

# Turning three overlapping thesauri into a Global Agricultural Concept Scheme

SWIB14, Bonn, 3 December 2014

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*.



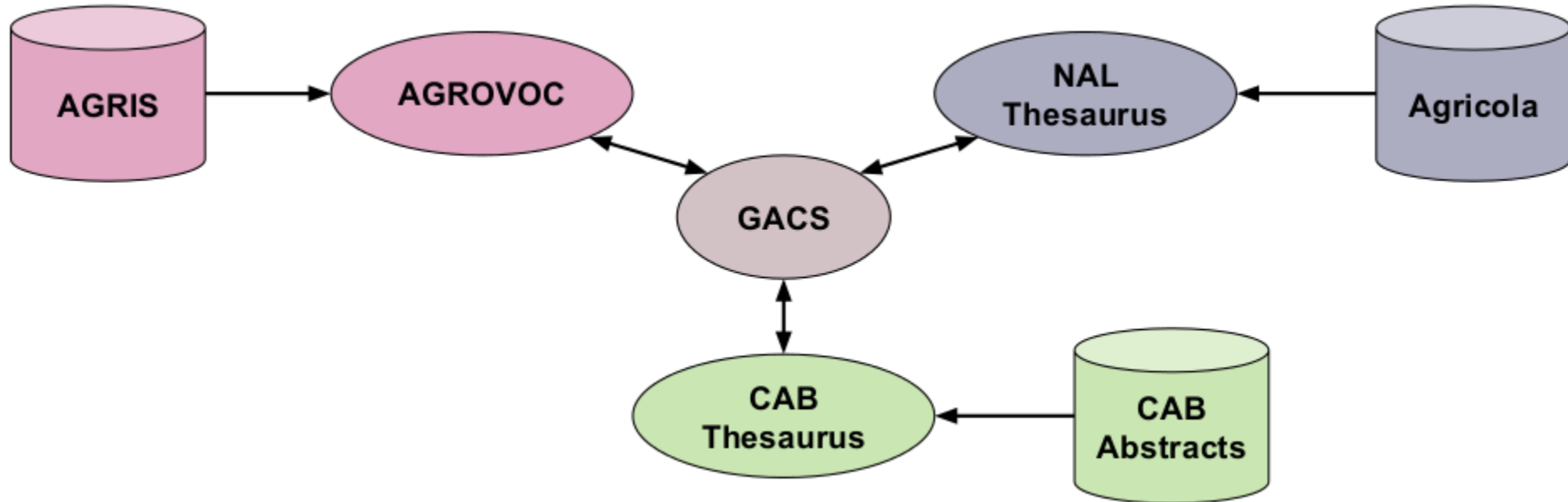
# **Global Agricultural Concept Scheme (GACS)**

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.

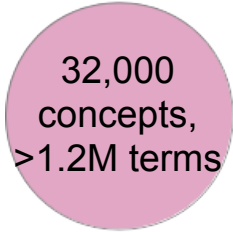
# Three Thesauri

# Separate thesauri, separate databases

Create GACS as a glue linking them together

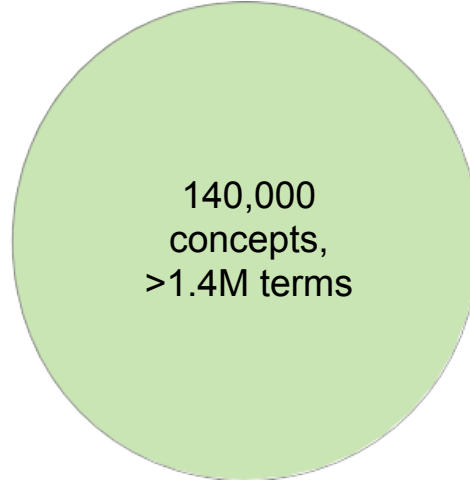


## AGROVOC



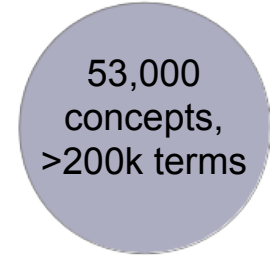
English, Spanish,  
Portuguese, German,  
Czech, Persian, Polish,  
Hindi, French, Italian,  
Russian, Japanese,  
Hungarian, Chinese,  
Slovak, Thai, Lao, Turkish,  
Korean, Arabic, Telugu ...

## CAB Thesaurus



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

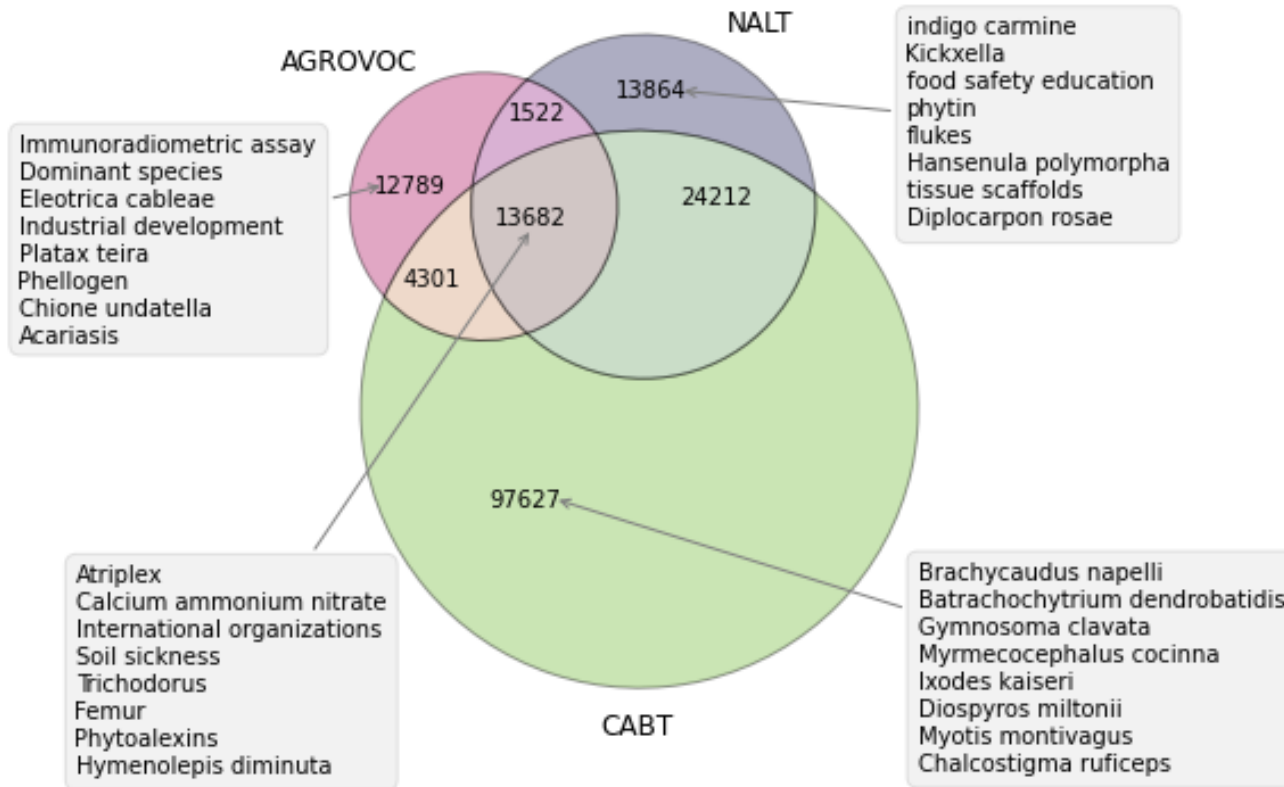
## NAL Thesaurus



English, Spanish

**All thesauri represented using SKOS**

# Overlap estimate

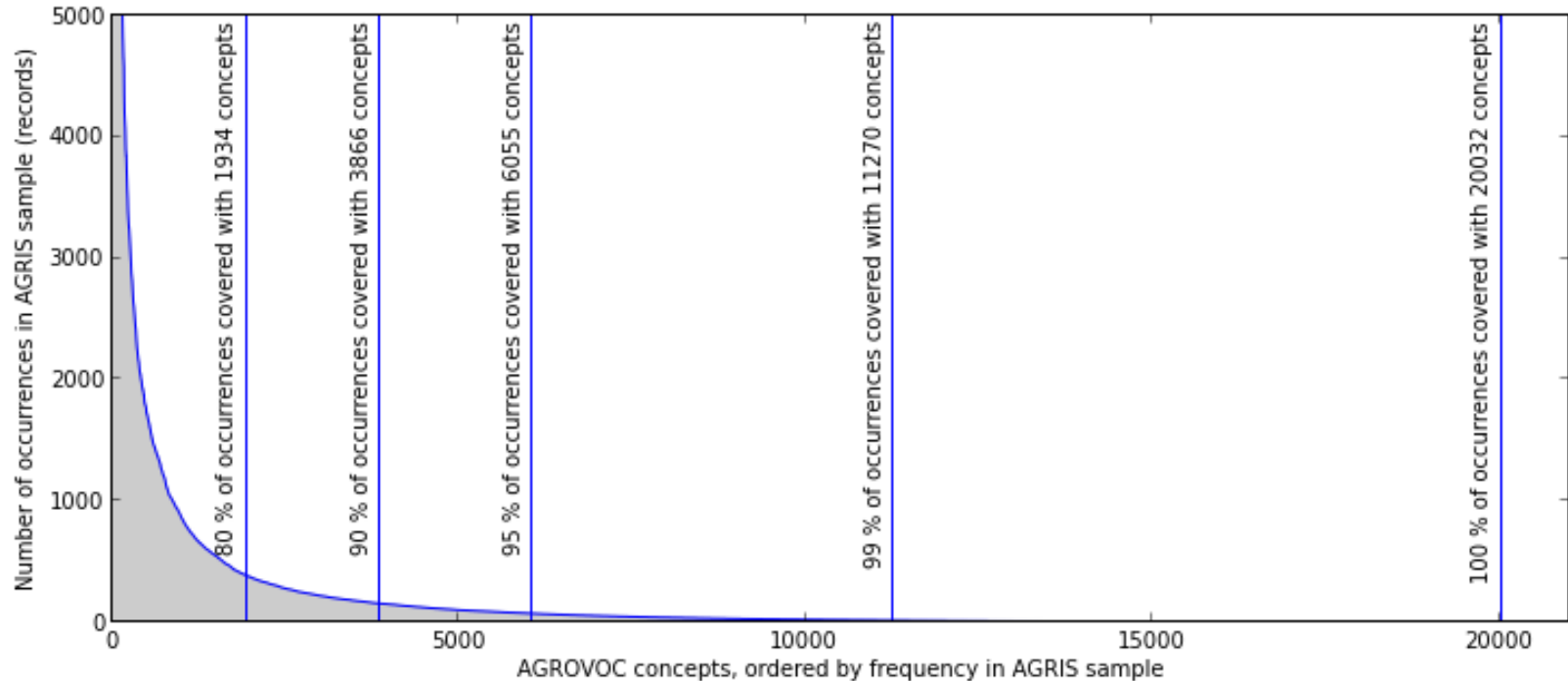


Obtained via automatic mappings created using AgreementMakerLight



# Long tail distribution (in AGRIS)

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



# Creating GACS

# Requirements and Wishes

1. An integrated view and bridge of existing thesauri
2. Reuses thesaurus development work, incl. translations
3. Compatible with existing databases
4. Based on RDF technologies: URIs, SKOS etc.
5. Available as Linked Open Data

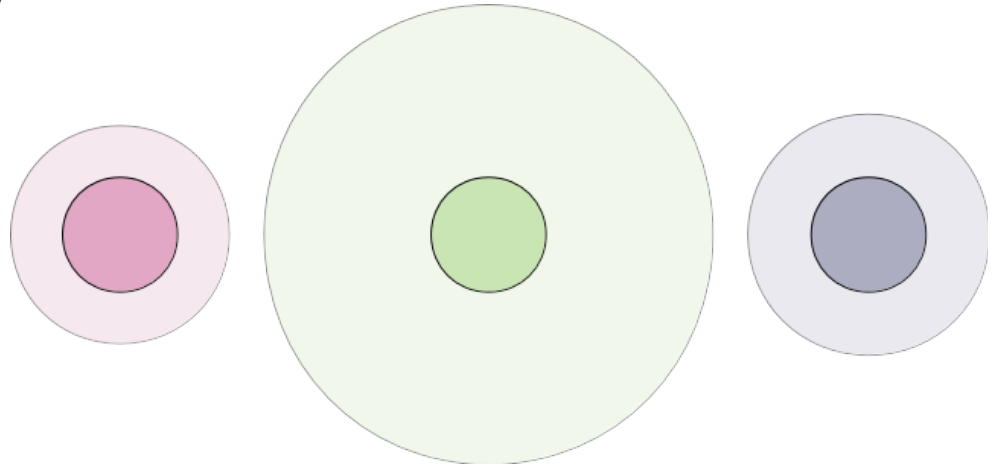
Currently building **GACS Beta**, a proof-of-concept implementation attempting to fulfill most requirements

# 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:

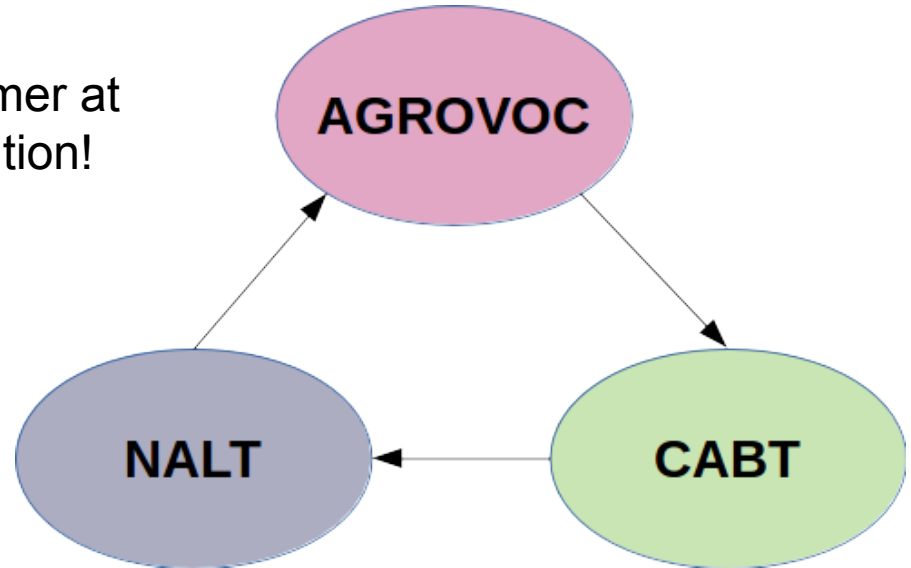
- added all countries (from AGROVOC)
- added organisms hierarchy all the way to the top



# Automated mappings

Created using AgreementMakerLight software between the full thesauri, for completeness

AgreementMakerLight was top performer at OAEI 2014 ontology mapping competition!



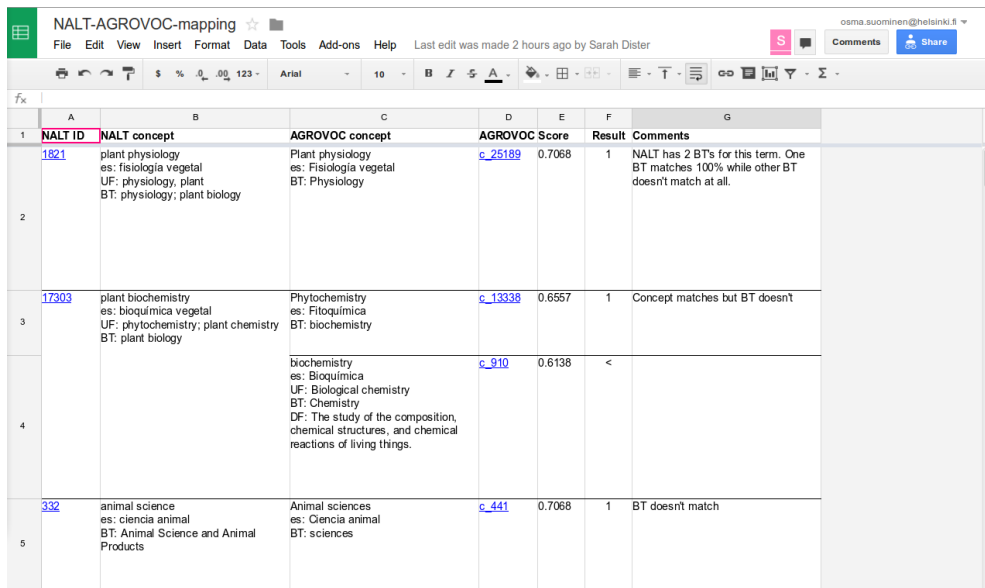
# Human evaluation of mappings

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, total evaluation time over 300 hours so far.

Currently projected to take 500-600 hours for GACS Beta.

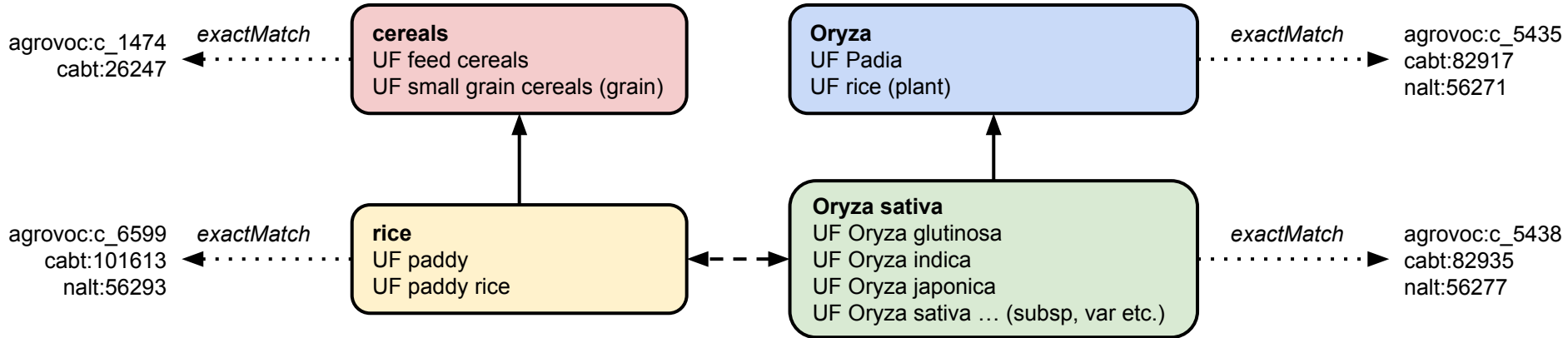


The screenshot shows a Google Docs spreadsheet titled "NALT-AGROVOC-mapping". The spreadsheet contains a table with the following columns: NALT ID, NALT concept, AGROVOC concept, AGROVOC Score, Result, and Comments. The table lists several rows of mappings, including:

	A	B	C	D	E	F	G
1	<a href="#">1821</a>	plant physiology es: fisiologia vegetal UF: fisiologia, plant BT: fisiologia; plant biology	Plant physiology es: Fisiologia vegetal BT: Fisiologia	<a href="#">c_25189</a>	0.7068	1	NALT has 2 BTs for this term. One BT matches 100% while other BT doesn't match at all.
3	<a href="#">17303</a>	plant biochemistry es: bioquímica vegetal UF: phytochemistry; plant chemistry BT: plant biology	Phytochemistry es: Fitoquímica BT: biochemistry	<a href="#">c_13338</a>	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.	<a href="#">c_910</a>	0.6138	<	
5	<a href="#">332</a>	animal science es: ciencia animal BT: Animal Science and Animal Products	Animal sciences es: Ciencia animal BT: sciences	<a href="#">c_441</a>	0.7068	1	BT doesn't match

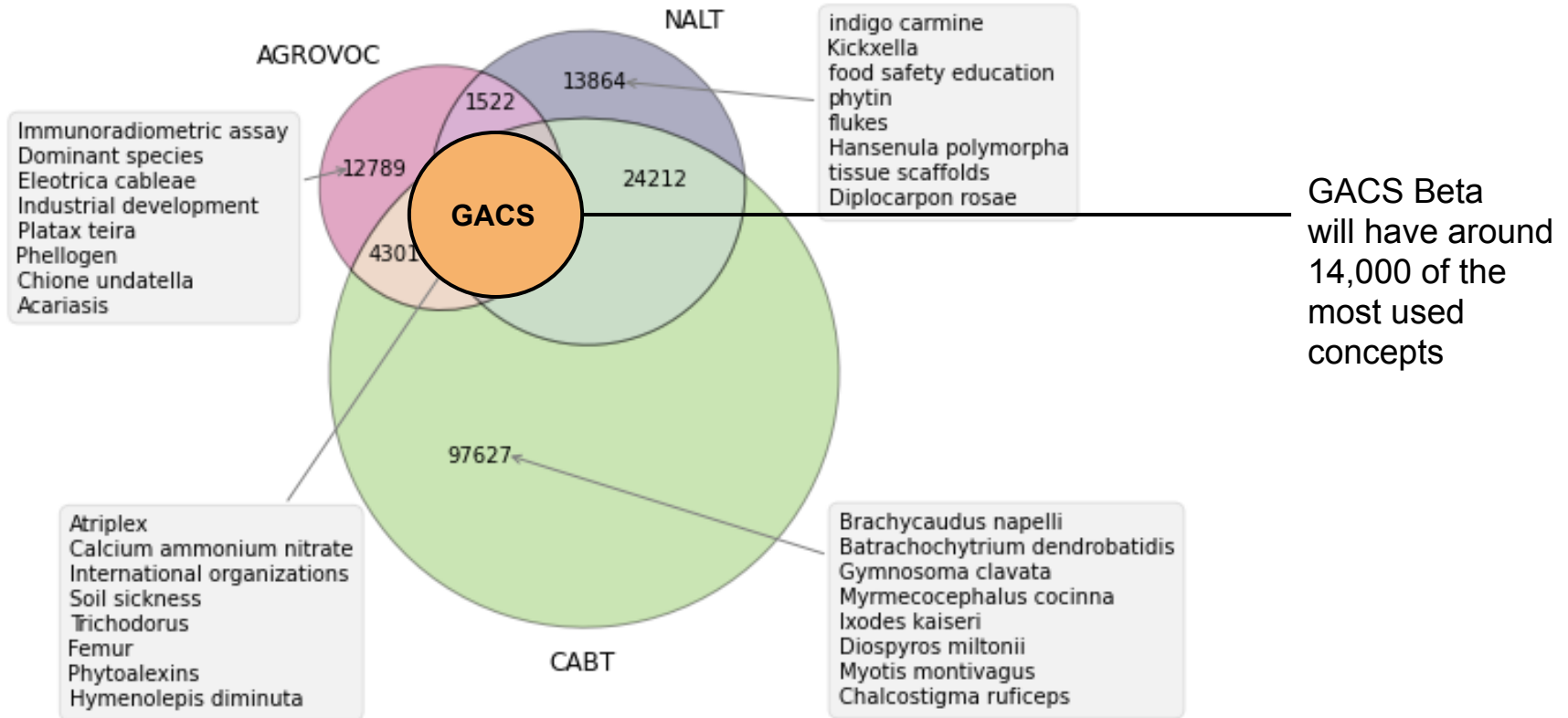
# Forming GACS concepts

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

# Demo of GACS Alpha in Skosmos

GACS Alpha Search within this vocabulary Any language Search

[Alphabetical](#) [Hierarchy](#)

- 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
    - Millet
    - 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	ryže ryže setá
Danish	ris
Dutch	rijst
Finnish	riisi
French	riz

# 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.

# Challenges

# Differences in modeling

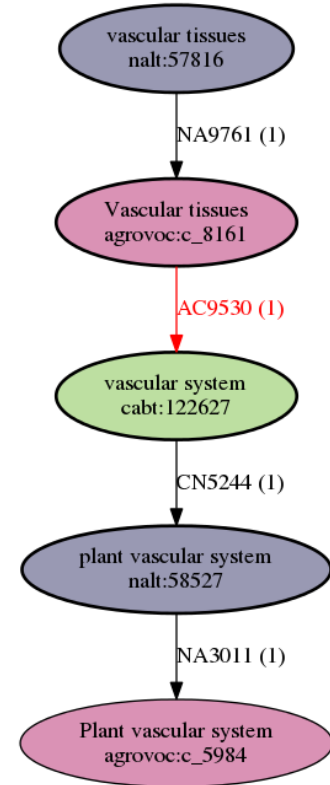
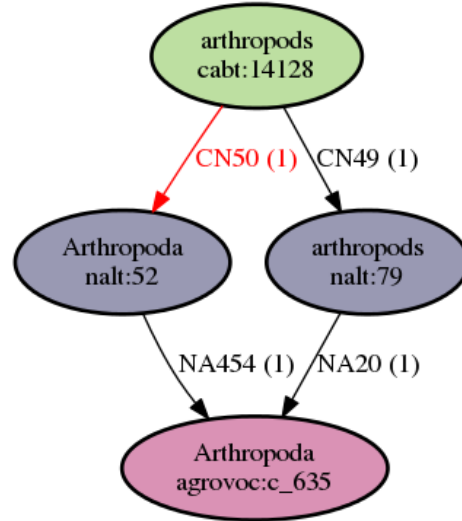
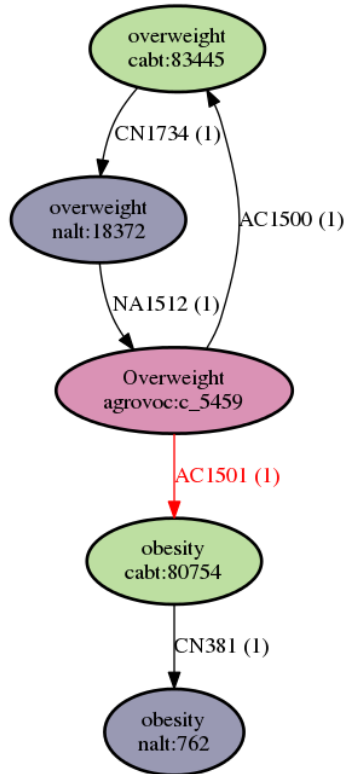
Q: Are taxonomic organism names (e.g. '*Bos taurus*') different concepts than the common names ('*cattle*')?

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

No final answer yet...

# Lumps

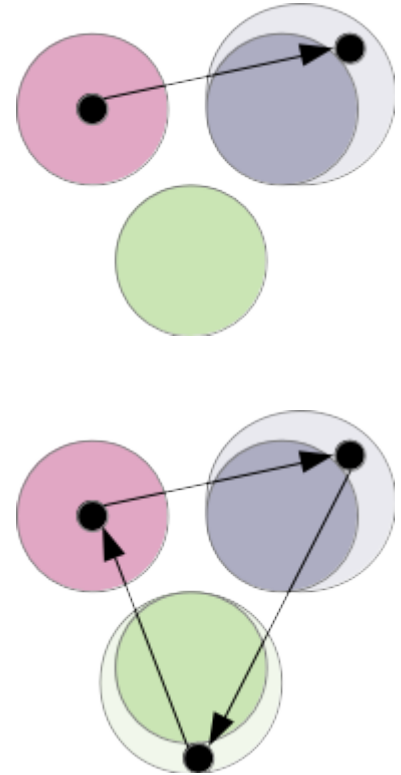
clusters of concepts mapped one-to-several, several-to-one, or in spirals



# **Next steps and future of GACS**

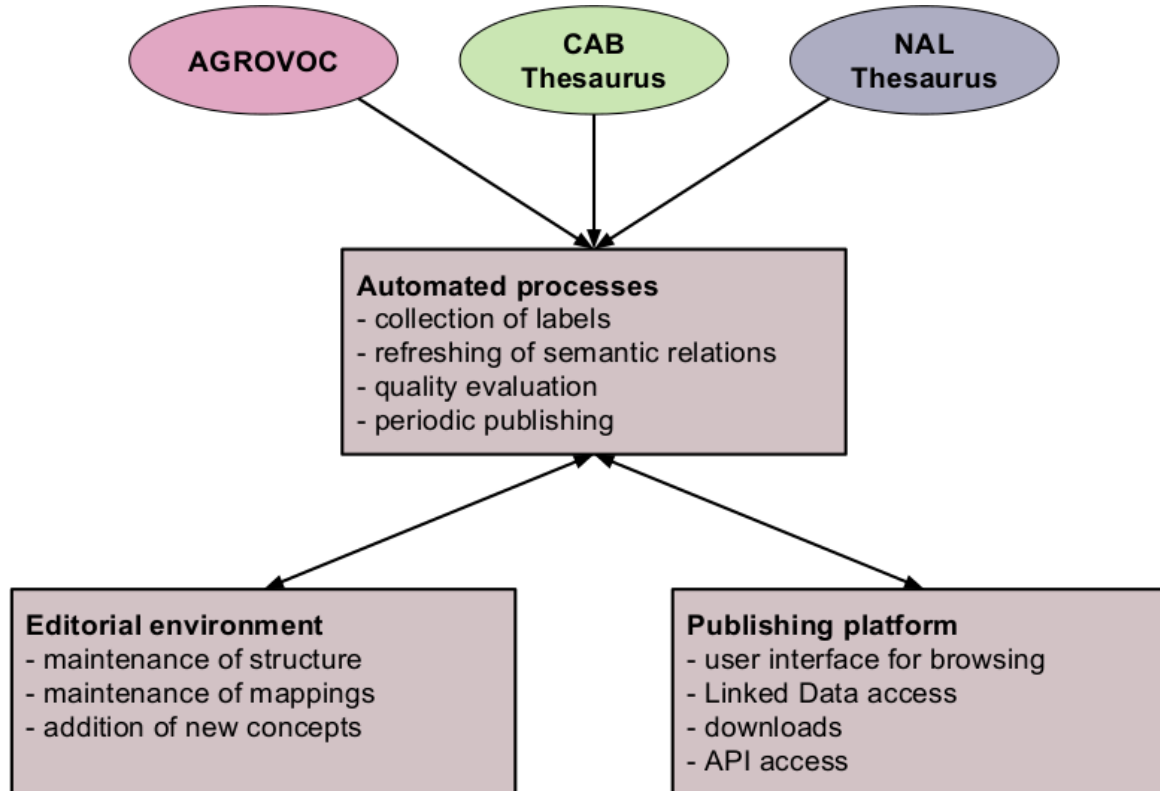
# Additional mapping rounds

Need to perform 2-3 more smaller mapping rounds in order to ensure that all necessary concepts have been fully mapped between all source thesauri





# GACS system infrastructure



# 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   [Advanced search](#) | [Last results](#)

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

**Concepts**  Show URI  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) Attribute (0) Notation (0) Relationship (0) History (0) Image (0) Scheme (1) Hierarchy

+ Add new term

Language	Term
English (en)	<input checked="" type="checkbox"/> Rice (Preferred) <input type="checkbox"/> W
	<input checked="" type="checkbox"/> Paddy <input type="checkbox"/> W

Legend Proposed Validated Published Revised Proposed deprecated Deprecated [Show more](#)

# 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

In future, more partners are expected and the scope of GACS can be adjusted.

# Thank you

Reports available on the FAO AIMS site:

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

These slides: <http://tinyurl.com/swib14-gacs>

[osma.suominen@helsinki.fi](mailto:osma.suominen@helsinki.fi)

[tom@tombaker.org](mailto:tom@tombaker.org)