts

by merging the source concepts and aggregating their information



agrovoc:c_1474
cabt:26247

exactMatch

cereals
UF feed cereals
UF small grain cereals (grain)

Oryza
UF Padia
UF rice (plant)

exactMatch

agrovoc:c_5435
cabt:82917
nalt:56271

agrovoc:c_6599
cabt:101613
nalt:56293

exactMatch

rice
UF paddy
UF paddy rice

Oryza sativa
UF Oryza glutinosa
UF Oryza indica
UF Oryza japonica
UF Oryza sativa … (subsp, var etc.)

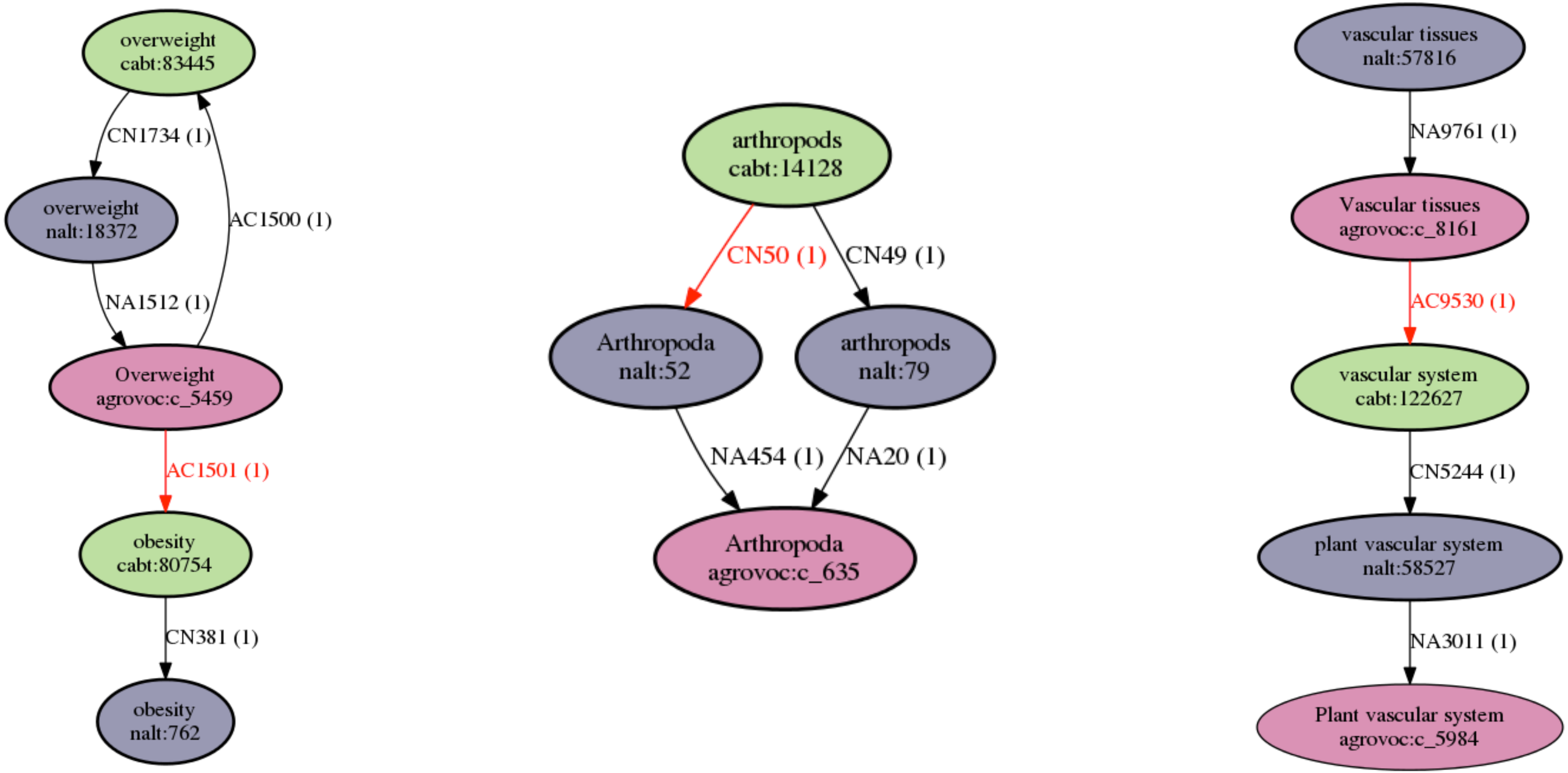
exactMatch

agrovoc:c_5438
cabt:82935
nalt:56277

(Note: GACS uses SKOS, not traditional thesaurus tags)

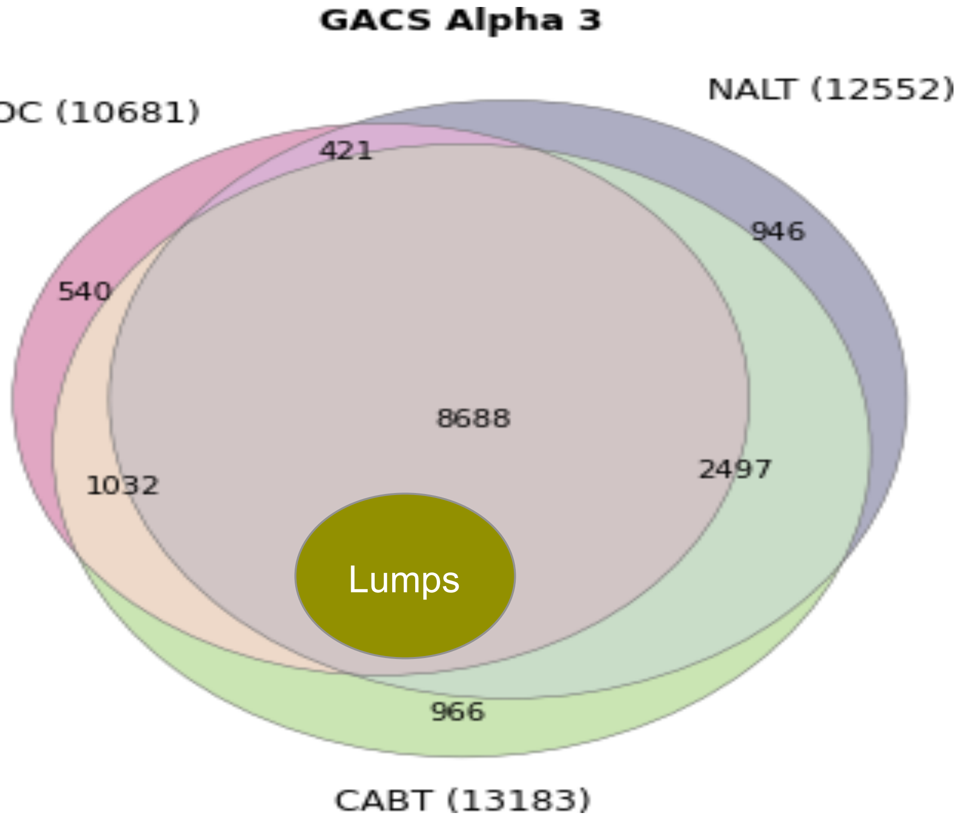
Lumps

clusters of concepts mapped one-to-several, several-to-one, or in spirals



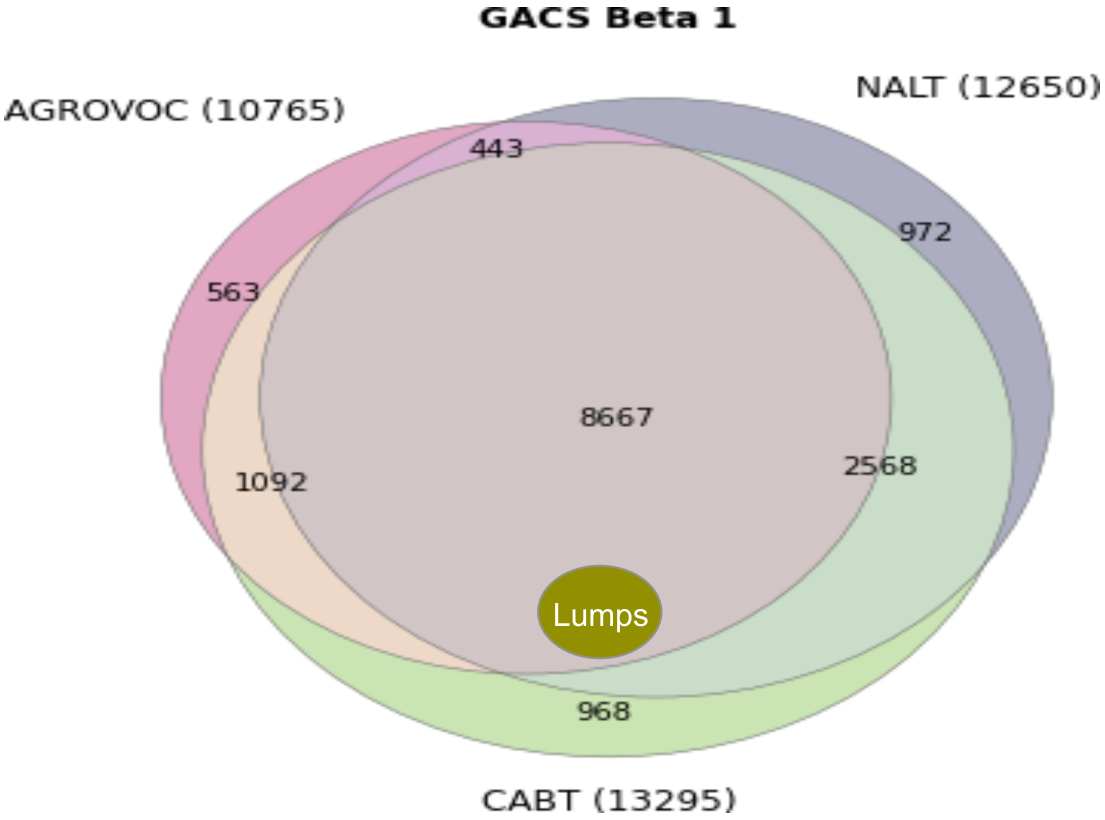
15,090 concepts; 972 lumps

March 2015



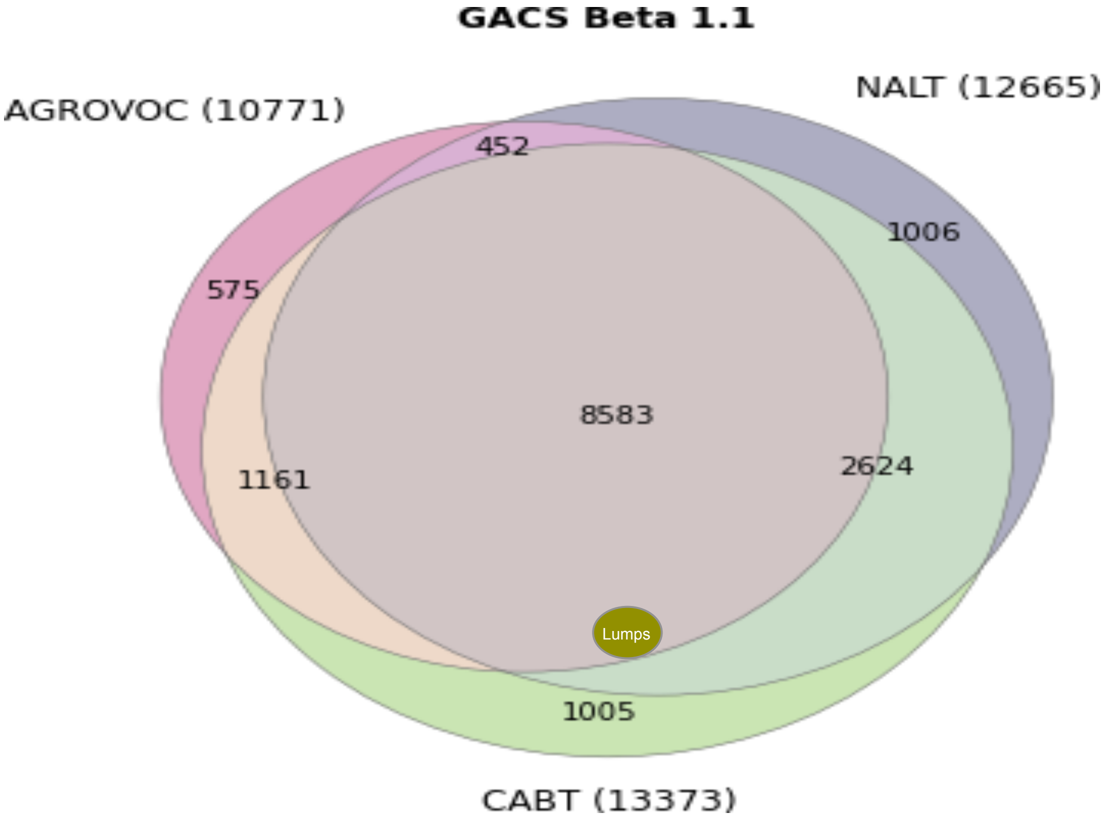
15,278 concepts; 339 lumps

April 2015



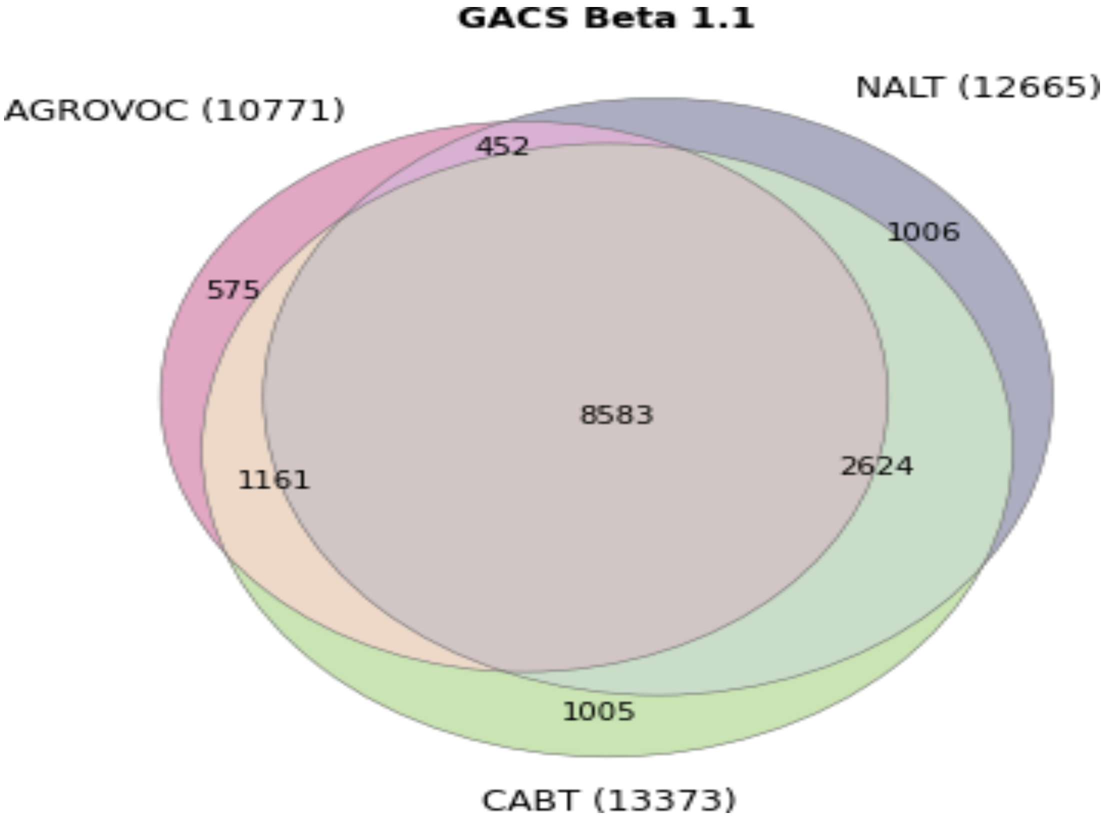
15,411 concepts; 84 lumps

October 2015

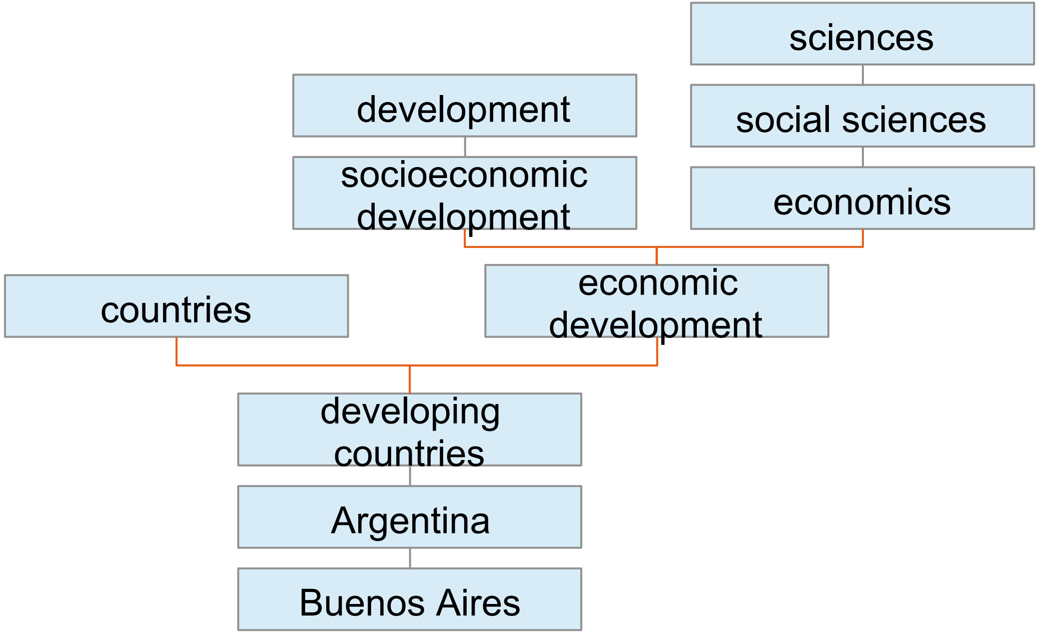


15,406 concepts; no lumps

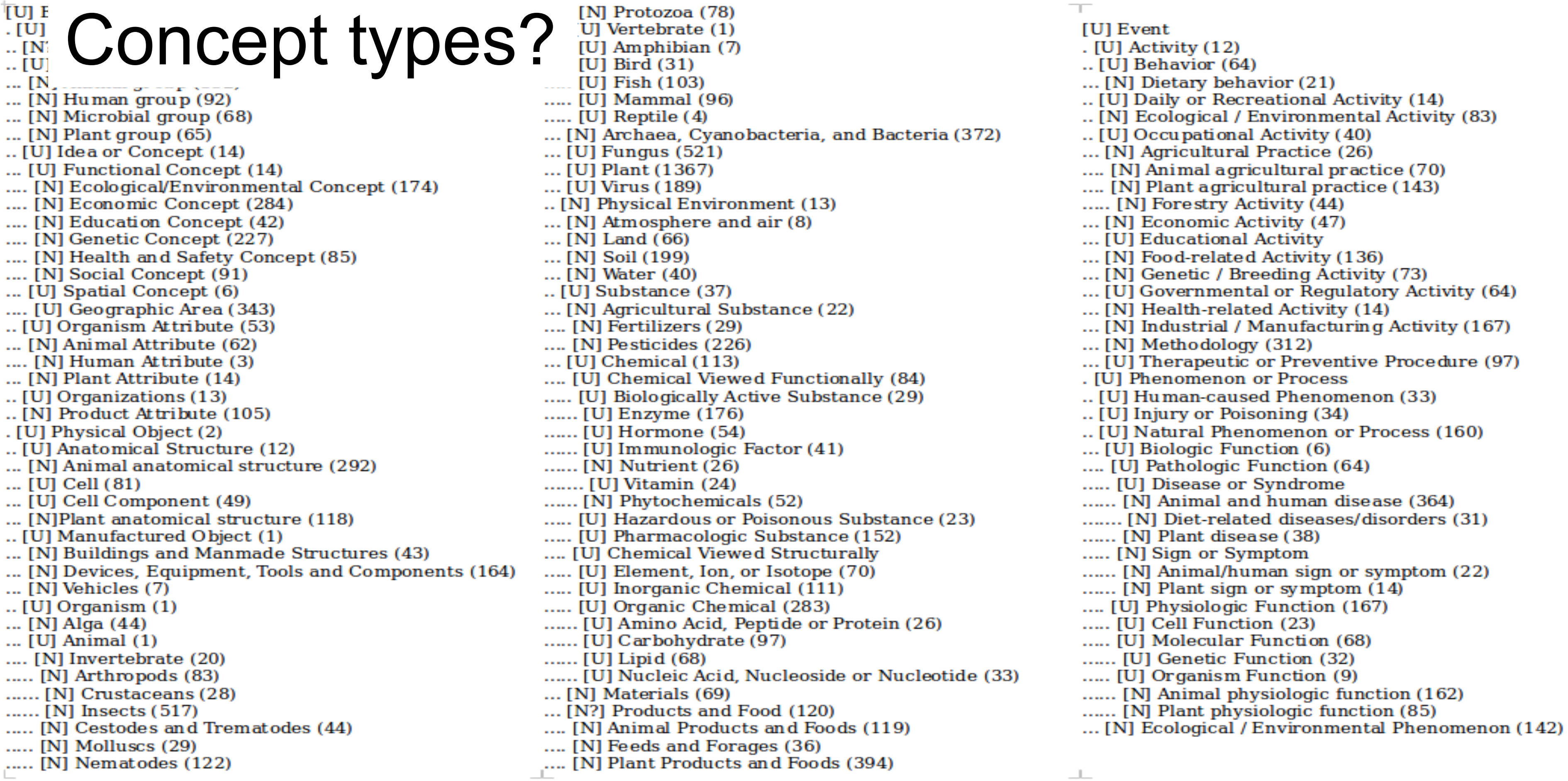
Last week



Polyhierarchy?



Concept types?

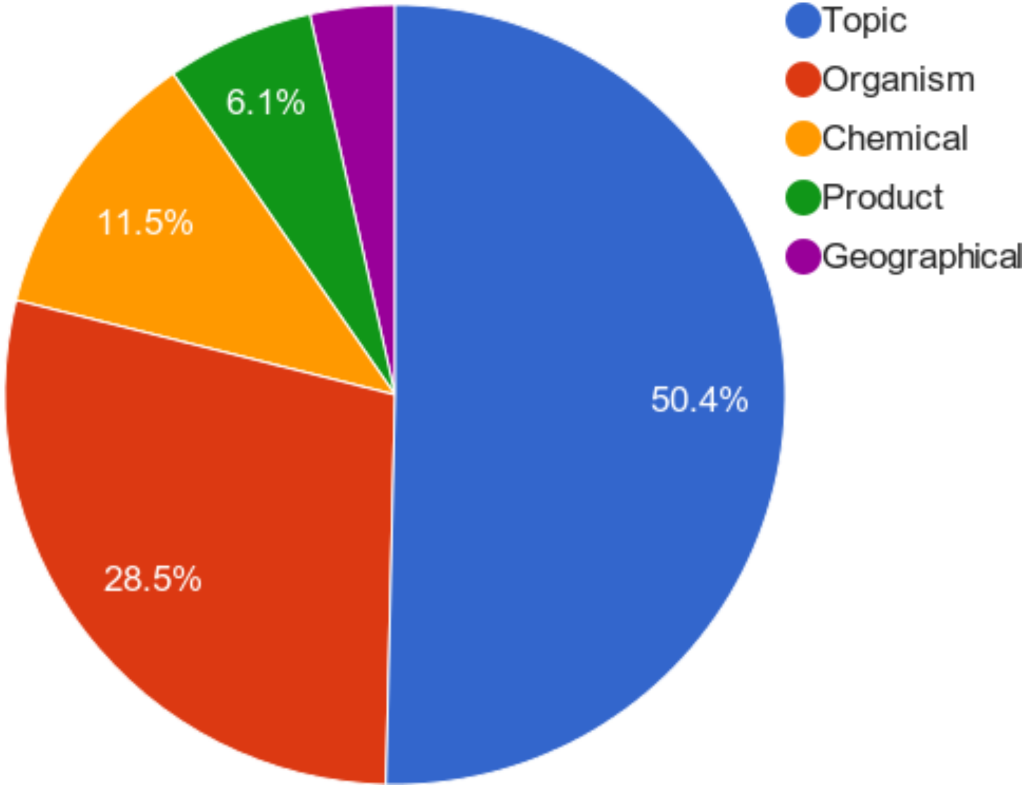


[U] E
. [U]
.. [N?
.. [U]
... [N]
... [N] Human group (92)
... [N] Microbial group (68)
... [N] Plant group (65)
.. [U] Idea or Concept (14)
... [U] Functional Concept (14)
.... [N] Ecological/Environmental Concept (174)
.... [N] Economic Concept (284)
.... [N] Education Concept (42)
.... [N] Genetic Concept (227)
.... [N] Health and Safety Concept (85)
.... [N] Social Concept (91)
... [U] Spatial Concept (6)
.... [U] Geographic Area (343)
.. [U] Organism Attribute (53)
... [N] Animal Attribute (62)
.... [N] Human Attribute (3)
... [N] Plant Attribute (14)
.. [U] Organizations (13)
.. [N] Product Attribute (105)
. [U] Physical Object (2)
.. [U] Anatomical Structure (12)
... [N] Animal anatomical structure (292)
... [U] Cell (81)
... [U] Cell Component (49)
... [N]Plant anatomical structure (118)
.. [U] Manufactured Object (1)
... [N] Buildings and Manmade Structures (43)
... [N] Devices, Equipment, Tools and Components (164)
... [N] Vehicles (7)
.. [U] Organism (1)
... [N] Alga (44)
... [U] Animal (1)
.... [N] Invertebrate (20)
..... [N] Arthropods (83)
...... [N] Crustaceans (28)
...... [N] Insects (517)
..... [N] Cestodes and Trematodes (44)
..... [N] Molluscs (29)
..... [N] Nematodes (122)

[N] Protozoa (78)
U] Vertebrate (1)
[U] Amphibian (7)
[U] Bird (31)
..... [U] Fish (103)
..... [U] Mammal (96)
..... [U] Reptile (4)
... [N] Archaea, Cyanobacteria, and Bacteria (372)
... [U] Fungus (521)
... [U] Plant (1367)
... [U] Virus (189)
.. [N] Physical Environment (13)
... [N] Atmosphere and air (8)
... [N] Land (66)
... [N] Soil (199)
... [N] Water (40)
.. [U] Substance (37)
... [N] Agricultural Substance (22)
.... [N] Fertilizers (29)
.... [N] Pesticides (226)
... [U] Chemical (113)
.... [U] Chemical Viewed Functionally (84)
..... [U] Biologically Active Substance (29)
...... [U] Enzyme (176)
...... [U] Hormone (54)
...... [U] Immunologic Factor (41)
...... [N] Nutrient (26)
.... [U] Vitamin (24)
...... [N] Phytochemicals (52)
..... [U] Hazardous or Poisonous Substance (23)
..... [U] Pharmacologic Substance (152)
.... [U] Chemical Viewed Structurally
..... [U] Element, Ion, or Isotope (70)
..... [U] Inorganic Chemical (111)
..... [U] Organic Chemical (283)
...... [U] Amino Acid, Peptide or Protein (26)
...... [U] Carbohydrate (97)
...... [U] Lipid (68)
...... [U] Nucleic Acid, Nucleoside or Nucleotide (33)
... [N] Materials (69)
... [N?] Products and Food (120)
.... [N] Animal Products and Foods (119)
.... [N] Feeds and Forages (36)
.... [N] Plant Products and Foods (394)

[U] Event
. [U] Activity (12)
.. [U] Behavior (64)
... [N] Dietary behavior (21)
.. [U] Daily or Recreational Activity (14)
.. [N] Ecological / Environmental Activity (83)
.. [U] Occupational Activity (40)
... [N] Agricultural Practice (26)
.... [N] Animal agricultural practice (70)
.... [N] Plant agricultural practice (143)
..... [N] Forestry Activity (44)
... [N] Economic Activity (47)
... [U] Educational Activity
... [N] Food-related Activity (136)
... [N] Genetic / Breeding Activity (73)
... [U] Governmental or Regulatory Activity (64)
... [N] Health-related Activity (14)
... [N] Industrial / Manufacturing Activity (167)
... [N] Methodology (312)
... [U] Therapeutic or Preventive Procedure (97)
. [U] Phenomenon or Process
.. [U] Human-caused Phenomenon (33)
.. [U] Injury or Poisoning (34)
.. [U] Natural Phenomenon or Process (160)
... [U] Biologic Function (6)
.... [U] Pathologic Function (64)
..... [U] Disease or Syndrome
...... [N] Animal and human disease (364)
.... [N] Diet-related diseases/disorders (31)
...... [N] Plant disease (38)
..... [N] Sign or Symptom
...... [N] Animal/human sign or symptom (22)
...... [N] Plant sign or symptom (14)
.... [U] Physiologic Function (167)
..... [U] Cell Function (23)
..... [U] Molecular Function (68)
...... [U] Genetic Function (32)
..... [U] Organism Function (9)
...... [N] Animal physiologic function (162)
...... [N] Plant physiologic function (85)
... [N] Ecological / Environmental Phenomenon (142)

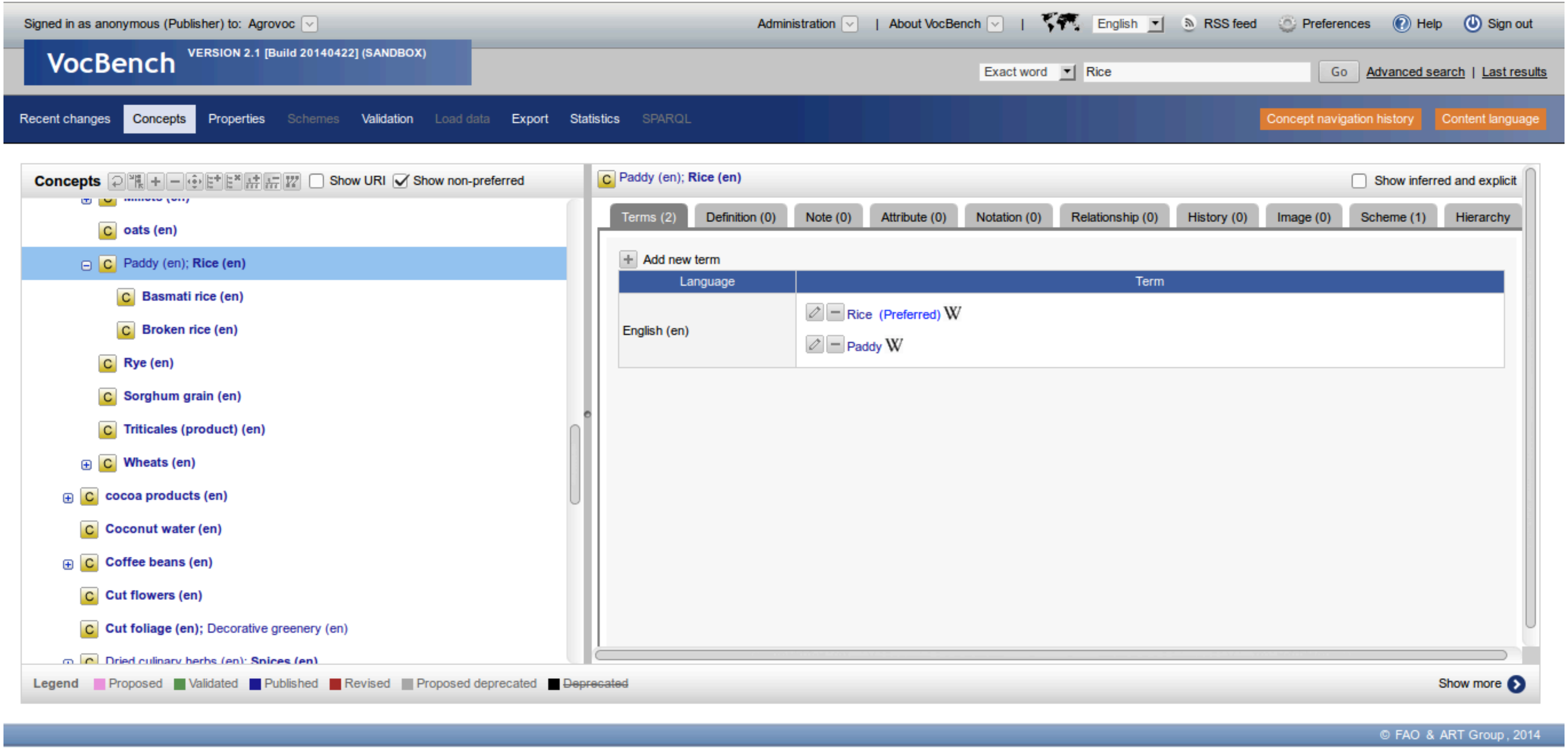
Concept types!



Towards GACS roll-out (2016)

- Concept scheme as Linked Data. Own publication and editorial platform.
- Quality improvements. Inconsistencies in hierarchy, choice of labels, scope notes and definitions.
- Own semantic structure. Common vs scientific names, custom relationships, concept types.

VocBench for editing

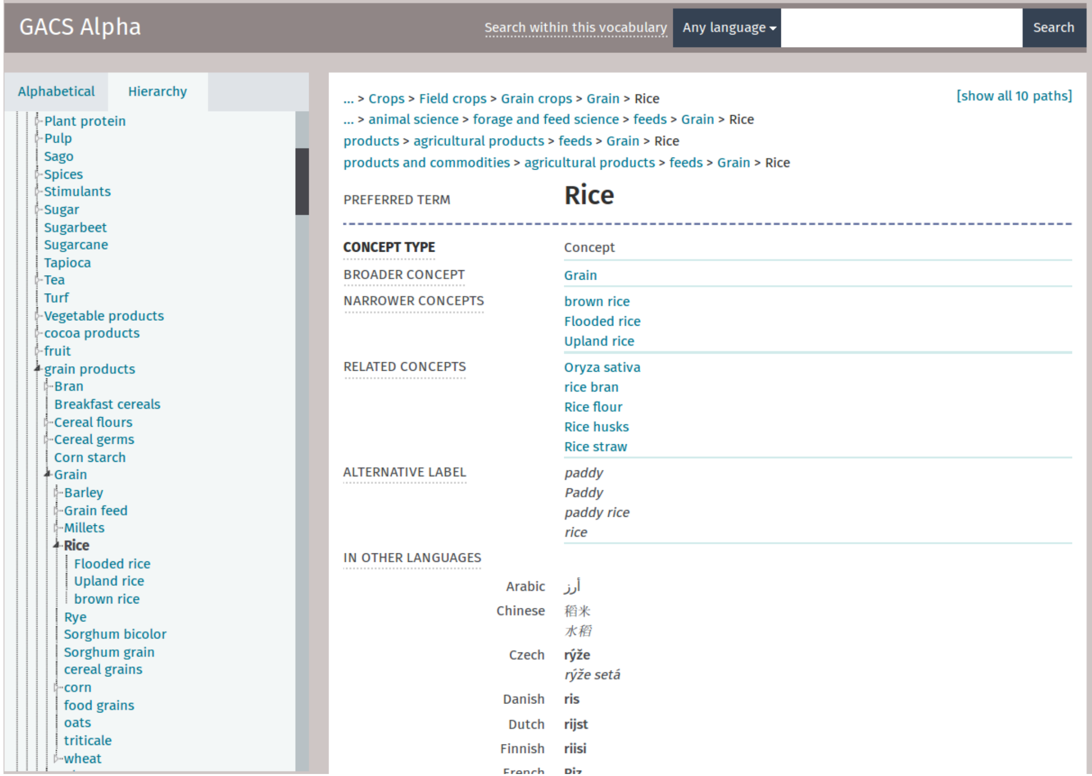


- C oats (en)
- C Paddy (en); Rice (en)
  - C Basmati rice (en)
  - C Broken rice (en)
- C Rye (en)
- C Sorghum grain (en)
- C Triticales (product) (en)
- C Wheats (en)
- C cocoa products (en)
- C Coconut water (en)
- C Coffee beans (en)
- C Cut flowers (en)
- C Cut foliage (en); Decorative greenery (en)
- C Dried culinary herbs (en); Spices (en)

Language	Term
English (en)	Rice (Preferred) W Paddy W

Legend 🟪 Proposed 🟩 Validated 🟦 Published 🟥 Revised ⬜ Proposed deprecated ⬛ ~~Deprecated~~

Skosmos for display and browsing



Plant protein
Pulp
Sago
Spices
Stimulants
Sugar
Sugarbeet
Sugarcane
Tapioca
Tea
Turf
Vegetable products
cocoa products
fruit
grain products
Bran
Breakfast cereals
Cereal flours
Cereal germs
Corn starch
Grain
Barley
Grain feed
Millets
Rice
Flooded rice
Upland rice
brown rice
Rye
Sorghum bicolor
Sorghum grain
cereal grains
corn
food grains
oats
triticale
wheat

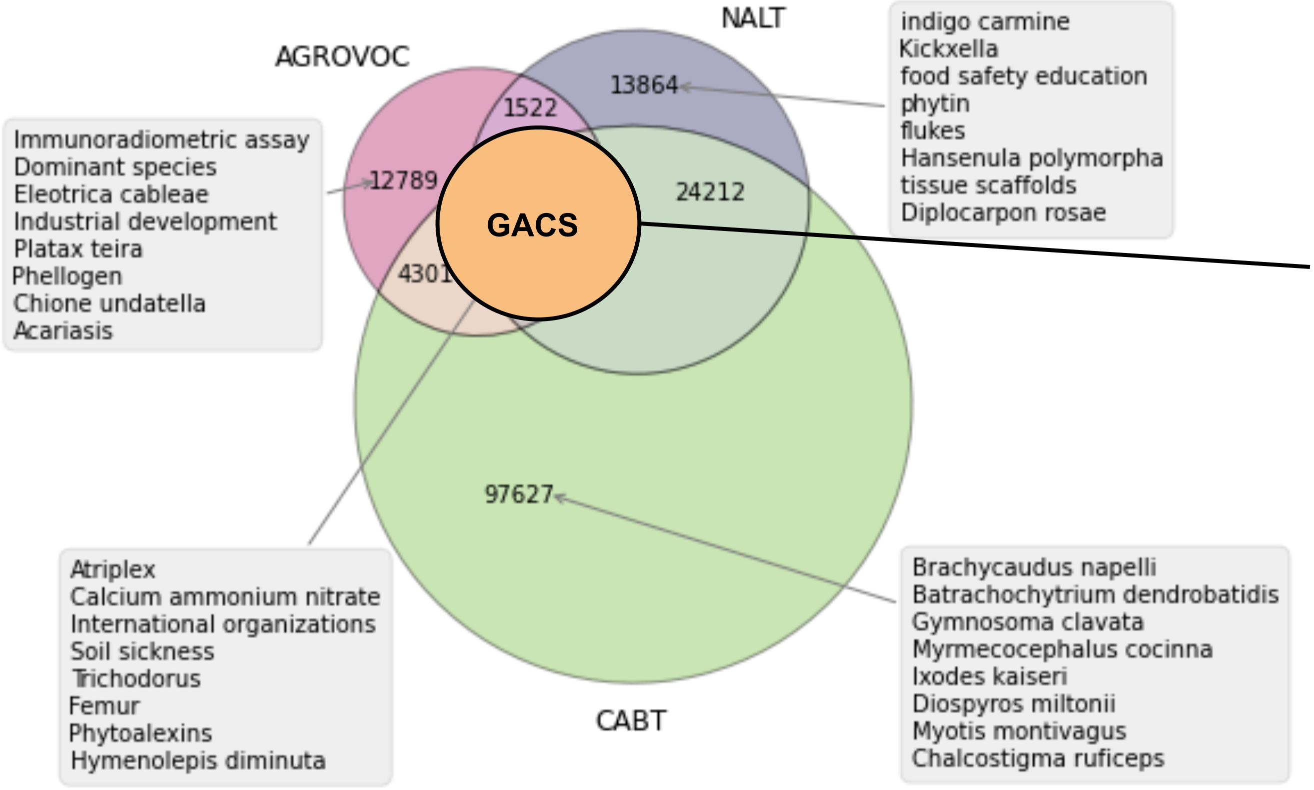
... > Crops > Field crops > Grain crops > Grain > Rice [show all 10 paths]
... > animal science > forage and feed science > feeds > Grain > Rice
products > agricultural products > feeds > Grain > Rice
products and commodities > agricultural products > feeds > Grain > Rice

PREFERRED TERM	**Rice**
CONCEPT TYPE	Concept
BROADER CONCEPT	Grain
NARROWER CONCEPTS	brown rice Flooded rice Upland rice
RELATED CONCEPTS	Oryza sativa rice bran Rice flour Rice husks Rice straw
ALTERNATIVE LABEL	*paddy* *Paddy* *paddy rice* *rice*

IN OTHER LANGUAGES

Arabic	أرز
Chinese	稻米 *水稻*
Czech	**rýže** *rýže setá*
Danish	**ris**
Dutch	**rijst**
Finnish	**riisi**
French	**Riz**

Size of GACS



Immunoradiometric assay
Dominant species
Eleotrica cableae
Industrial development
Platax teira
Phellogen
Chione undatella
Acariasis

indigo carmine
Kickxella
food safety education
phytin
flukes
Hansenula polymorpha
tissue scaffolds
Diplocarpon rosae

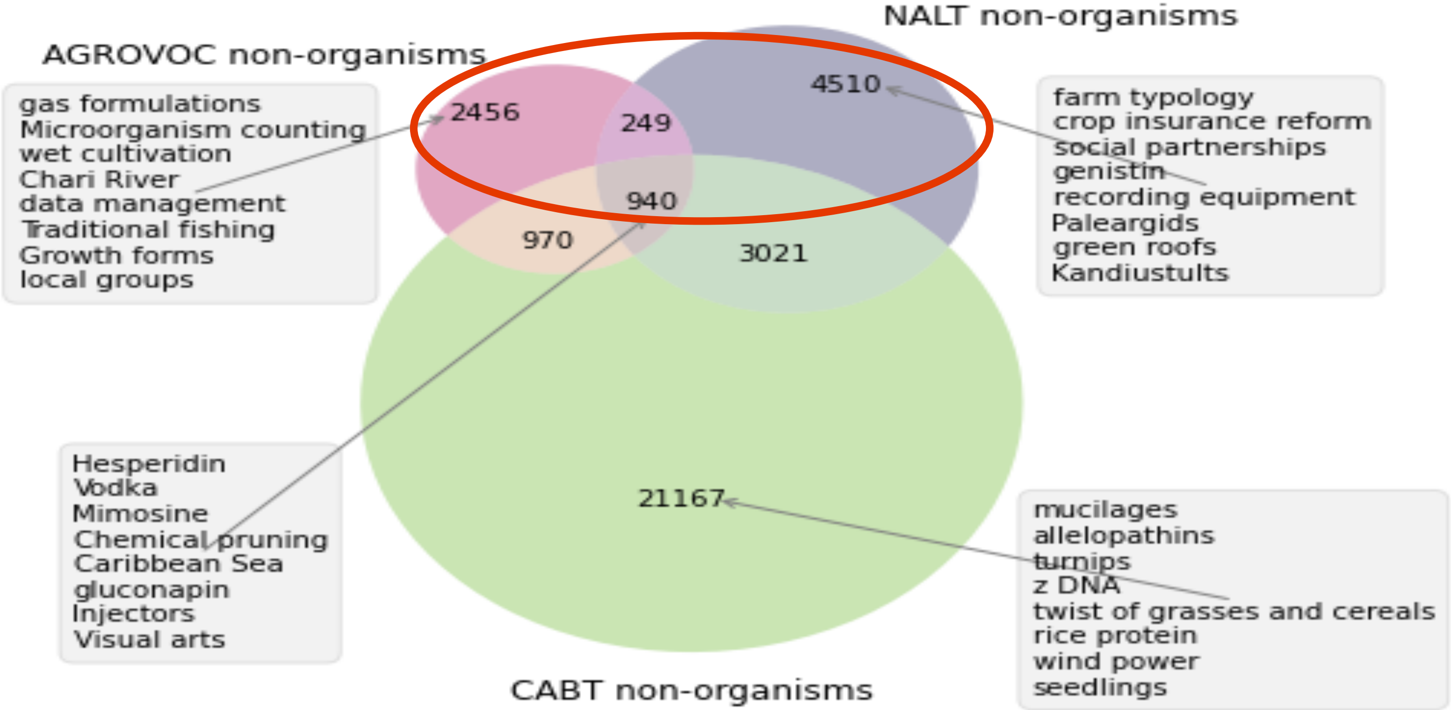
Atriplex
Calcium ammonium nitrate
International organizations
Soil sickness
Trichodorus
Femur
Phytoalexins
Hymenolepis diminuta

Brachycaudus napelli
Batrachochytrium dendrobatidis
Gymnosoma clavata
Myrmecocephalus cocinna
Ixodes kaiseri
Diospyros miltonii
Myotis montivagus
Chalcostigma ruficeps

GACS Beta 1.1

- 15,406 concepts
- 398,216 labels in 28 languages

Extension module for what remains?



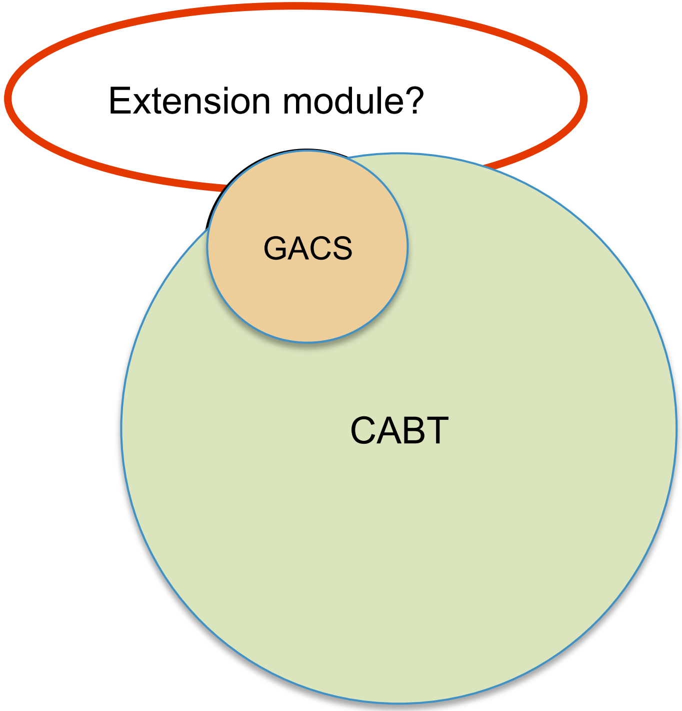
gas formulations
Microorganism counting
wet cultivation
Chari River
data management
Traditional fishing
Growth forms
local groups

farm typology
crop insurance reform
social partnerships
genistin
recording equipment
Paleargids
green roofs
Kandiustults

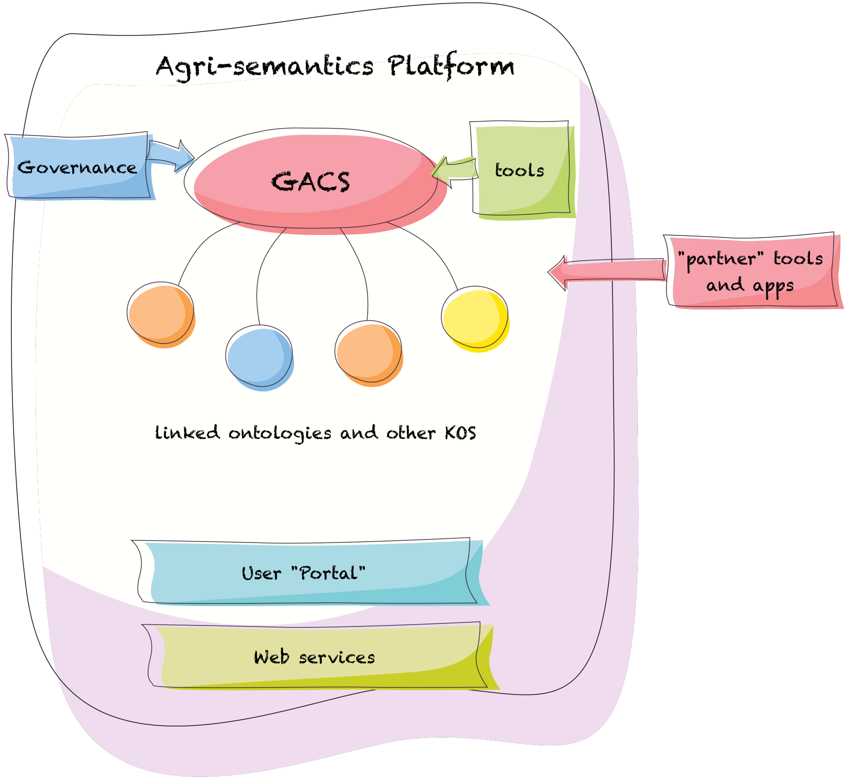
Hesperidin
Vodka
Mimosine
Chemical pruning
Caribbean Sea
gluconapin
Injectors
Visual arts

mucilages
allelopathins
turnips
z DNA
twist of grasses and cereals
rice protein
wind power
seedlings

AGROVOC and NALT may be phased out



Agrisemantics



http://aims.fao.org/sites/default/files/Report_workshop_Agrisemantics.pdf

Global food security and climate change

- GACS as hub for agricultural code lists, taxonomies, statistical indicators...
- Simplify data normalization and integration
- More coherent datasets and research results
- Help farmers become more efficient

Reports available on the FAO AIMS site

http://aims.fao.org/community/agrovoc/blogs/phase-one-gacs-approved-read-reports
http://aims.fao.org/sites/default/files/Report_workshop_Agrisemantics.pdf

osma.suominen@helsinki.fi
tom@tombaker.org

Abstract

The Food and Agricultural Organization of the United Nations (FAO), CAB International (CABI), and the USDA National Agricultural Library (NAL), maintainers of three large thesauri of agricultural terminology that largely overlap in scope, have partnered to create a shared Global Agricultural Concept Scheme (GACS). Duplication of effort has proven to be both inefficient and a barrier to users wishing to search across databases indexed with their terms. Expressing AGROVOC, CAB Thesaurus, and NAL Thesaurus in RDF and SKOS, as Linked Data, facilitates mappings, but mappings among three large, continually moving targets are difficult to maintain.

Starting with algorithmically generated mappings among three sets of the terms most frequently used to index the AGRICOLA, CAB Abstracts, and AGRIS databases, thesaurus managers in the GACS Working Group have manually vetted the mappings for quality and are currently correcting logical inconsistencies. In a final iteration, these mappings will be used to generate a Global Agricultural Concept Scheme with its own identifiers, and GACS will be moved into its own distributed editorial environment and jointly maintained by the three partners.

Targeted for beta release in early 2016, GACS aggregates the complementary strengths of its sources, such as expertise in particular areas and labels in twenty languages. Formulating consistent policies for GACS on issues such as scientific versus common names for organisms requires balancing scientific, commercial, educational, and mass-market perspectives. The challenge of global food security under conditions of climate change will require the integration of data at all levels. GACS can serve as a focal point in the broader ecosystem of vocabularies, code lists, database schemas, ontologies, statistical indicators, and taxonomies required to drive agricultural research and innovation.