GACS Survey results, part 2: Structural questions

Osma Suominen
GACS workshop @MTSR
Göttingen, Germany
November 23, 2016

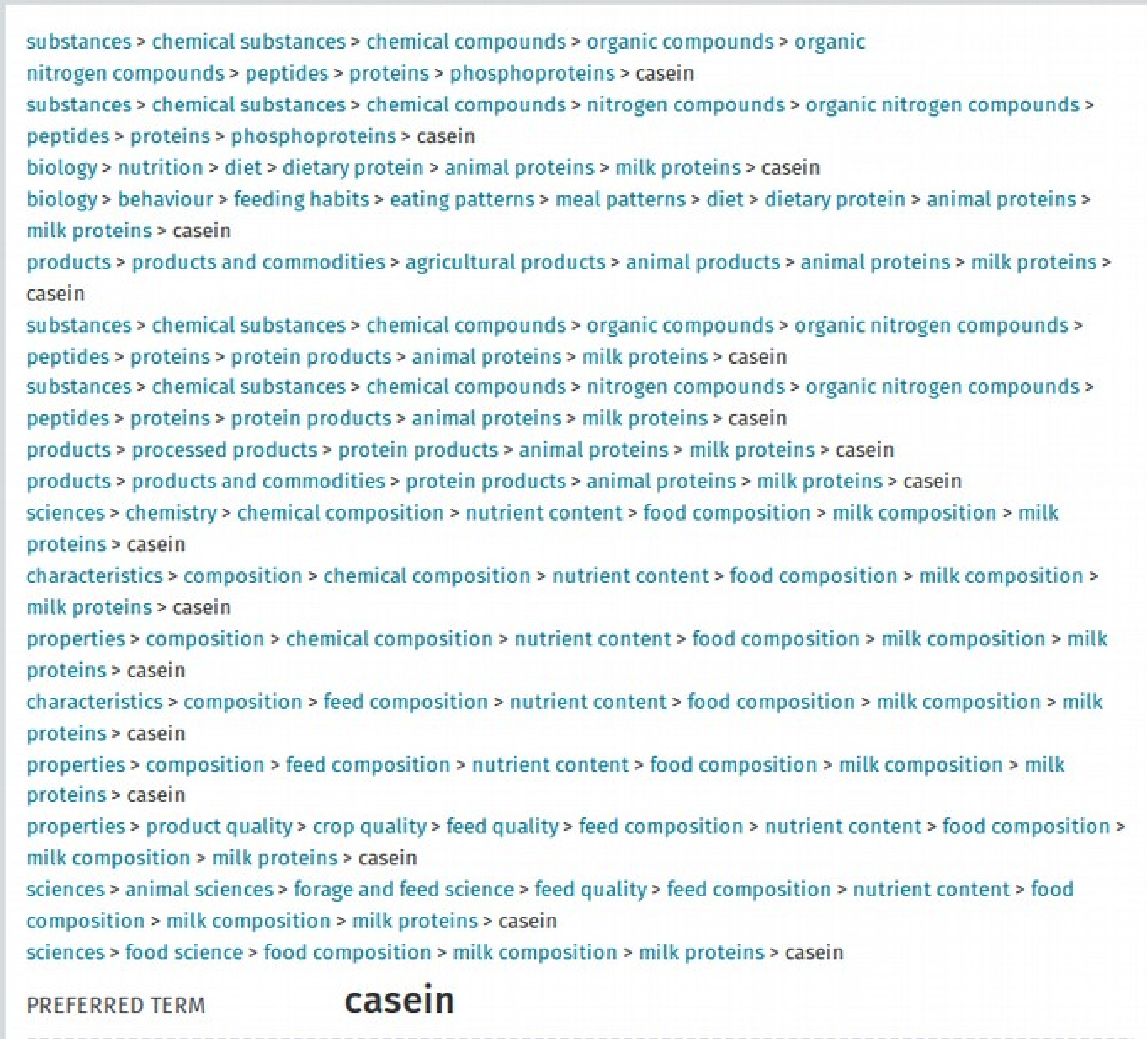
About the survey

- We wanted to hear thoughts about how GACS should be **structured** so that it best serves the **needs of users** as well as **applications** that make use of GACS.
- Blog post on AIMS with background & scenarios
- Actual survey using Google Forms
  - open November 3 to November 20 (21)
- Advertised on many forums including: MTSR workshop participants, RDA Agrisemantics WG, NKOS community
- Commissioned with funding from CABI

GACS Core Beta 3.1 hierarchy stats

- 15427 concepts
- 20072 BT/NT relationships (1.3 per concept)
- 4226 concepts (27%) have more than one BT
- 578 top concepts

Hierarchy worst case: *casein* has 17 paths from the top



substances > chemical substances > chemical compounds > organic compounds > organic nitrogen compounds > peptides > proteins > phosphoproteins > casein

substances > chemical substances > chemical compounds > nitrogen compounds > organic nitrogen compounds > peptides > proteins > phosphoproteins > casein

biology > nutrition > diet > dietary protein > animal proteins > milk proteins > casein

biology > behaviour > feeding habits > eating patterns > meal patterns > diet > dietary protein > animal proteins > milk proteins > casein

products > products and commodities > agricultural products > animal products > animal proteins > milk proteins > casein

substances > chemical substances > chemical compounds > organic compounds > organic nitrogen compounds > peptides > proteins > protein products > animal proteins > milk proteins > casein

substances > chemical substances > chemical compounds > nitrogen compounds > organic nitrogen compounds > peptides > proteins > protein products > animal proteins > milk proteins > casein

products > processed products > protein products > animal proteins > milk proteins > casein

products > products and commodities > protein products > animal proteins > milk proteins > casein

sciences > chemistry > chemical composition > nutrient content > food composition > milk composition > milk proteins > casein

characteristics > composition > chemical composition > nutrient content > food composition > milk composition > milk proteins > casein

properties > composition > chemical composition > nutrient content > food composition > milk composition > milk proteins > casein

characteristics > composition > feed composition > nutrient content > food composition > milk composition > milk proteins > casein

properties > composition > feed composition > nutrient content > food composition > milk composition > milk proteins > casein

properties > product quality > crop quality > feed quality > feed composition > nutrient content > food composition > milk composition > milk proteins > casein

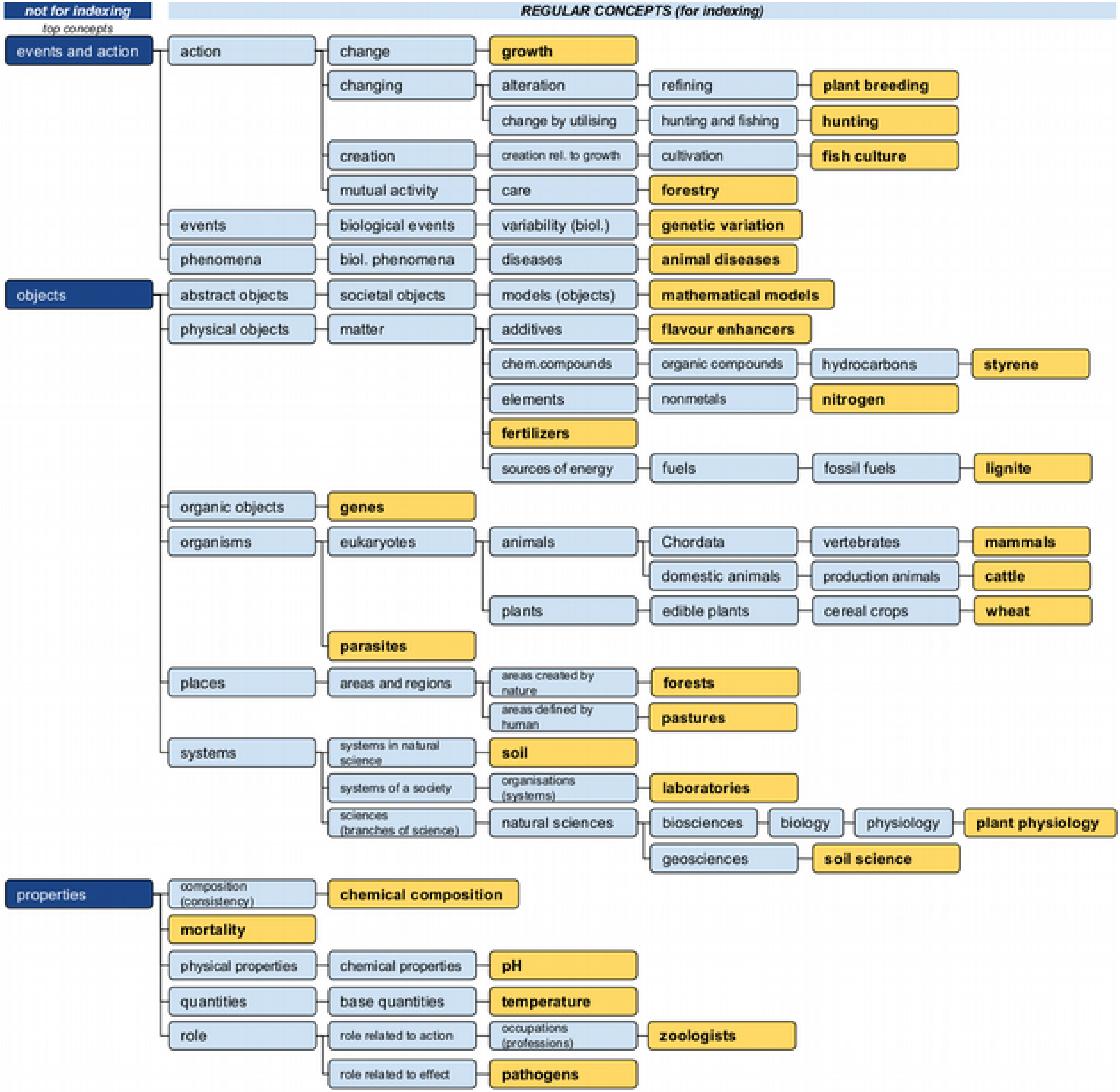
sciences > animal sciences > forage and feed science > feed quality > feed composition > nutrient content > food composition > milk composition > milk proteins > casein

sciences > food science > food composition > milk composition > milk proteins > casein

PREFERRED TERM casein

Scenario A

based on YSO
and DOLCE



not for indexing
top concepts

REGULAR CONCEPTS (for indexing)

- events and action
  - action
    - change
      - **growth**
    - changing
      - alteration
        - refining
          - **plant breeding**
      - change by utilising
        - hunting and fishing
          - **hunting**
    - creation
      - creation rel. to growth
        - cultivation
          - **fish culture**
    - mutual activity
      - care
        - **forestry**
  - events
    - biological events
      - variability (biol.)
        - **genetic variation**
  - phenomena
    - biol. phenomena
      - diseases
        - **animal diseases**
- objects
  - abstract objects
    - societal objects
      - models (objects)
        - **mathematical models**
  - physical objects
    - matter
      - additives
        - **flavour enhancers**
      - chem.compounds
        - organic compounds
          - hydrocarbons
            - **styrene**
      - elements
        - nonmetals
          - **nitrogen**
      - **fertilizers**
      - sources of energy
        - fuels
          - fossil fuels
            - **lignite**
  - organic objects
    - **genes**
  - organisms
    - eukaryotes
      - animals
        - Chordata
          - vertebrates
            - **mammals**
        - domestic animals
          - production animals
            - **cattle**
      - plants
        - edible plants
          - cereal crops
            - **wheat**
    - **parasites**
  - places
    - areas and regions
      - areas created by nature
        - **forests**
      - areas defined by human
        - **pastures**
  - systems
    - systems in natural science
      - **soil**
    - systems of a society
      - organisations (systems)
        - **laboratories**
    - sciences (branches of science)
      - natural sciences
        - biosciences
          - biology
            - physiology
              - **plant physiology**
        - geosciences
          - **soil science**
- properties
  - composition (consistency)
    - **chemical composition**
  - **mortality**
  - physical properties
    - chemical properties
      - **pH**
  - quantities
    - base quantities
      - **temperature**
  - role
    - role related to action
      - occupations (professions)
        - **zoologists**
    - role related to effect
      - **pathogens**

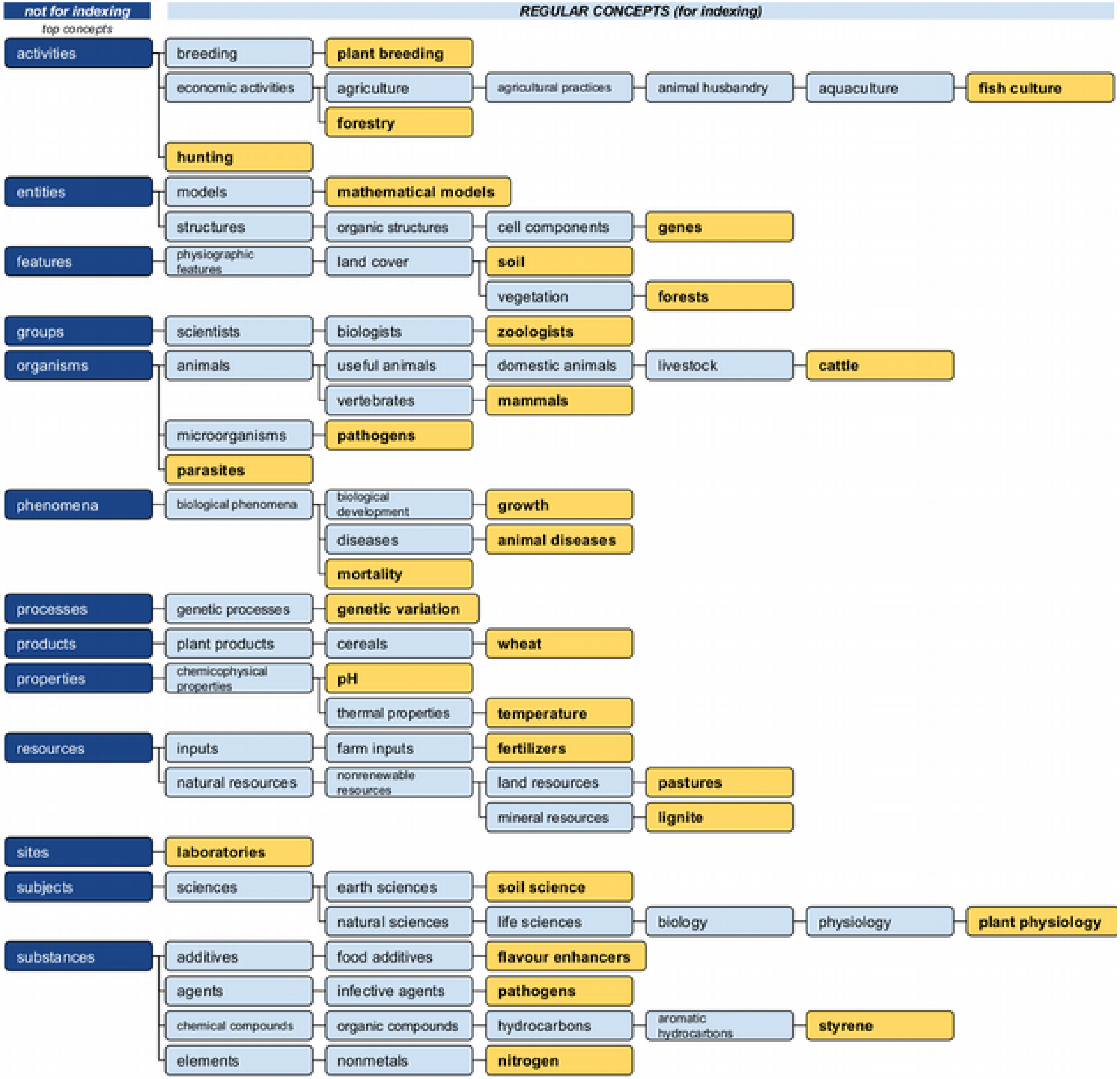
Strengths and Weaknesses of Scenario A

- clear meaning of concepts (3)
- good for reuse and alignment (3)
- based on DOLCE, not ad hoc (2)
- not too many top concepts (2)
- stable and thus easy to maintain (2)
- top concepts provide hints about the nature of concepts
- simple model
- flexible, everything has its place
- IS-A relationships gives strong semantics
- easy to apply
- disjoint top-level classes
- easy to distinguish indexing concepts from non-indexing concepts
- well organized
- easy to grasp
- represents properties as such
- conceptual space enables software to use it without awkward workarounds
- corresponds to high levels of other industry vocabularies

- too many layers of artificial, abstract concepts (5)
- too abstract (3)
- too focused at top level (3)
- difficult to browse for normal people (3)
- difficult for maintainers
- difficult for indexers
- top level is not useful
- too deep and complex when including all concepts
- risk losing oversight of top level concepts when adding concepts deep into the hierarchy
- not intuitive to have forestry and forests in completely different branches of the hierarchy
- hierarchy not specific enough for user needs
- ambiguity problems
- super classes not reasonable, e.g. Zoologists under properties?
- breadth of coverage is not apparent
- no practical use for the general IS-A types
- linguistic based categories cannot be used to drive semantics for information retrieval
- conjunctions create problems, e.g. ”plant fats and oils”
- tries to adopt an ontology-like approach but using the modelling tools of thesauri

Scenario B

based on
AGROVOC



not for indexing
top concepts

REGULAR CONCEPTS (for indexing)

- activities
  - breeding
    - **plant breeding**
  - economic activities
    - agriculture
      - agricultural practices
        - animal husbandry
          - aquaculture
            - **fish culture**
    - **forestry**
  - **hunting**
- entities
  - models
    - **mathematical models**
  - structures
    - organic structures
      - cell components
        - **genes**
- features
  - physiographic features
    - land cover
      - **soil**
      - vegetation
        - **forests**
- groups
  - scientists
    - biologists
      - **zoologists**
- organisms
  - animals
    - useful animals
      - domestic animals
        - livestock
          - **cattle**
    - vertebrates
      - **mammals**
  - microorganisms
    - **pathogens**
  - **parasites**
- phenomena
  - biological phenomena
    - biological development
      - **growth**
    - diseases
      - **animal diseases**
    - **mortality**
- processes
  - genetic processes
    - **genetic variation**
- products
  - plant products
    - cereals
      - **wheat**
- properties
  - chemicophysical properties
    - **pH**
    - thermal properties
      - **temperature**
- resources
  - inputs
    - farm inputs
      - **fertilizers**
  - natural resources
    - nonrenewable resources
      - land resources
        - **pastures**
      - mineral resources
        - **lignite**
- sites
  - **laboratories**
- subjects
  - sciences
    - earth sciences
      - **soil science**
    - natural sciences
      - life sciences
        - biology
          - physiology
            - **plant physiology**
- substances
  - additives
    - food additives
      - **flavour enhancers**
  - agents
    - infective agents
      - **pathogens**
  - chemical compounds
    - organic compounds
      - hydrocarbons
        - aromatic hydrocarbons
          - **styrene**
  - elements
    - nonmetals
      - **nitrogen**

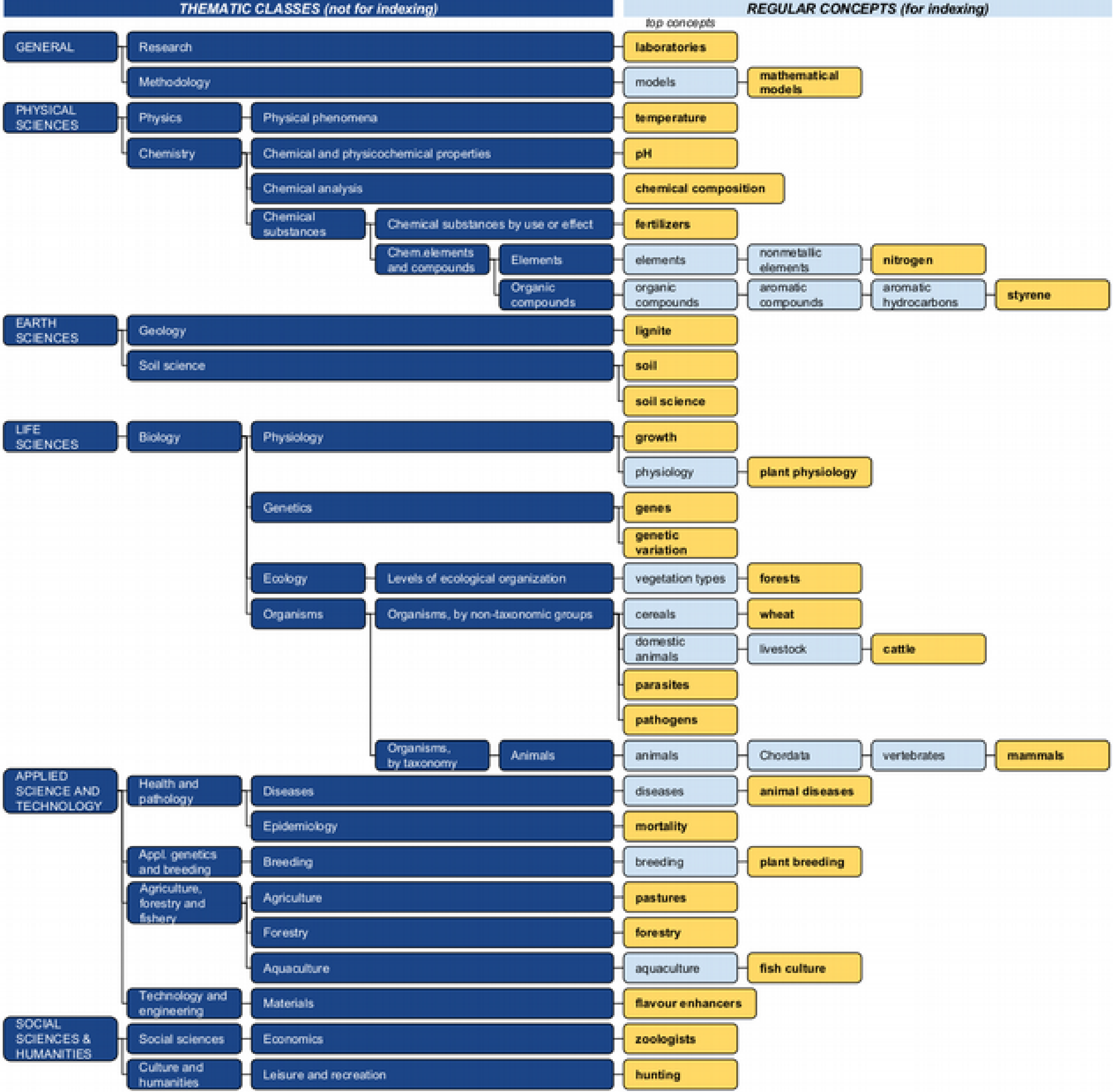
Strengths and Weaknesses of Scenario B

- clear top level concepts representing agricultural domain (4)
- fewer abstract layers than in A (2)
- good for browsing as hierarchy makes sense to users
- categorized from a user perspective
- eases reuse of GACS
- easy to align to other resources
- close to AGROVOC, which we are used to
- keeps the difference between a vocabulary and an ontology
- useful for information retrieval
- represents properties as such
- good balance between height and depth
- appropriate level of organization on lower levels

- ambiguity problems / unclear distinctions e.g. phenomena vs. entities (7)
- difficulty selecting and formalizing the categories (2)
- difficult to maintain (2)
- quite abstract at the top
- too messy and ad hoc
- resources required to execute
- more high level modelling effort required
- not friendly for human browsing
- not good for collecting terms in a subject area
- breadth of coverage is not apparent
- risk losing oversight of top level concepts when adding concepts deep into the hierarchy
- loses generic approach of DOLCE in A
- would probably need upper ontologies to link to other domain resources
- categories really need to be reworked
- too much stuff on the upper level
- tries to adopt an ontology-like approach but using the modelling tools of thesauri

Scenario C

based on CAB Classified Thesaurus



THEMATIC CLASSES (not for indexing)

REGULAR CONCEPTS (for indexing)

top concepts

GENERAL – Research – laboratories

Methodology – models – mathematical models

PHYSICAL SCIENCES – Physics – Physical phenomena – temperature

Chemistry – Chemical and physicochemical properties – pH

Chemical analysis – chemical composition

Chemical substances – Chemical substances by use or effect – fertilizers

Chem.elements and compounds – Elements – elements – nonmetallic elements – nitrogen

Organic compounds – organic compounds – aromatic compounds – aromatic hydrocarbons – styrene

EARTH SCIENCES – Geology – lignite

Soil science – soil

soil science

LIFE SCIENCES – Biology – Physiology – growth

physiology – plant physiology

Genetics – genes

genetic variation

Ecology – Levels of ecological organization – vegetation types – forests

Organisms – Organisms, by non-taxonomic groups – cereals – wheat

domestic animals – livestock – cattle

parasites

pathogens

Organisms, by taxonomy – Animals – animals – Chordata – vertebrates – mammals

APPLIED SCIENCE AND TECHNOLOGY – Health and pathology – Diseases – diseases – animal diseases

Epidemiology – mortality

Appl. genetics and breeding – Breeding – breeding – plant breeding

Agriculture, forestry and fishery – Agriculture – pastures

Forestry – forestry

Aquaculture – aquaculture – fish culture

Technology and engineering – Materials – flavour enhancers

SOCIAL SCIENCES & HUMANITIES – Social sciences – Economics – zoologists

Culture and humanities – Leisure and recreation – hunting

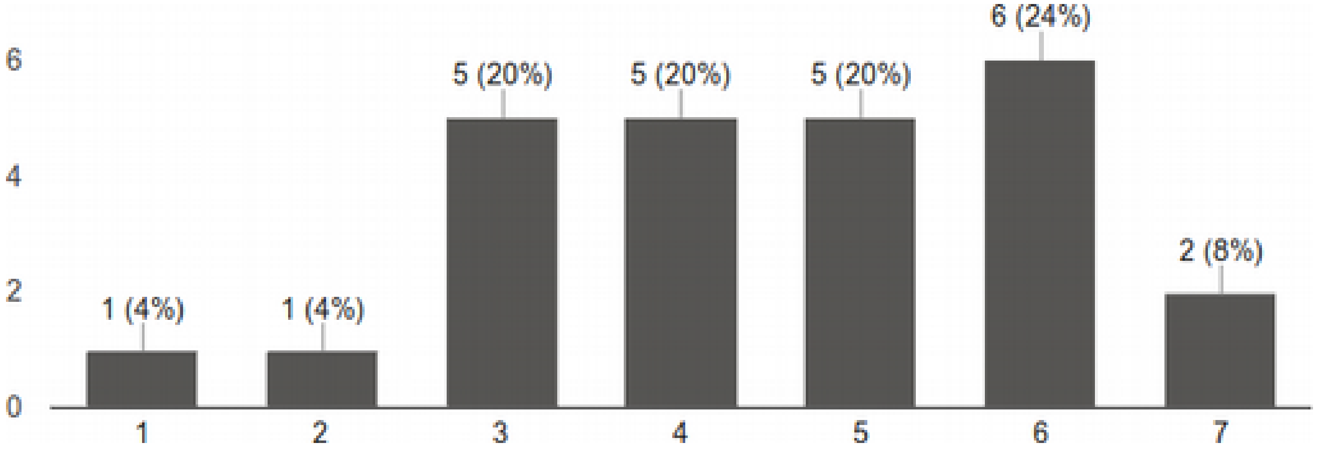
Strengths and Weaknesses of Scenario C

- intuitive to average users (3)
- conventional way of representing human knowledge (2)
- easy to maintain (2)
- less emphasis on concept hierarchy reduces messiness
- small footprint
- easy to reconcile with thematic classes
- not perfect but reasonable hierarchical paths
- already mostly accomplished in GACS
- good for collecting terms in a subject area
- breadth of coverage is apparent
- classificatory approach works even if modelling is shallow
- thematic classes outline the fundamental meaning of concepts
- ”forestry” and ”forests” close by, as expected
- narrows the margin of error with many maintainers working on hierarchy
- reflects the current version of GACS
- low depth
- clear and extensible thematic conceptualization layer
- doesn't make strong assumptions modelling upper and medium level concepts

- cannot represent interdisciplinary complexity (4)
- classified schemes are old-fashioned (2)
- difficult to choose discipline for many concepts (2)
- only limited use for search expansion / information retrieval (2)
- don't like division between indexable and non-indexable concepts – let users decide! (2)
- represents only scientific approach to domain (2)
- non-rigorous
- concepts could fit in more than one group
- weak semantics, no true IS-A relationships
- ”General” category is bad
- if you go for this, good luck with the modelling...
- losing the opportunity to express generic concepts
- continues current confusion
- some intermediary terms are confusing (e.g. materials)
- BT/NT not formal enough to perform automatic alignments
- difficult to use in cross-domain contexts
- top-level hierarchy needs work
- thematic classification is not a good basis for hierarchical structure
- there is no real hierarchy, so cannot be used to annotate data
- confusing entry points
- disorganized, doesn't give a lot of guidance

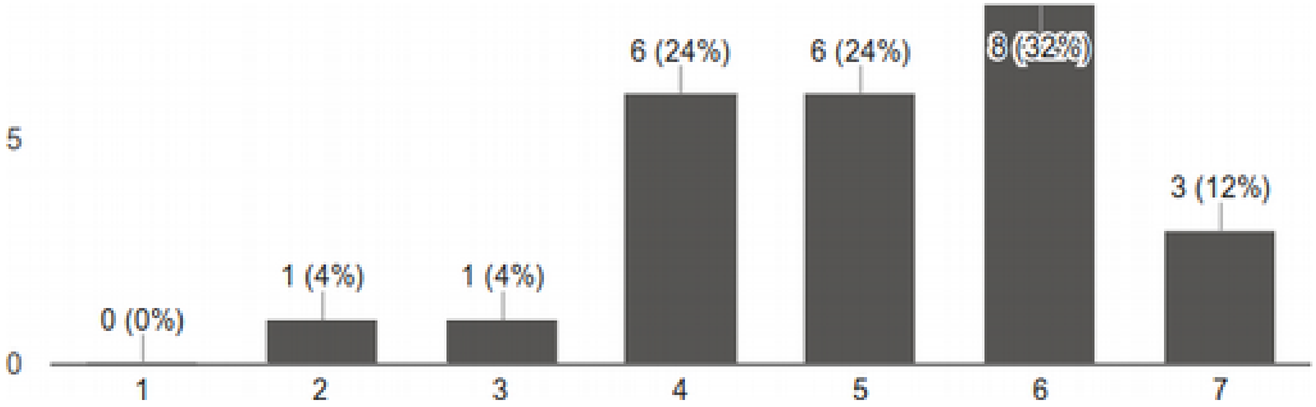
Vote results

A



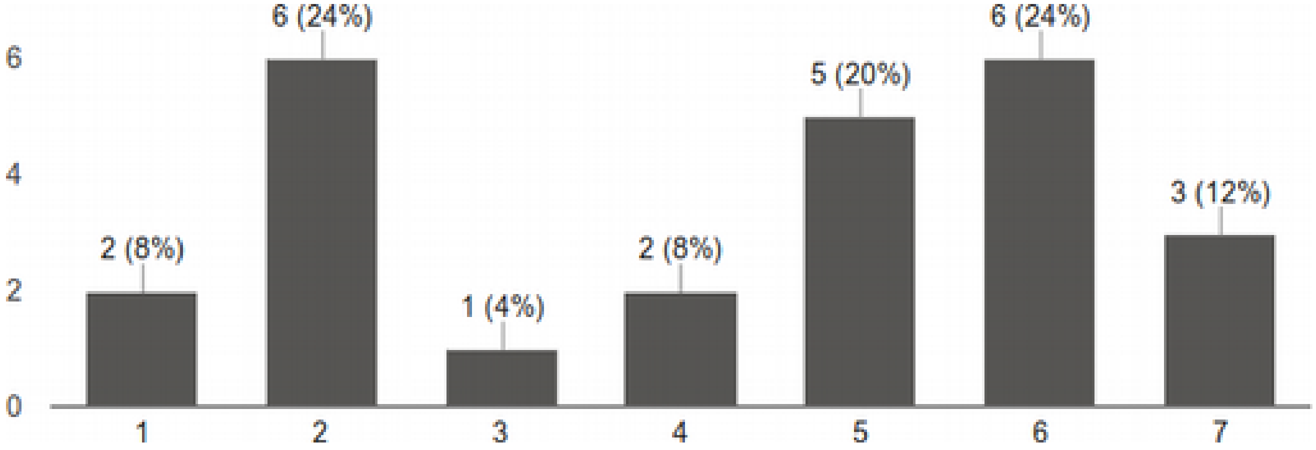
Median 5
Average 4.52

B



Median 5
Average 5.12

C



Median 5
Average 4.28

Suggestions for the hierarchy (1/2)

- these are complementary approaches that should be combined (5)
- need a multi-faceted approach (2)
- not a single big hierarchy, but several resources with light yet precise structures

- should have IS-A relationship view of concepts (3)
- should have three or so top concepts which make abstract distinctions which are easy to grasp, not only in English
- there is no need to have very small root levels (e.g. 3 concepts)
- should avoid a very large root level (e.g. 300 concepts)
- should skip intermediate levels of abstract, not-for-indexing concepts
- should have a balanced structure with 10-20 top concepts, like B
- should be broad and shallow, like B
- should have a thematic view
- I would look at the Basic Formal Ontology approach

If I had to design a hierarchy... (2/2)

- I would do it like Scenario A
- It would be similar to approach A with thematic classes resembling top concepts of B
- it would probably be close to structure B
- I would use a solution resembling B, while also incorporating elements of C
- I would choose a thematic top level categorization, like C
- I would probably mix A and B; B is the most accurate, but the DOLCE alignment would be interesting to avoid some model pitfalls

- does there need to be a hierachy in GACS core, or should it be the application developer's responsibility to develop their own to suit their requirements?
- I probably wouldn't design a hierarchy, I usually do things bottom-up

Some draft conclusions for the hierarchy

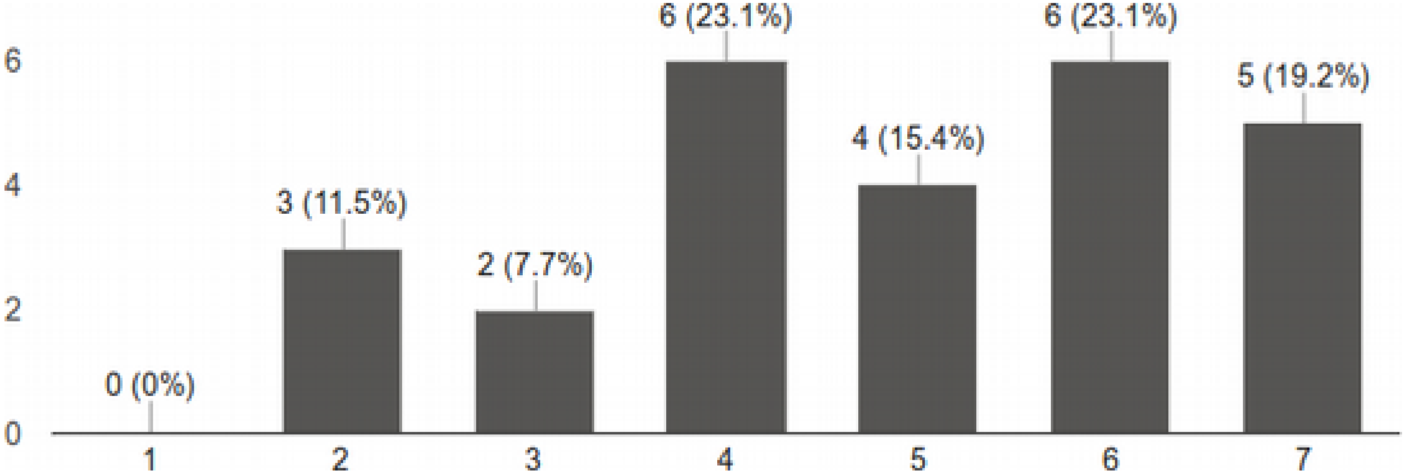
(Osma's thoughts, not directly from the survey)

1) We should keep the thematic groups as an additional view, with possibly some tweaks
   - allow multiple groups for a concept, with guidance
   - classify the remaining 20% of concepts which currently are not

2) We should consider adopting the current concept types (Organism, Chemical, Geographical, Product, Topic) as top concepts
   - maybe split off Property (and others?) from Topic
   - maybe create an additional layer of organization below these top concepts, especially for Topics which are quite many
   - could mean dropping the notion of concept types, to avoid encoding the same information in two different ways

3) In any case we need to continue cleaning up the hierarchy
   - reduce unwarranted polyhierarchy, for example by calculating scores for each BT/NT relationships based on metrics (e.g. shared among source thesauri?) and removing the worst ones
   - flag and correct situations where BT and NT have different concept types (~1300)
   - it would help to move some leaf concepts out of GACS Core

How useful are the GACS thematic groups to you?

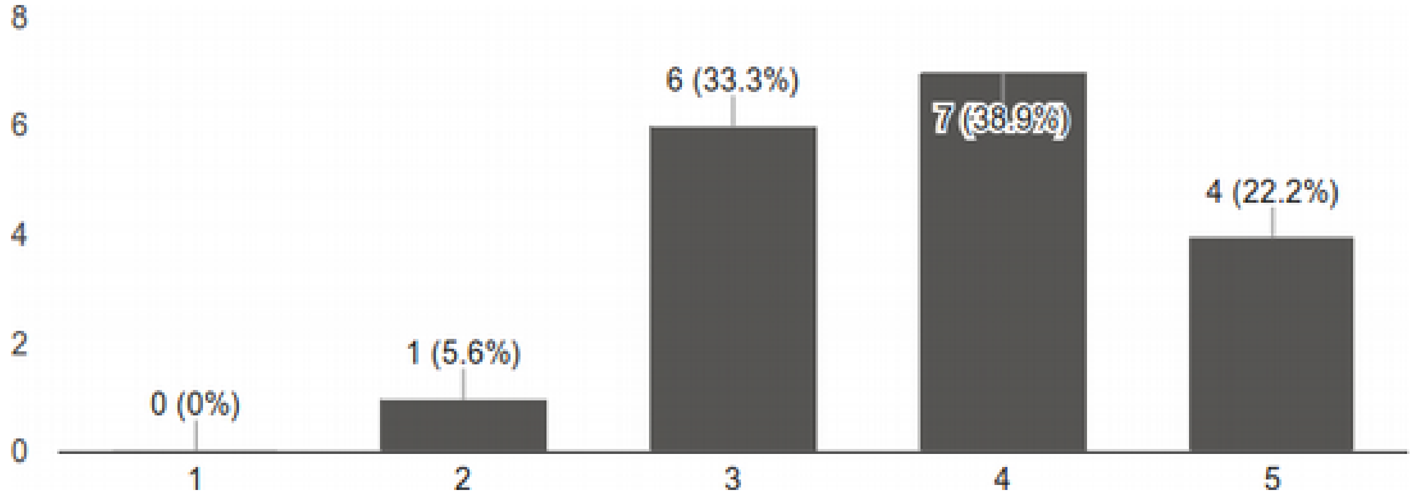


Median 5
Average 4.89

Suggestions for improvement of the thematic groups

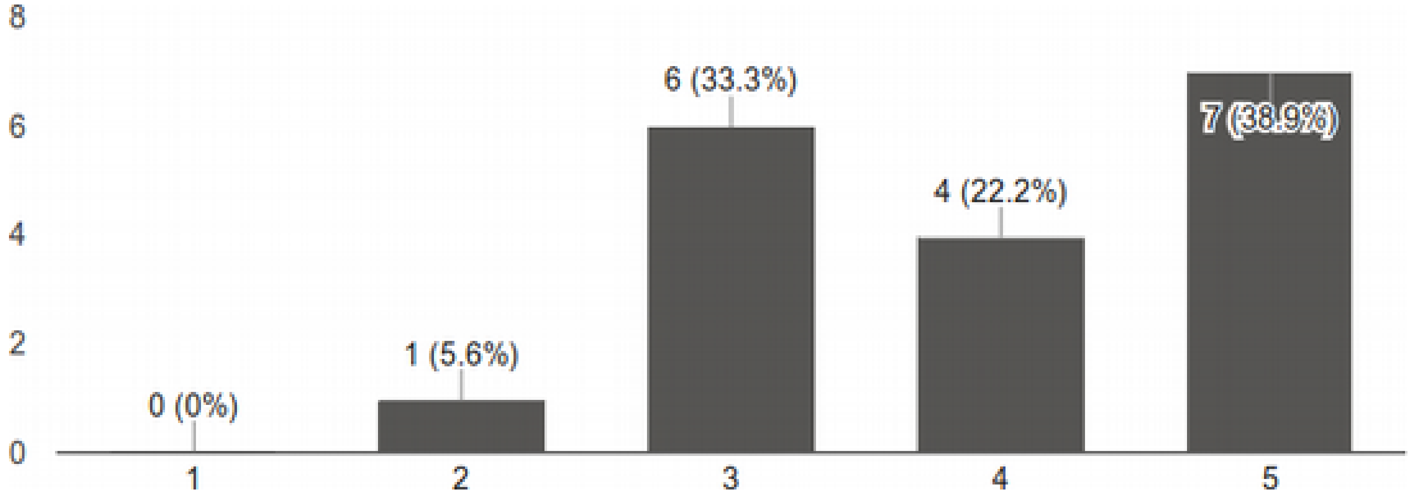
- implementing would require polyhierarchy and guidance, i.e. a concept can be in multiple groups (2)
- eliminate General category as it is not helpful (2)
- ”WJ policy, politics, government and law” is extremely broad, should be split
- should link to existing global classifications, e.g. DDC or UDC, or FAO's AGRIS subject categories

Is identifying part/whole relationships (meronyms) important to you?



Median 4
Average 3.78

Is identifying homographs in the vocabulary important to you?



Median 4
Average 3.94