

Global Agricultural Concept Scheme

Rome, 2 July 2015

Osma Suominen and Thomas Baker

Outline

1. Background
2. Starting point: three thesauri
3. Creating GACS
4. Challenges
5. Next steps and future of GACS

Background

- Food and Agriculture Organization of the UN
- CABI (UK)
- National Agricultural Library (US)

Each organization maintains a thesaurus of terms and concepts related to agriculture -- concepts like ***rice***, ***ricefield aquaculture***, and ***plant pests***.



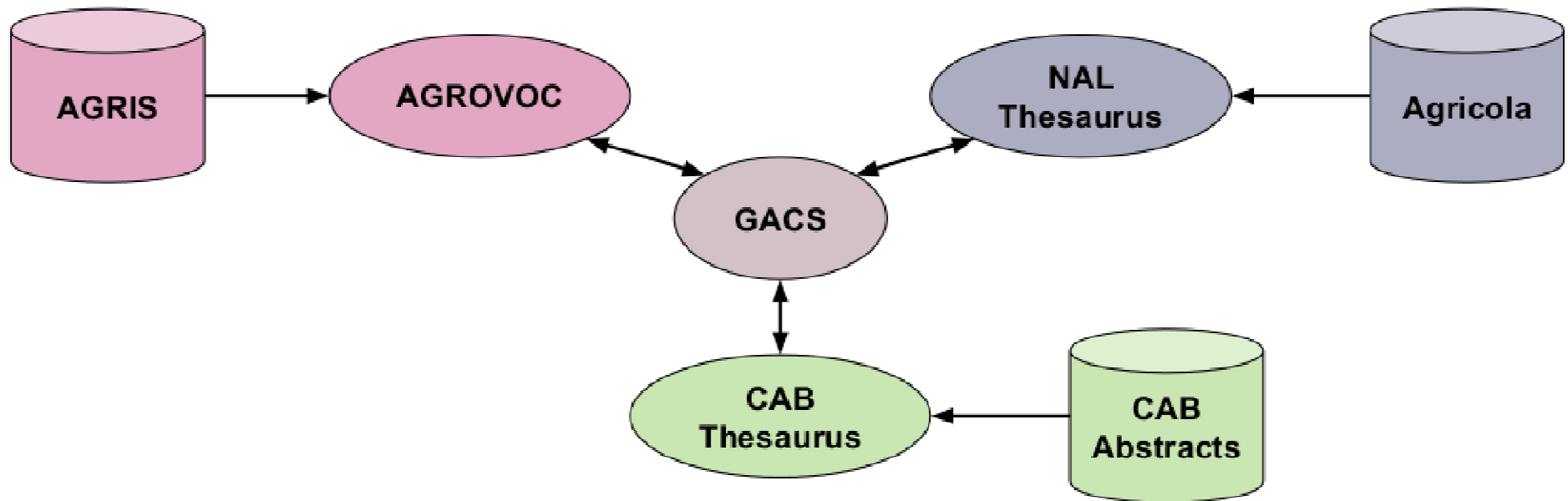
Global Agricultural Concept Scheme (GACS)

1. To improve the semantic interoperability of thesauri maintained by FAO, CABI, and NAL.
2. To provide core concepts broadly supported across the three thesauri.
3. To achieve efficiencies of scale by maintaining the core concepts in cooperation.

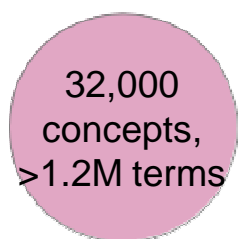
Three Thesauri

Separate thesauri, separate databases

Create GACS as a glue linking them together

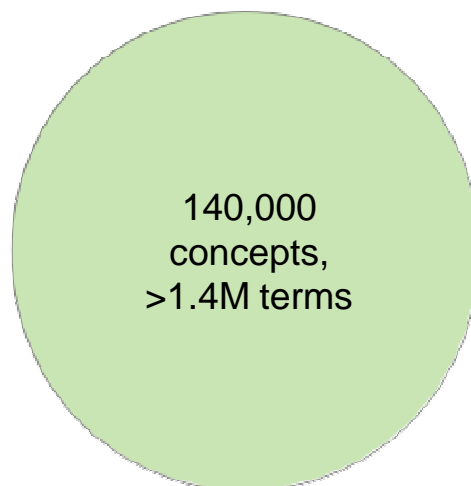


AGROVOC



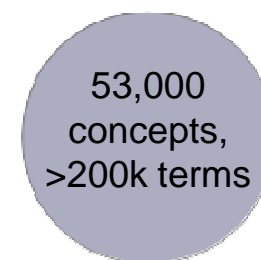
English, Spanish,
Portuguese, German,
Czech, Persian, Polish,
Hindi, French, Italian,
Russian, Japanese,
Hungarian, Chinese,
Slovak, Thai, Lao, Turkish,
Korean, Arabic, Telugu ...

CAB Thesaurus



English, Spanish,
Portuguese, Dutch
+ many languages with
lower coverage

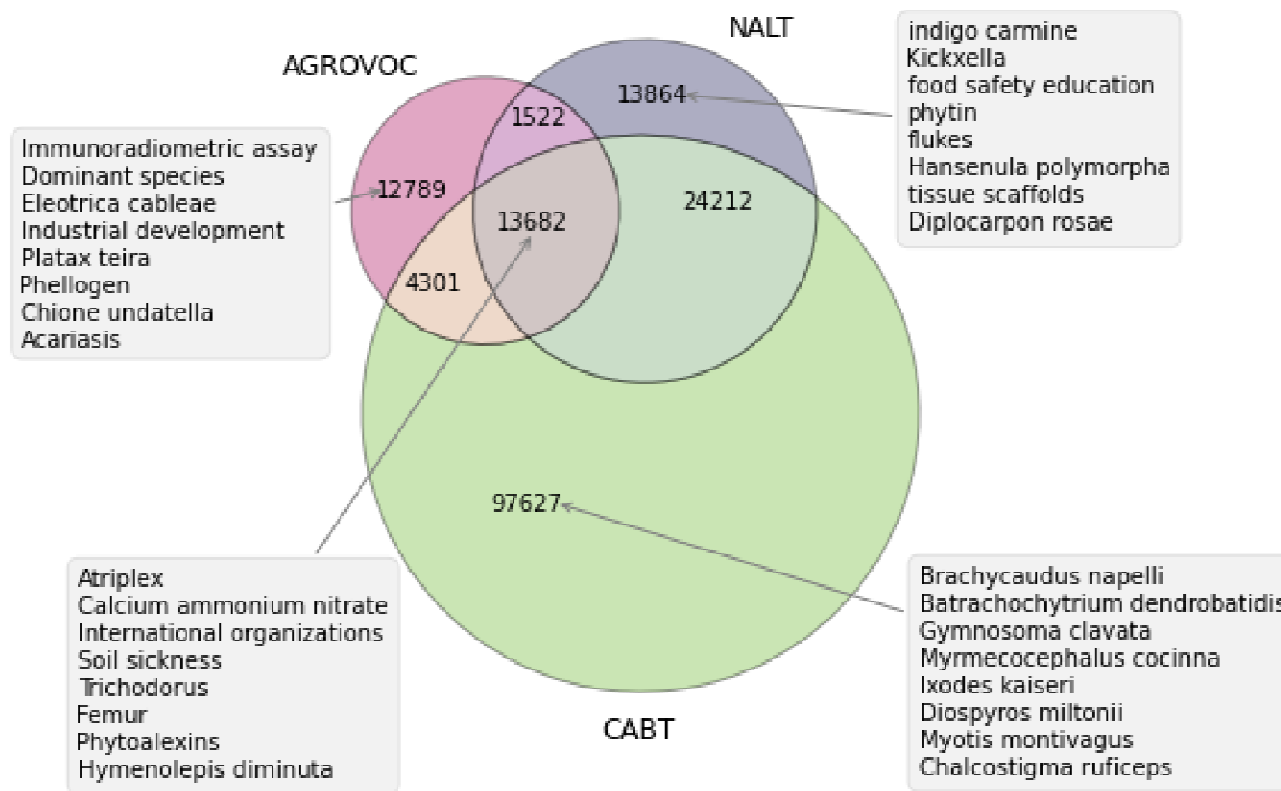
NAL Thesaurus



English, Spanish

All thesauri represented using SKOS

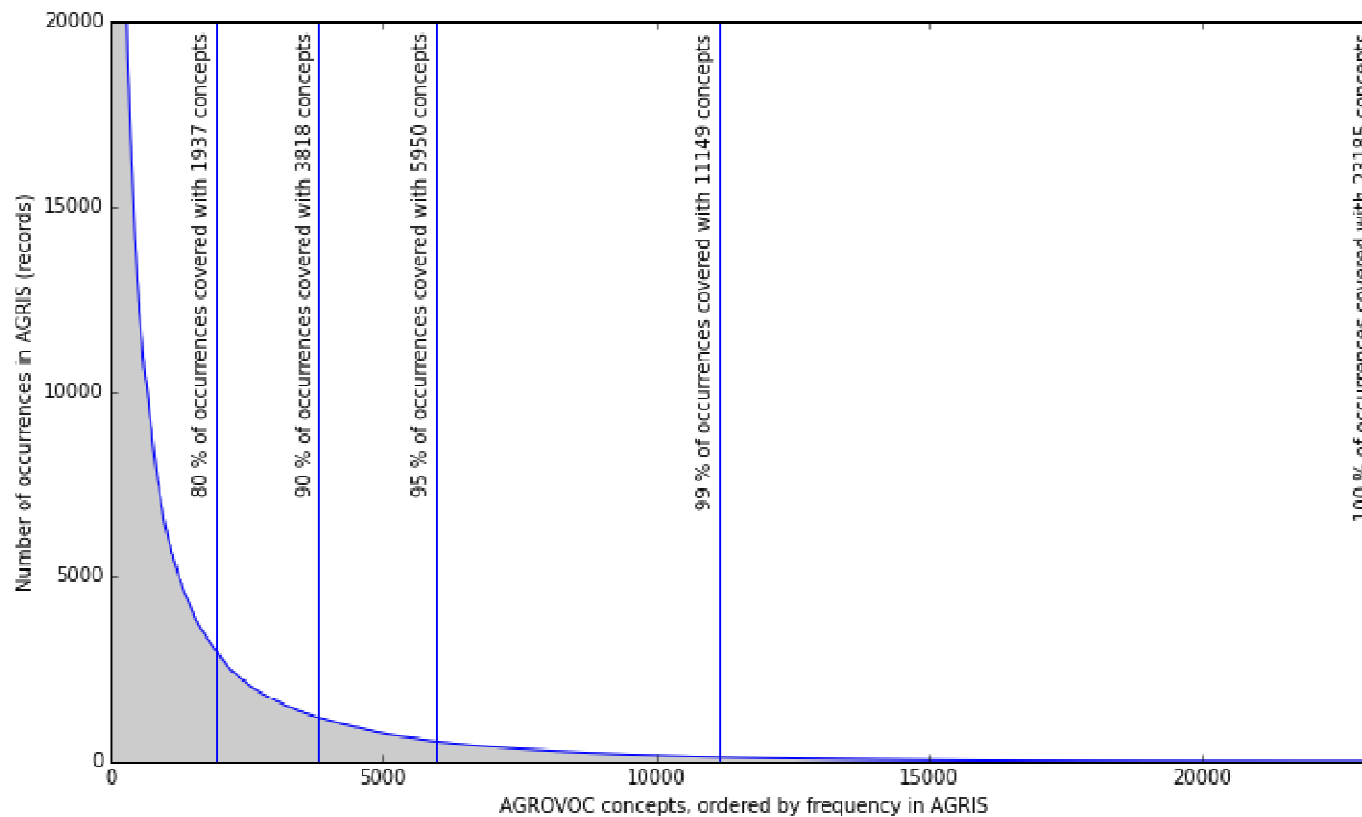
Overlap estimate



Obtained via automatic mappings created using AgreementMakerLight

Long tail distribution (in AGRIS)

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



Creating GACS

Requirements and Wishes

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

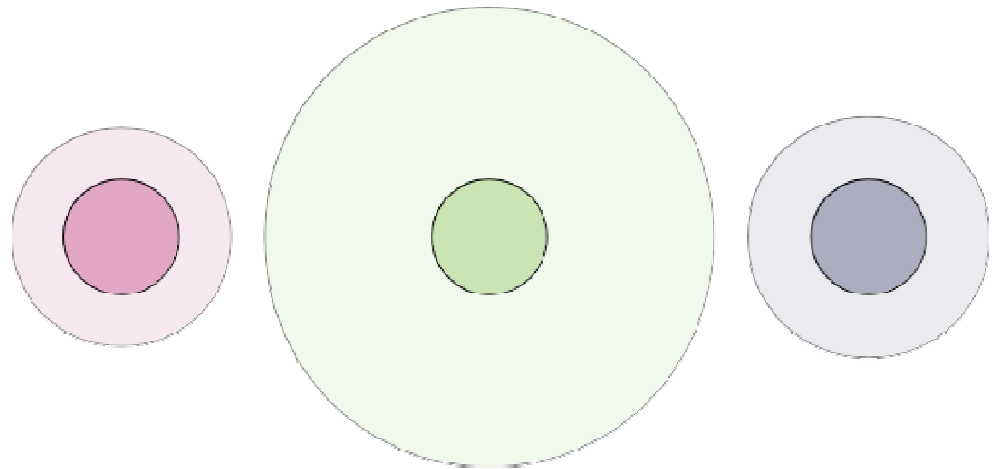
GACS Beta proof-of-concept meets most requirements

Selection of top 10,000 concepts

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

These lists of concepts were modified as follows:

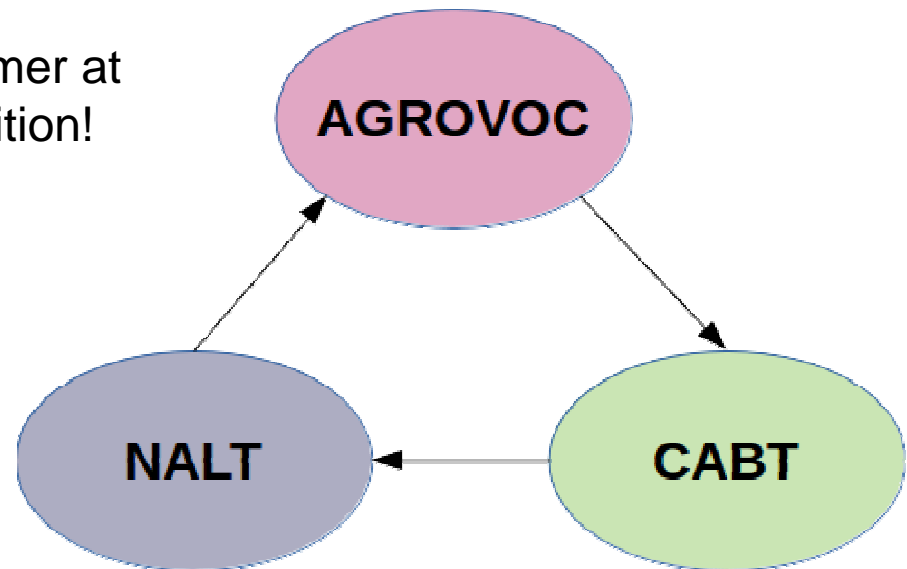
- added all countries (from AGROVOC)
- added organisms hierarchy all the way to the top



Automated mappings

Created using AgreementMakerLight software
between the full thesauri, for completeness

AgreementMakerLight was top performer at
OAEI 2014 ontology mapping competition!

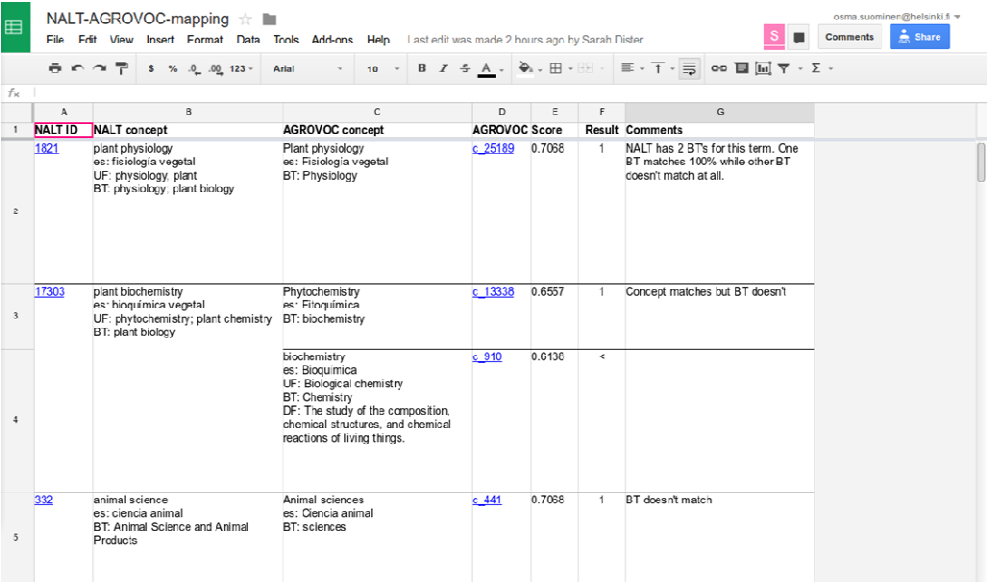


Human evaluation of mappings

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.



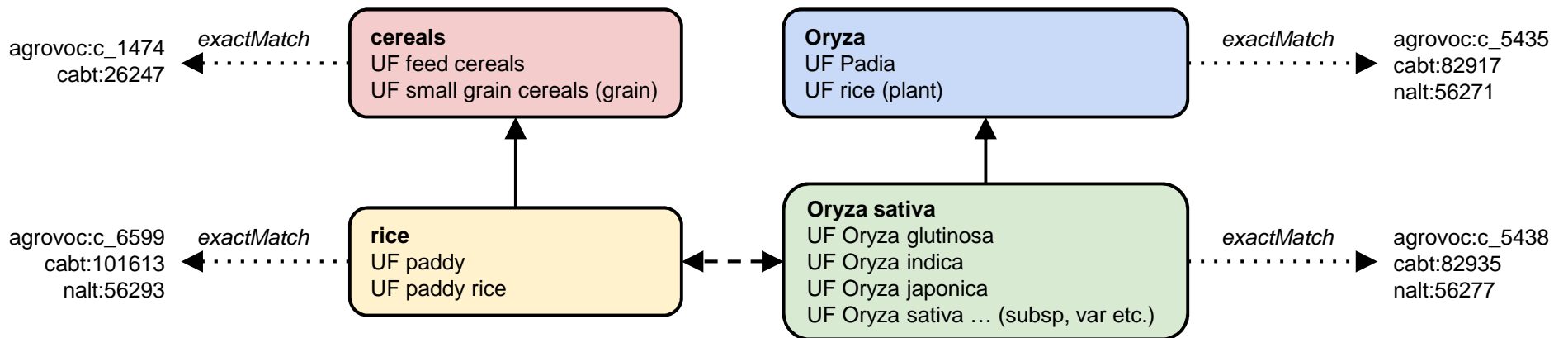
	A	B	C	D	E	F	G
	NALT ID	NALT concept	AGROVOC concept	AGROVOC Score		Result	Comments
1	1821	plant physiology es: fisiología vegetal UF: physiology, plant BT: physiology, plant biology	Plant physiology es: Fisiología vegetal BT: Physiology	c_25189	0.7058	1	NALT has 2 BT's for this term. One BT matches 100% while other BT doesn't match at all.
2							
3	17303	plant biochemistry es: bioquímica vegetal UF: phytochemistry; plant chemistry BT: plant biology	Phytochemistry es: Fitquímica 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	0.6130	<	
5	332	animal science es: ciencia animal BT: Animal Science and Animal Products	Animal sciences es: Ciencia animal BT: sciences	c_441	0.7058	1	BT doesn't match

Slide 14

- 1 The hours used for evaluation should be verified from partners
Osma Suominen,

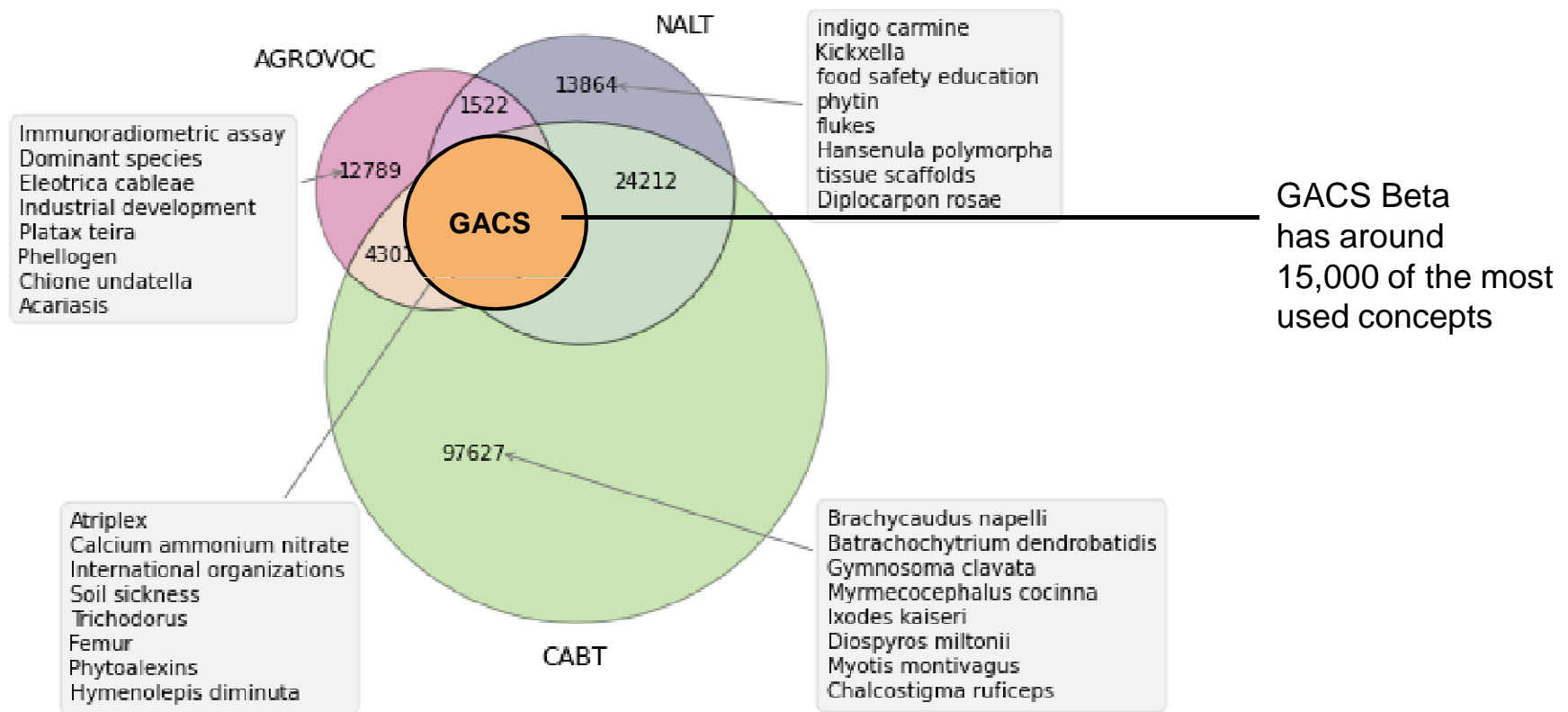
Forming GACS concepts

by merging the source concepts and aggregating their information



(actually we use SKOS, not traditional thesaurus tags)

Size of GACS



Quality evaluation

Using the qSKOS and Skosify tools that can find and correct problems in SKOS vocabularies [1], we can detect

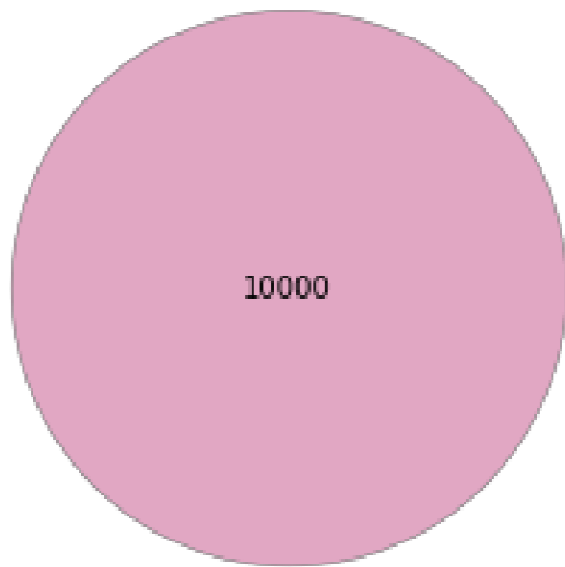
- missing, invalid or overlapping concept labels
- anomalies in concept hierarchy, e.g. cycles
- ...and many other kinds of problems.

Many problems are expected due to merging of concepts within GACS, but most should be automatically corrected.

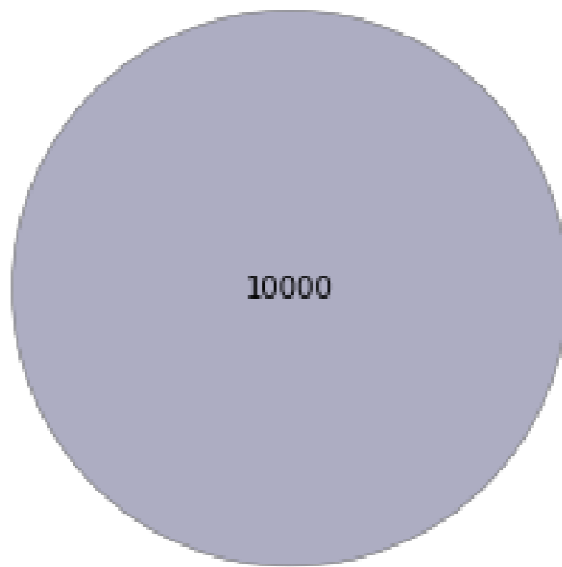
[1] Osmo Suominen and Christian Mader: **Assessing and Improving the Quality of SKOS Vocabularies**. JoDS, 3(1) 2014.

Starting point

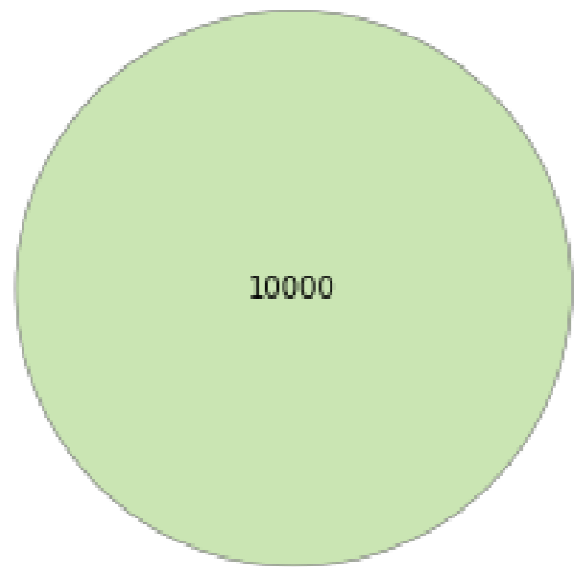
Before mapping



AGROVOC (10000)

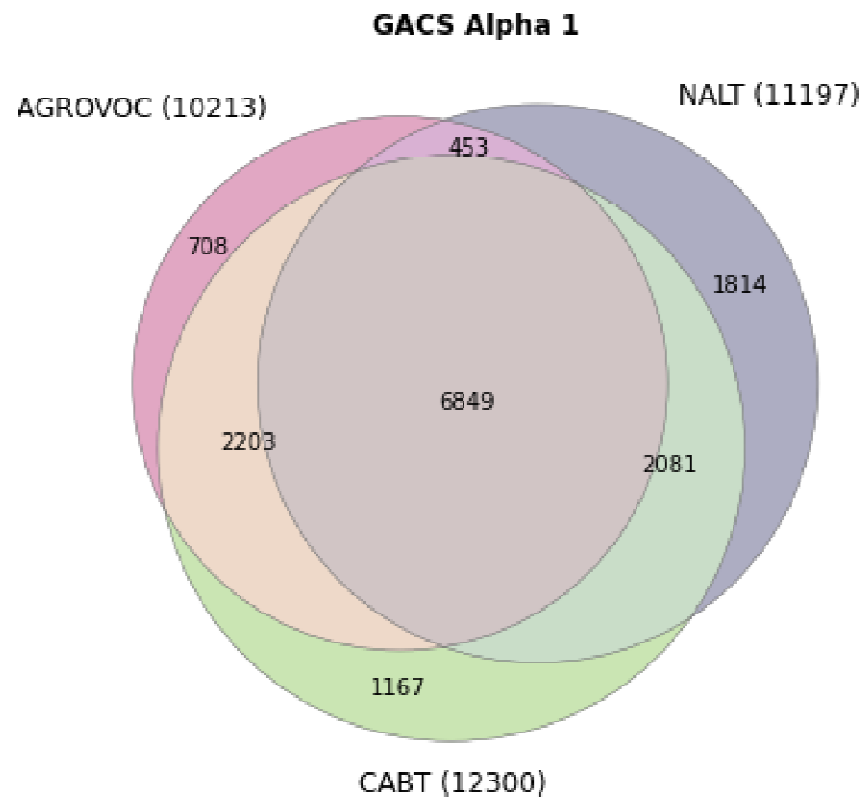


NALT (10000)

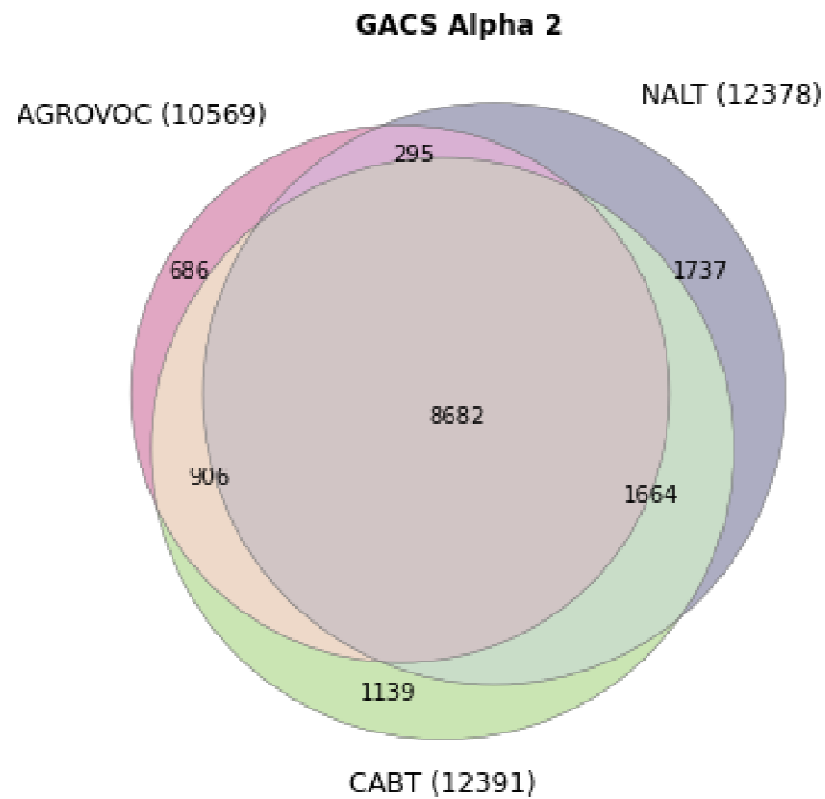


CABT (10000)

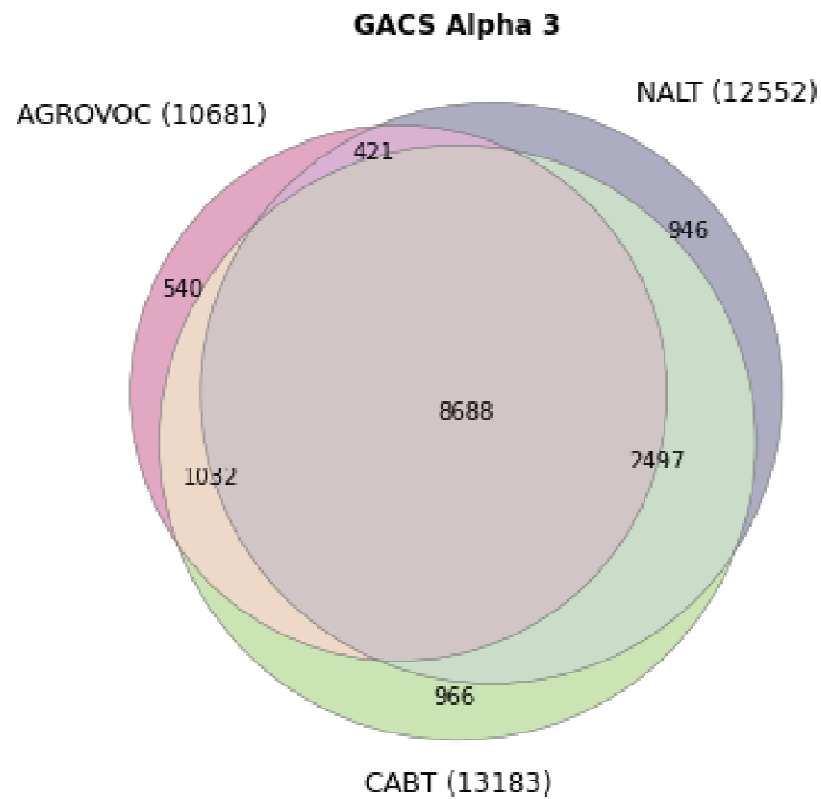
30,000 mappings later...



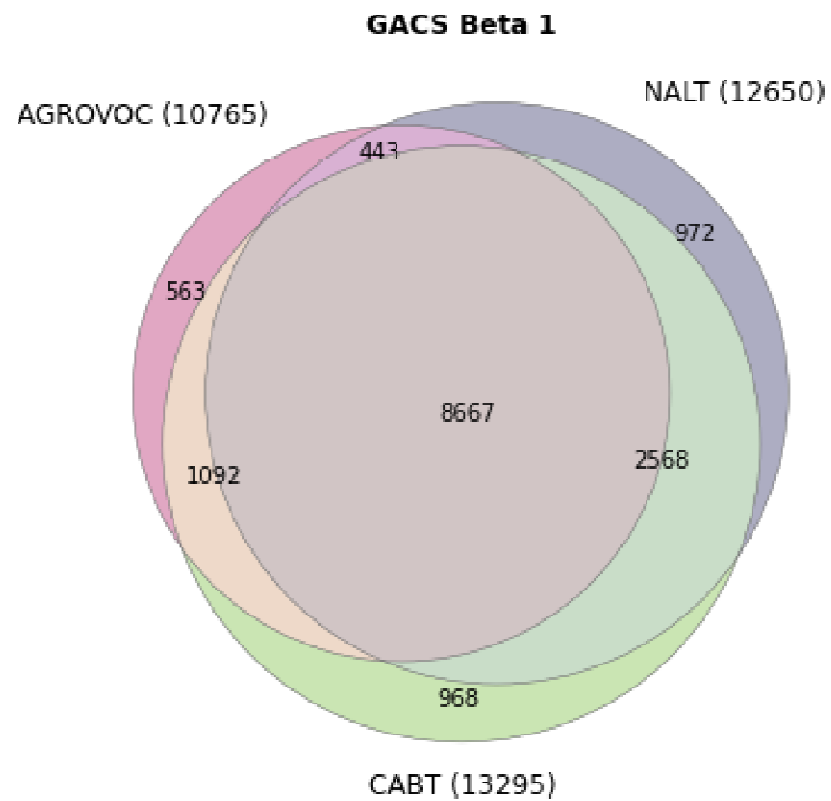
4,689 mappings later...



5,522 mappings later...

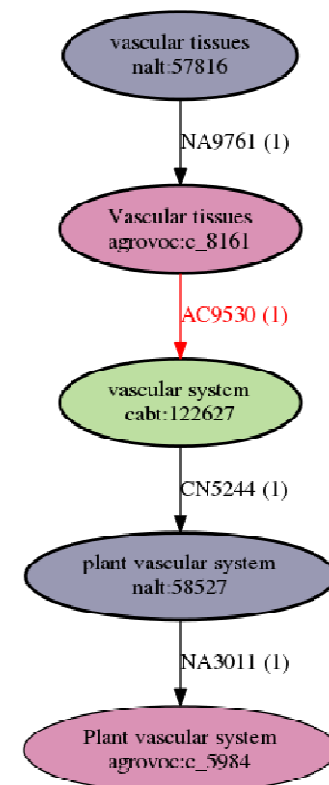
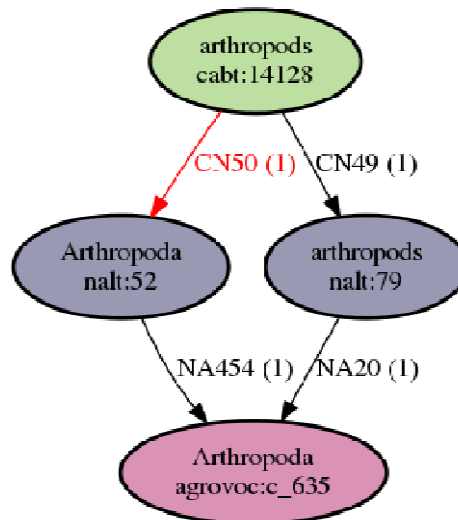
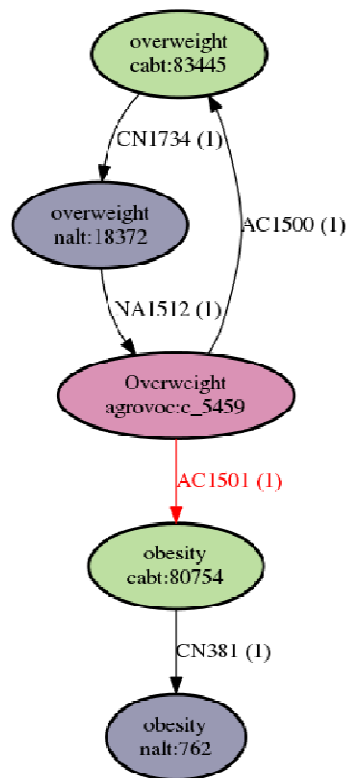


625 resolved lumps later...



Lumps

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



Lessons already learned

- It is hard to sustain focus on mapping beyond circa five hours per day.
- Mapping reveals issues with both the source and target thesauri -- areas for improvement, or errors, fixable in collaboration.
- Starting with the 10,000 most-used concepts shines a light on parts of thesauri that may long have lacked attention.
- Starting small, with a core, avoids the potential stress of over-committing resources.
- Mapping provides an incentive to adopt open-data technologies that can have prove beneficial in other areas.

Differences in modeling

Q: Are taxonomic organism names (e.g. '*Bos taurus*') different concepts than the common names ('*cattle*')?

- sometimes there is no 1:1 match and/or context of use is different
- the source thesauri all have different policies

VocBench for editing

Signed in as anonymous (Publisher) to: Agrovoc Administration | About VocBench | English | RSS feed | Preferences | Help | Sign out

VocBench VERSION 2.1 [Build 20140422] (SANDBOX)

Exact word: Go [Advanced search](#) | [Last results](#)

[Recent changes](#) | **Concepts** | [Properties](#) | [Schemes](#) | [Validation](#) | [Load data](#) | [Export](#) | [Statistics](#) | [SPARQL](#) Concept navigation history Content language

Concepts Show LIRI ☒ Show non-preferred ☐

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

Paddy (en); Rice (en) Show inferred and explicit ☐

Terms (2) | **Definition (0)** | **Note (0)** | **Attributes (0)** | **Notation (0)** | **Relationship (0)** | **History (0)** | **Image (0)** | **Scheme (1)** | **Hierarchy**

Add new term

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

Legend: Proposed Validated Published Revised Proposed deprecated Deprecated Show more

© FAO & ART Group, 2014

Skosmos for display and browsing

GACS Alpha

Search within this vocabulary

Any language

Search

Alphabetical

Hierarchy

Plant protein

Pula

Sago

Spices

Stimulants

Sugar

Sugarbeet

Sugarcane

Tapioca

Tea

Turf

Vegetable products

cocoa products

fruit

grain products

Sida

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

Arabic

Chinese

稻米

水稻

Czech

ryže

ryže setá

Denish

ris

Dutch

rijst

Finnish

riisi

French

riz

Next steps and future of GACS

GACS Phase 3

- Publish GACS Beta 3 by end-2015
- Concept scheme with own semantic structure
 - Own publication and editorial platform
 - Based on, mapped to, but independent of, its three source thesauri
- Quality improvements
 - Inconsistencies in hierarchy, choice of labels, scope notes and definitions
- Enriching GACS structure
 - Common vs scientific names, custom relationships, concept types (as in UMLS Semantic Network), thematic groups

Beyond GACS Beta?

Q: Can GACS replace existing agricultural thesauri?

- definitely not with GACS Beta due to smaller scope/size
- a future GACS may be an alternative for some scenarios, but not all uses of existing thesauri because
 - they cover areas beyond agriculture
 - existing systems and processes (publication, automatic indexing...) depend on current thesauri

Extend to more partners?

Thank you

Reports available on the FAO AIMS site:

<http://aims.fao.org/community/agrovoc/blogs/phase-one-gacs-approved-read-reports>

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