Proposal for restructuring Scientific Names and Common Names of Organisms in AGROVOC

Margherita Sini[1], Dagobert Soergel[2], Gudrun Johannsen[1],
[1] **Food and Agriculture organization of the United Nations**
[2] **Maryland University, US**

Contributions from:
Aree Thunkijjanukij (KU, Thailand),
Vimlesh Yadav (IITK, India)
Magnus Grylle (FAO-FO)

Last release: April 2009

Introduction

Scientific names and common names in AGROVOC are not consistently used: some scientific names are descriptors and some are non-descriptors (even if they are the main accepted scientific name for an organism); likewise, some common names are descriptors and some are non-descriptors.
The main reason for this is that during indexing somebody may decide to tag a document with the scientific name while somebody else may decide to tag it with the common name.
In addition, experts that use AGROVOC only for searching or navigation (not for indexing) would like to use their preferred term, be it the scientific name or the common name.

The right approach would be to have an indexing system which does not use terms for indexing but rather a unique reference to a concept. However, this is not always possible, especially for legacy systems, therefore the right approach is to allow tagging of any document about an organism with its scientific name if it exists or with its common name if it exists, being sure that in AGROVOC they both refer to the same concept, or, if the specific organism cannot be ascertained, with a generic common name if the corresponding scientific name does not exist (e.g. "Stem borer" – which corresponds to many species and many genera). The solution described in this document will allow this through a specific definition of descriptors and non-descriptors.

There is also among some the practice of using the scientific name for scientific documents about an organism and the common name for documents about an organism that are addressed to practitioners such as farmers. However, a more systematic approach to finding documents at the desired level ("scientific" or "non-scientific" documents) is to filter results using the metadata element "type of document": for example, when somebody wants to retrieve scientific information he/she may decide to filter by just selecting from the types of document "scientific paper", while somebody that wants to find a common article may just select other "document types" from this specific metadata element.

The "Good Practices Guidelines for Indexing with AGROVOC" that documents about a given organism, should be tagged with the scientific name if scientific-oriented, while documents about the product derived from those organisms (e.g. the apple fruit from an apple tree, or goat as a meat product versus the goat as a living organism / livestock) should be tagged with a different term which identifies a different concept. However, this proposal will allow the indexing of document by a common name, being sure that this term refers to the same concept as the corresponding scientific name.

As a consequence, we should remember that different types of "concepts" in AGROVOC should be represented with different descriptors, while terms representing the same concept should be grouped together. For example, "potato" which refers to the potato plant, should be "grouped together" with the scientific name of the potato plant, i.e. Solanum tuberosum, while "potato" as a product, should be a different concept. Disambiguation must be done at the term level, if needed.

By "grouped together" we mean assigning a unique place in the hierarchy to a concept, even if this may have more than one preferred descriptor. Other terms which may be used for this concept should be non-descriptors (for these main descriptors).

In order to achieve this, we would like to add some modifications to the AGROVOC structure as explained later in this paper.

During search the system should support the user in finding relevant information. For such systems, we may have two cases:

1. traditional indexing systems which tag documents using terms in one or more languages; such a system can work in one of two ways:
    a. the system may decide to use only the query string for searching; in this case while indexing it may be possible that the system adds all non-descriptors and all other descriptors to the document together with the descriptor selected by the user (least preferred option);
    b. the system checks the query string in the thesaurus before retrieving documents; if the user types a non-descriptor, or another preferred descriptor, the system maps the user's search term to the corresponding main-descriptor and finds the documents indexed with that descriptor; in this case while indexing the system tags documents only with this main descriptor;

Note: by "main descriptor" we mean the one that will have attached a hierarchy; this may be a scientific name or a common name.

2. concept-based or URI-based systems:
    a. in general these kind of systems match the query string in the thesaurus/ontology and map the user's search terms to concepts (which may be expressed by URIs) and then search for the concepts; term disambiguation may part of the mapping..

In both cases, while searching,

- if the user types the scientific name, the system will retrieve all documents related to that organism;
- if the user types a common name, a sophisticate system will ask whether the user means the corresponding organism or the corresponding product (e.g. "do you mean the potato plant or potatoes (product)?").

Note: Latin names are repeated in AGROVOC for many languages, with consequent problem over performance, but this is considered to be ok within this current structure. This should perhaps be fixed later.

In view of creating a concept server from the AGROVOC thesaurus we have initiated a project called "AGROVOC revision and refinement"[1] with ICRISAT. ICRISAT has already given some instructions on how to improve AGROVOC, e.g. revise the hierarchical structure of terms in order to reduce the number of top terms (TT), and make sure that the hierarchies under those TT are consistent. This document amends the proposal for TASK 4 "revision of scientific terms and common names" of this project.

Other users from India (e.g. IITK) and from Thailand (e.g. KU Agris Center) are also working on enriching scientific names and common names in AGROVOC and creating other ontologies including scientific information.

Therefore, we are here proposing a new structure for scientific names and common names in AGROVOC, which would also facilitate further creation of sub-domain-specific ontologies.

This document aims at creating consensus among AGROVOC users on this new proposed method for managing taxonomies and common names.

Current situation

Currently in AGROVOC some common names of organisms are descriptors, some are non-descriptors, and some scientific names are descriptors and others non-descriptors.

This may be correct only under specific circumstances which are explained in this document: it is actually preferable to have **all commonly accepted scientific names of organisms as descriptors and all most used common-names as descriptors also (see below)**. Other non-commonly used or deprecated scientific name will be kept as non-descriptors.

Both the scientific name and the common name of an organism may be descriptors; they will linked with a specific new relationships called "sameAs". ONLY one of the descriptors will have a specific hierarchy.

Common-names that do not have a corresponding scientific name **can be descriptors**.

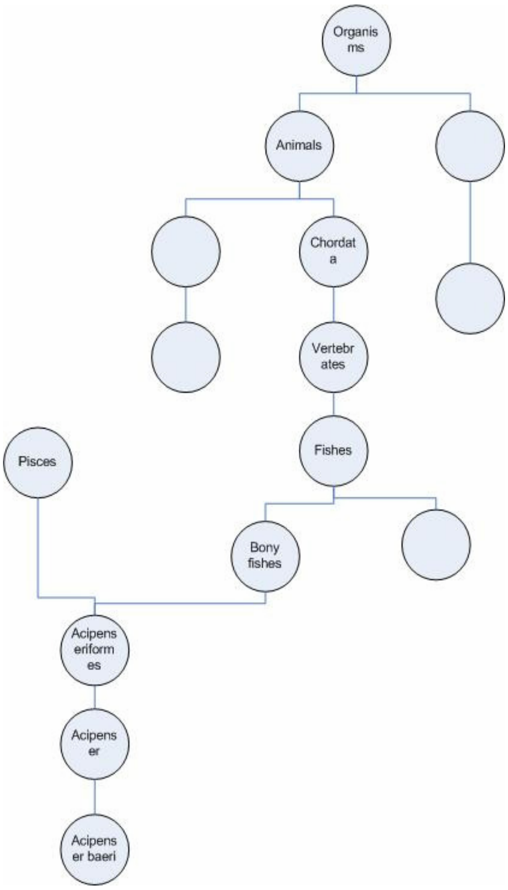
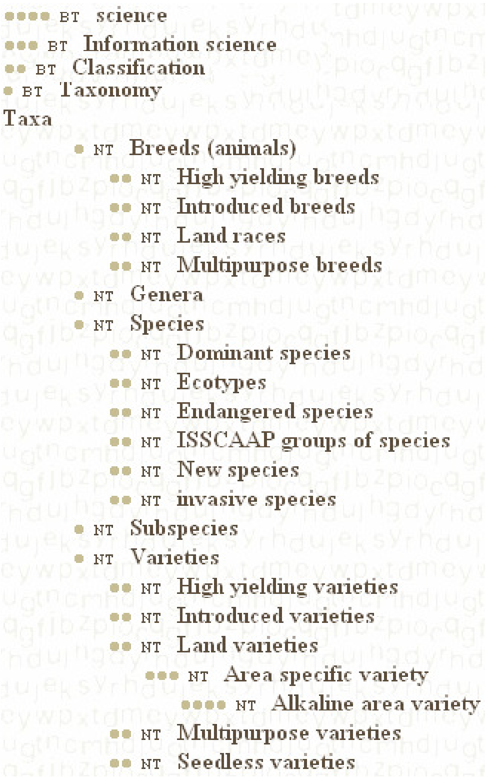


Figure 1

[1] See also http://agrovoc-revision-refinement.blogspot.com/

E.g. in Figure 1, all mentioned terms are descriptors in the current AGROVOC, but “Pisces” and “Fishes” correspond to the same concept.

In addition, in AGROVOC there is a structure of Taxonomic entity types (taxonomic levels) which is not fully developed nor fully exploited, i.e. we cannot retrieve all genera, all species, etc. See picture on the right (Figure 2).



```
●●●● BT science
●●● BT Information science
●● BT Classification
● BT Taxonomy
Taxa
     ● NT Breeds (animals)
         ●● NT High yielding breeds
         ●● NT Introduced breeds
         ●● NT Land races
         ●● NT Multipurpose breeds
     ● NT Genera
     ● NT Species
         ●● NT Dominant species
         ●● NT Ecotypes
         ●● NT Endangered species
         ●● NT ISSCAAP groups of species
         ●● NT New species
         ●● NT invasive species
     ● NT Subspecies
     ● NT Varieties
         ●● NT High yielding varieties
         ●● NT Introduced varieties
         ●● NT Land varieties
             ●●● NT Area specific variety
                 ●●●● NT Alkaline area variety
         ●● NT Multipurpose varieties
         ●● NT Seedless varieties
```

Figure 2

In AGROVOC we cannot currently identify what is a kingdom, a class or a family, because there is no specification of this information. For example, we cannot specify that “Lepidoptera” isOrderOf “Insecta” or “Scirpophaga” hasKingdom “Animalia”.

Also, while moving from AGROVOC to the Concept Server[2] we need to be able to state better relationships between scientific names and common names, identifying species, genus, families, synonyms of common names or scientific names, etc.

But most important we need to be able to create unique concepts for organisms from the mixture of scientific names and common names currently used. I.e., we would like to realize the structure represented in the following picture:

[2] http://www.fao.org/aims/

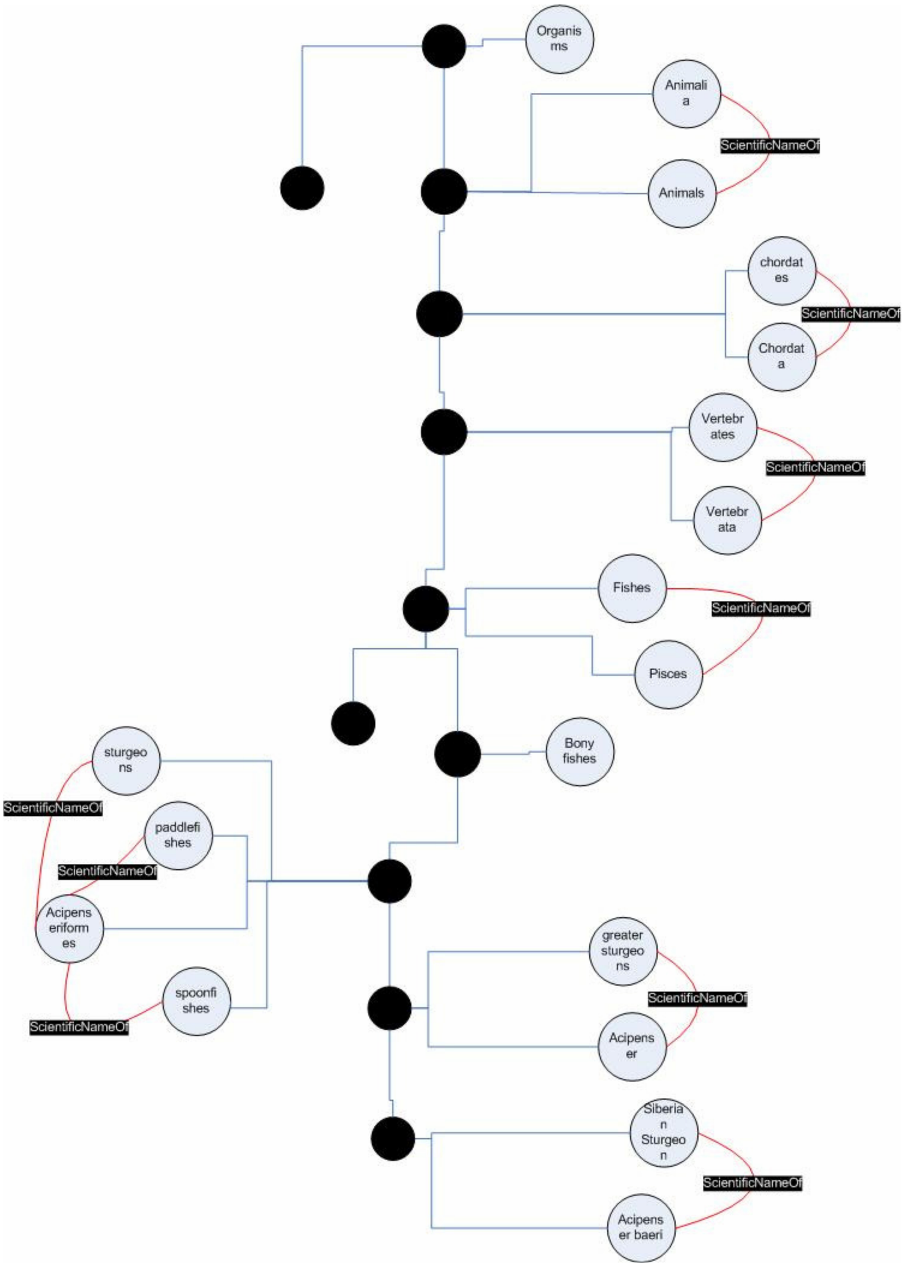


Figure 3

We have prepared a proposal for ICRISAT on restructuring the current AGROVOC; however, we have recently identified some additional problems, and we are therefore proposing this new solution.

Proposed solution

The proposed solution allows multiple descriptors (multiple terms that can be used in indexing) for identifying an organism. However, only one term will be the main descriptor which will be linked into the hierarchy of organisms. Other (alternate) descriptors will be linked to the main descriptor via the relationship "correspondsTo"; conversely, the main descriptor will be linked to alternate descriptors via the relationship "sameAs. Usually the main descriptor will be the most accepted scientific name, and the most used common-name may be an alternate descriptor; however, it may also be the other way around. (NOTE: see below for exceptions).

Again:	correspondsTo	from alternate descriptor to main descriptor
	sameAS	from main descriptor to alternate descriptor

Example:

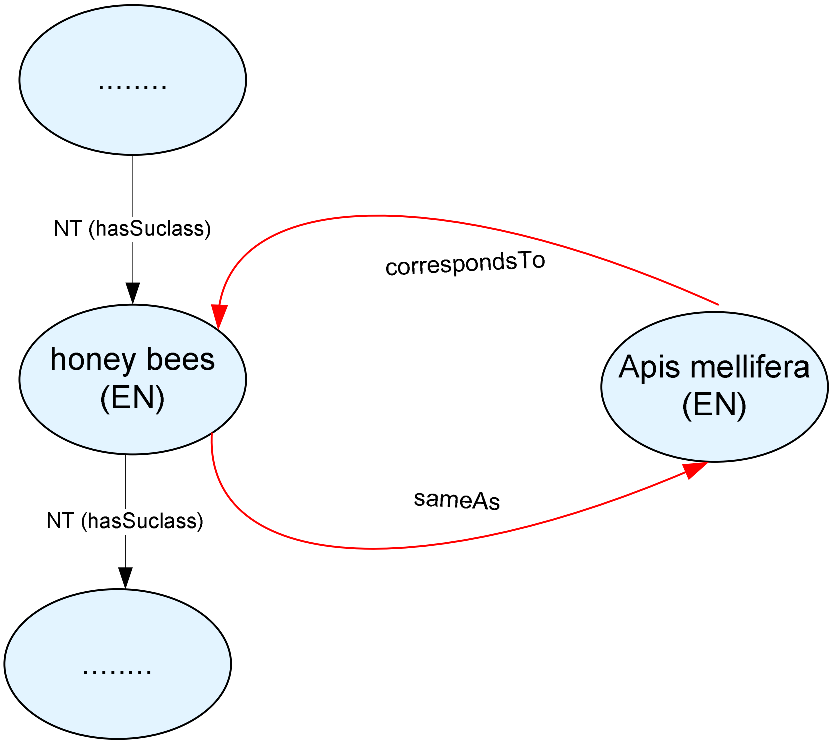


Figure 4

or

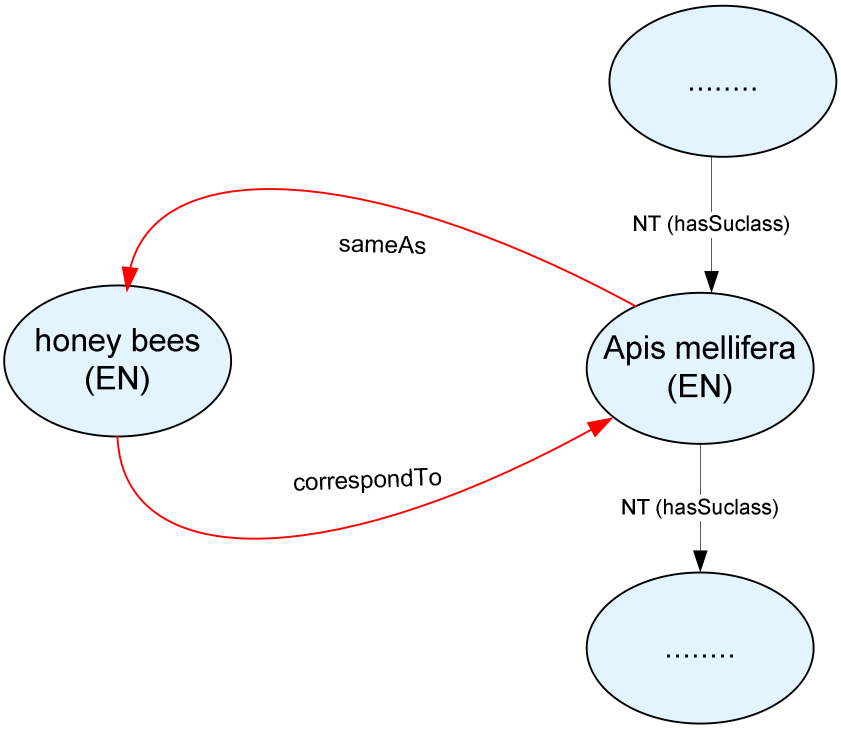


Figure 5

The picture above are both perfectly valid and all terms are DESCRIPTORS.

Currently in AGROVOC scientific names are assigned with a scope as follows:

TA	Scientific name for animals
TF	Scientific name for fungi
TP	Scientific name for plants
TB	Scientific name for bacteria
TV	Scientific name for virus

All common names will be assigned with a new scope as follows:

CA	Common name for animals
CF	Common name for fungi
CP	Common name for plants
CB	Common name for bacteria
CV	Common name for virus

In addition: all main descriptors for organisms should have a relationship isTaxonomicLevel (unrefined RT) to a concept from taxa (order, species, family, etc.); this would allow the expression of hierarchical relationships between organisms (hasSpecies, hasFamily, hasOrder, ... instead of just hasSubclass).

For all scientific names only the latest accepted scientific name should be descriptor. All other scientific and non-scientific names for the same Species or Genus or Family, etc. should be non-descriptors properly refined (e.g. “isHistoricalNameOf / hasHistoricalName”, etc.).

All common names **NOT identifying an organism** but **a product** should be in the hierarchy of products, which will refer to top level concepts such as "animal product" or "plant product". Note: Products do not have scientific names. When the same common name may refer to both an organism and a product derived from this organism, it must be disambiguated, e.g., potato plant and potatoes (product) or apple (plant) and apple (product).

Some common names designate an organism or a class of organisms that does **NOT have a scientific name; such common names may be descriptors** and can be incorporated in the hierarchy of the organisms. These may have scientific names as BT or NTs, and can also have other concepts as BTs (e.g. Stem borer case, see following pictures). This can create what we call polyhierarchies in AGROVOC.

Scientific names for organisms should be capitalized based on scientific rules; common names for organisms should not be capitalized. The common names identifying a product should NOT be capitalized (e.g. rice).

In summary, this proposal suggests that:

- Of all scientific names for an organism, the most common accepted one is selected as the main (or possibly the alternate) descriptor; the other scientific names are non-descriptors linked by USE relationships, refined as synonym or something more appropriate (there are usually many scientific names for the same organism, such as old scientific names that have been changed, and spelling variants).
  Note: Polyhierarchies are possible. Therefore we may have a hierarchy in which we have a mixture of generic concepts and scientific names.
- All the common names identifying an organism may also be descriptors (generally alternate descriptors, one common name may be the main descriptor); these may include local names, etc.
- only the main descriptor can (and should) have BTs, NTs, RTs. All (main or alternate) descriptors may have USED-FOR relationships.
- The relationships between scientific names and common names should be sameAs, refined as hasScientificTaxonomicName, isScientificTaxonomicNameOf.
- Every main descriptor for an organism (usually the most common scientific name) should have a isTaxonomicLevel (unrefined: RT) with one of the taxa groups: Species, Subspecies, Genera, (others will be created); e.g.

Oryza sativa	isTaxonomicLevel (RT)	species
Oryza glaberrima	isTaxonomicLevel (RT)	species
Animalia	isTaxonomicLevel (RT)	Kingdom
Plantae	isTaxonomicLevel (RT)	Kingdom

etc.

- The BT of common names representing PRODUCTS is different from the hierarchy of organisms, and will be organised independently. NOTE: it may be possible that disambiguation would be needed (e.g. "potatoes (plant)" and "potatoes (product)", the second being under vegetables as products).
- The common names representing PRODUCTS may be in RT with the corresponding organism (e.g. "potatoes (product)" RT (refined as: derivedFrom) "Solanum tuberosum").
- Organisms for which the main descriptor is a common name form part of the organism hierarchy; the common name will appear between scientific names.

During the revision, we should keep in mind that the hierarchy built with BT may be used in indexing systems for query expansion. Therefore a query with a species may be expanded by looking to all the organisms of the upper and/or lower level taxonomic group.

In this context, we do not have problems:

- if some organisms do not have a common name; they will be descriptor in the right position of the taxonomic hierarchy;
- if a common name such as “stem borer” designates and organism or a class of organism that does not have a scientific name, this common name will be included in the organism (poly-) hierarchy; for example, “stem borer” is a descriptor that has as NT all the organisms (mostly species) that fall under this broad general category; stem borer will be an NT of the lowest taxonomic class that includes all the organisms that fall under stem borer.
- the connection of scientific names by specific relationships such as hasSpecies, hasFamily, hasOrder, etc., can be built automatically.

In this context, there is a problem, if we want to create a non-descriptor for a non-descriptor: e.g., it will not be possible to say that “Stemborer” is a spelling variant of “Stem borer” as both are non-descriptors. This is solved within the AGROVOC Concept Server.
But we can identify better non-descriptors for a main descriptor and non-descriptors for corresponding other descriptors.

See figure below for an overview of the proposal.

An example of an organism which has no scientific name in AGROVOC is the concept “cattle”: this should be a descriptor, in the hierarchy of the organisms and be also a concept in the livestock hierarchy. Currently it has 2 BTs “Bovinae” and “Livestock”, an example of polyhierarchy. Note: Many other organisms or taxa may need a BT to livestock.

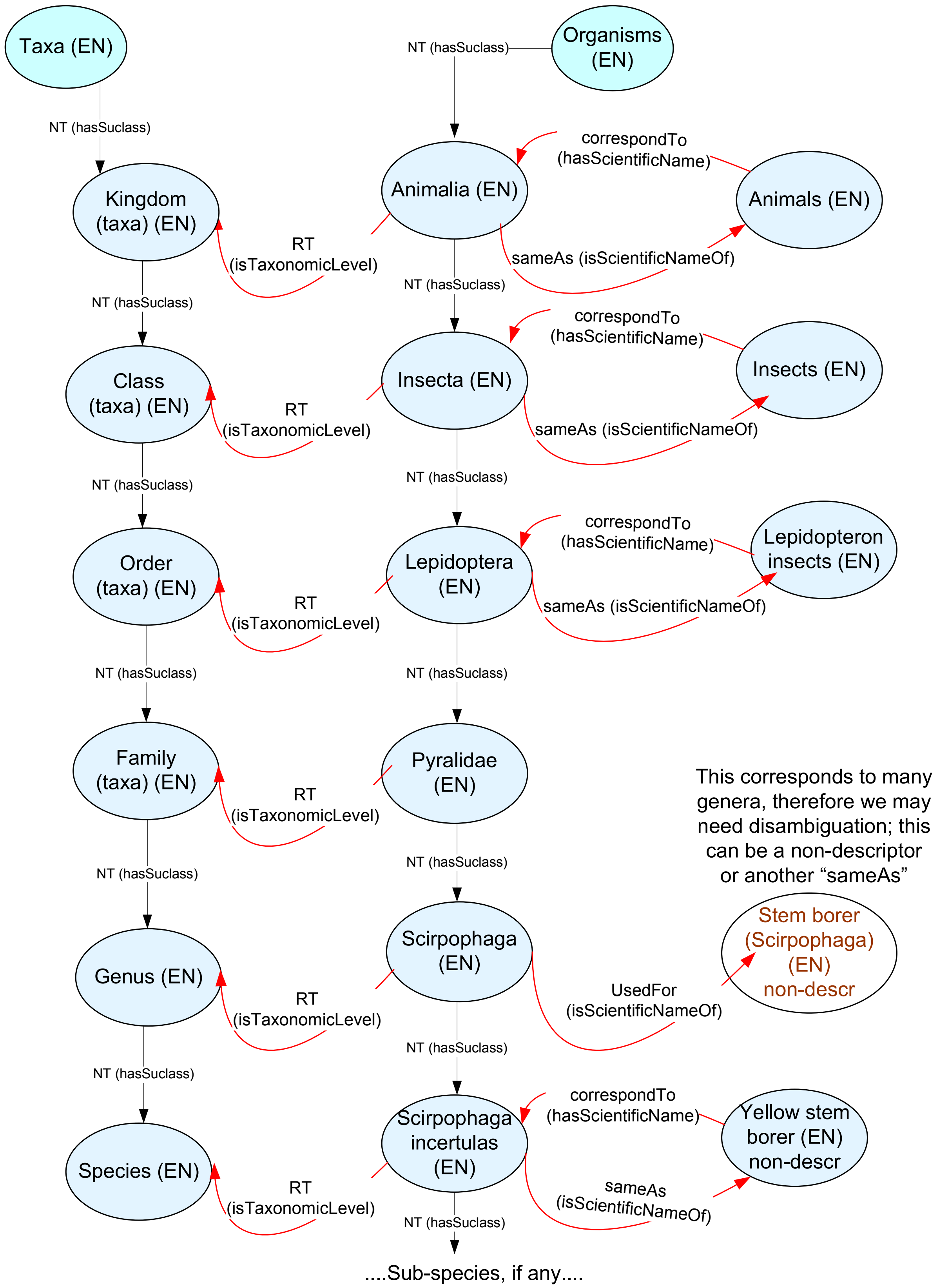

Taxa (EN)
NT (hasSuclass)
Organisms (EN)
NT (hasSuclass)
Kingdom (taxa) (EN)
RT (isTaxonomicLevel)
Animalia (EN)
correspondTo (hasScientificName)
Animals (EN)
sameAs (isScientificNameOf)
NT (hasSuclass)
NT (hasSuclass)
Class (taxa) (EN)
RT (isTaxonomicLevel)
Insecta (EN)
correspondTo (hasScientificName)
Insects (EN)
sameAs (isScientificNameOf)
NT (hasSuclass)
NT (hasSuclass)
Order (taxa) (EN)
RT (isTaxonomicLevel)
Lepidoptera (EN)
correspondTo (hasScientificName)
Lepidopteron insects (EN)
sameAs (isScientificNameOf)
NT (hasSuclass)
NT (hasSuclass)
Family (taxa) (EN)
RT (isTaxonomicLevel)
Pyralidae (EN)
This corresponds to many genera, therefore we may need disambiguation; this can be a non-descriptor or another “sameAs”
NT (hasSuclass)
NT (hasSuclass)
Genus (EN)
RT (isTaxonomicLevel)
Scirpophaga (EN)
UsedFor (isScientificNameOf)
Stem borer (Scirpophaga) (EN) non-descr
NT (hasSuclass)
NT (hasSuclass)
Species (EN)
RT (isTaxonomicLevel)
Scirpophaga incertulas (EN)
correspondTo (hasScientificName)
Yellow stem borer (EN) non-descr
sameAs (isScientificNameOf)
NT (hasSuclass)
....Sub-species, if any....

Figure 6

And the following picture for the hierarchy of products:

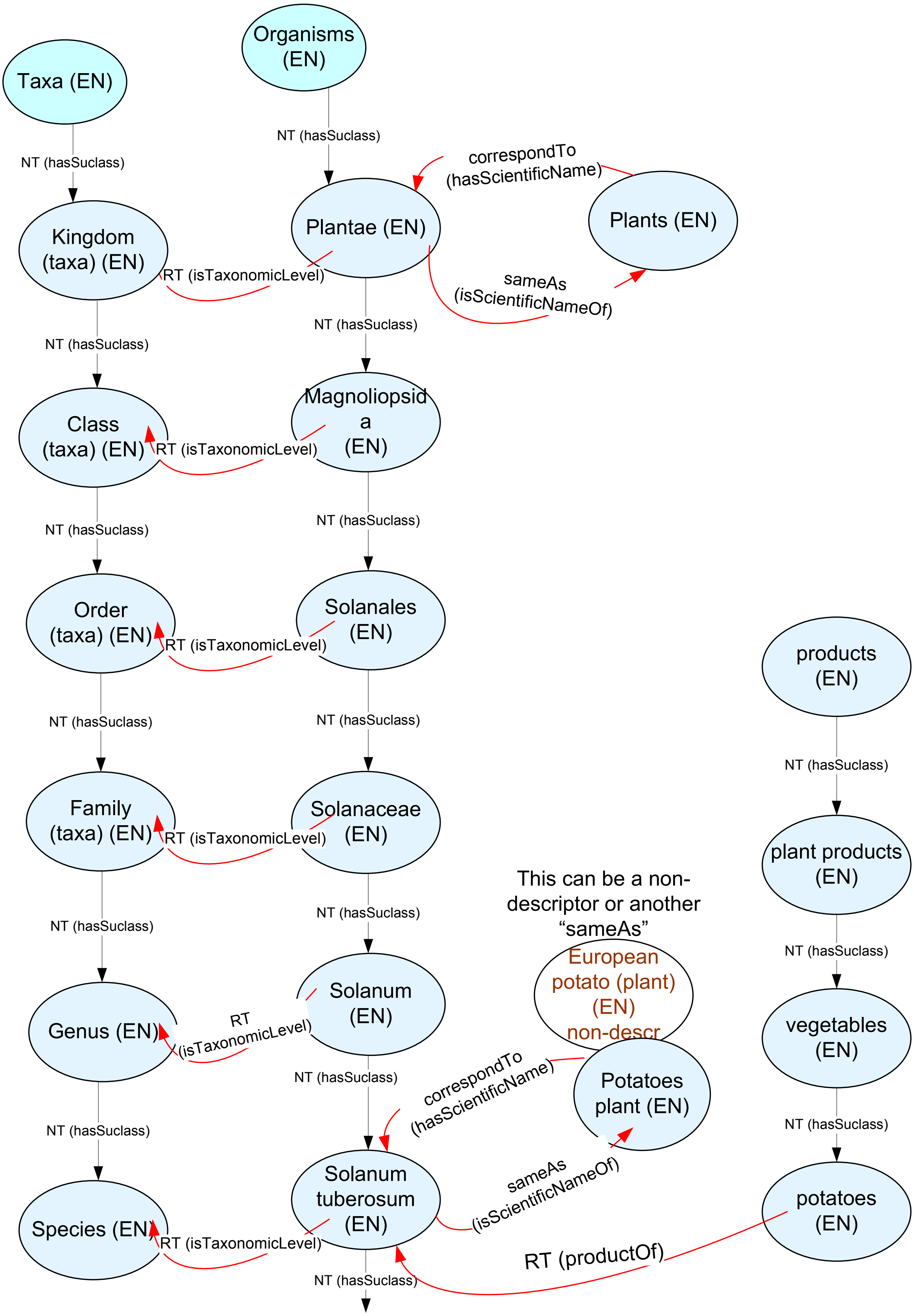

Taxa (EN)
Organisms (EN)
NT (hasSuclass)
NT (hasSuclass)
correspondTo (hasScientificName)
Kingdom (taxa) (EN)
Plantae (EN)
Plants (EN)
RT (isTaxonomicLevel)
sameAs (isScientificNameOf)
NT (hasSuclass)
NT (hasSuclass)
Class (taxa) (EN)
Magnoliopsida (EN)
RT (isTaxonomicLevel)
NT (hasSuclass)
NT (hasSuclass)
Order (taxa) (EN)
Solanales (EN)
RT (isTaxonomicLevel)
products (EN)
NT (hasSuclass)
NT (hasSuclass)
NT (hasSuclass)
Family (taxa) (EN)
Solanaceae (EN)
RT (isTaxonomicLevel)
plant products (EN)
This can be a non-descriptor or another “sameAs”
NT (hasSuclass)
NT (hasSuclass)
NT (hasSuclass)
European potato (plant) (EN) non-descr
Genus (EN)
Solanum (EN)
RT (isTaxonomicLevel)
vegetables (EN)
NT (hasSuclass)
correspondTo (hasScientificName)
Potatoes plant (EN)
NT (hasSuclass)
NT (hasSuclass)
sameAs (isScientificNameOf)
Species (EN)
Solanum tuberosum (EN)
RT (isTaxonomicLevel)
potatoes (EN)
RT (productOf)
NT (hasSuclass)

Figure 7

This hierarchy to which organisms are related may be related to products, or other categories of elements derived from organisms.

The hierarchy of organisms includes classes of organism that do not fall into the taxonomic hierarchy; such classes are often designated by a common name. In the example below, the different genera (more often species will be appropriate here) should be NT of *stem borer*. If a user does a search for *stem borer*, s/he should find documents indexed with any of the specific organisms that are considered stem borers.
For example see the picture below:

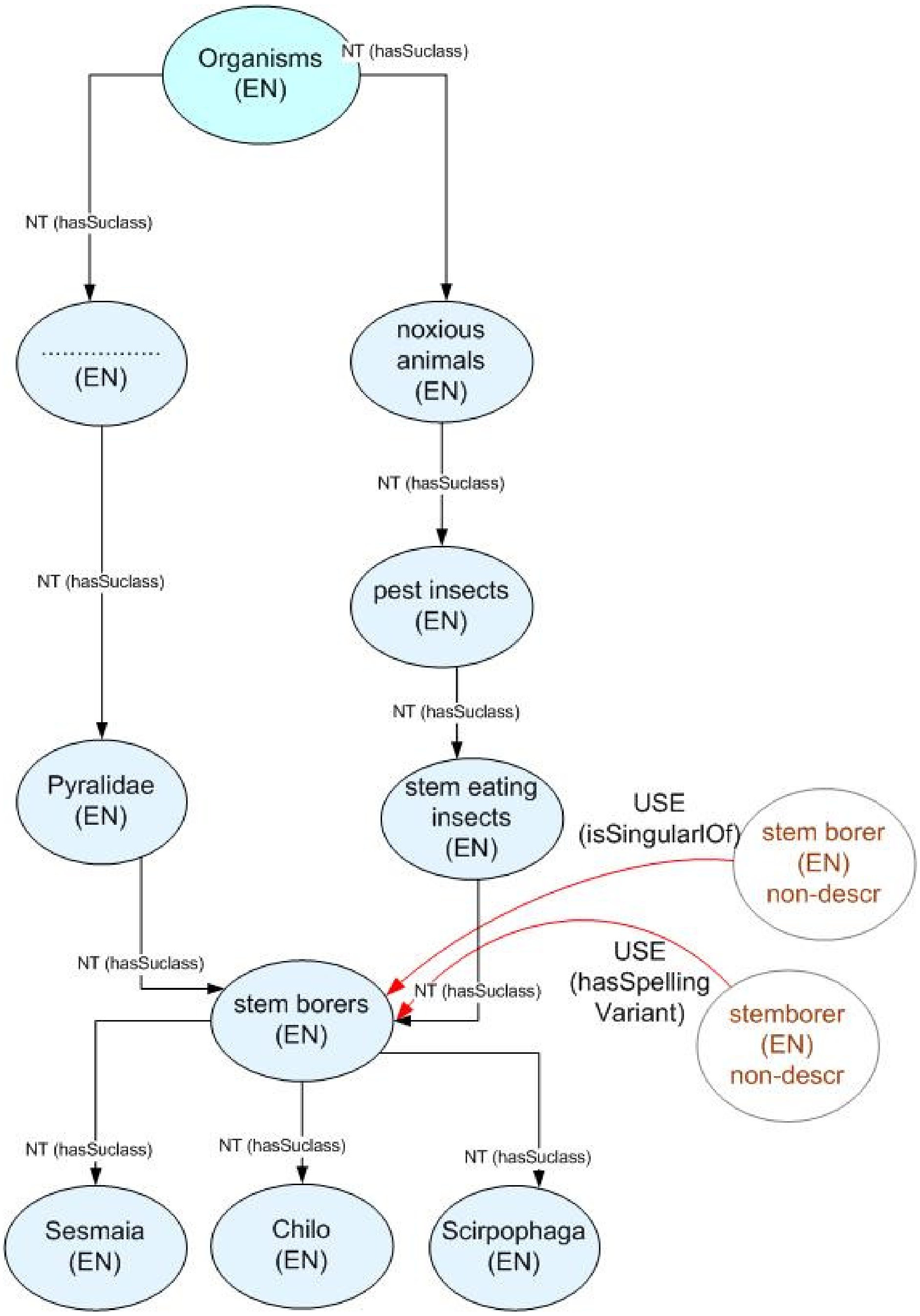


Figure 8

If needed, we may also add the following top level relationships:

– “plant product”	derivedFrom (RT)	“plantae”;
– “animal product”	derivedFrom (RT)	“animalia”;
– “noxious animals”	subclassOf (BT)	“organisms”.

Further example of the proposal

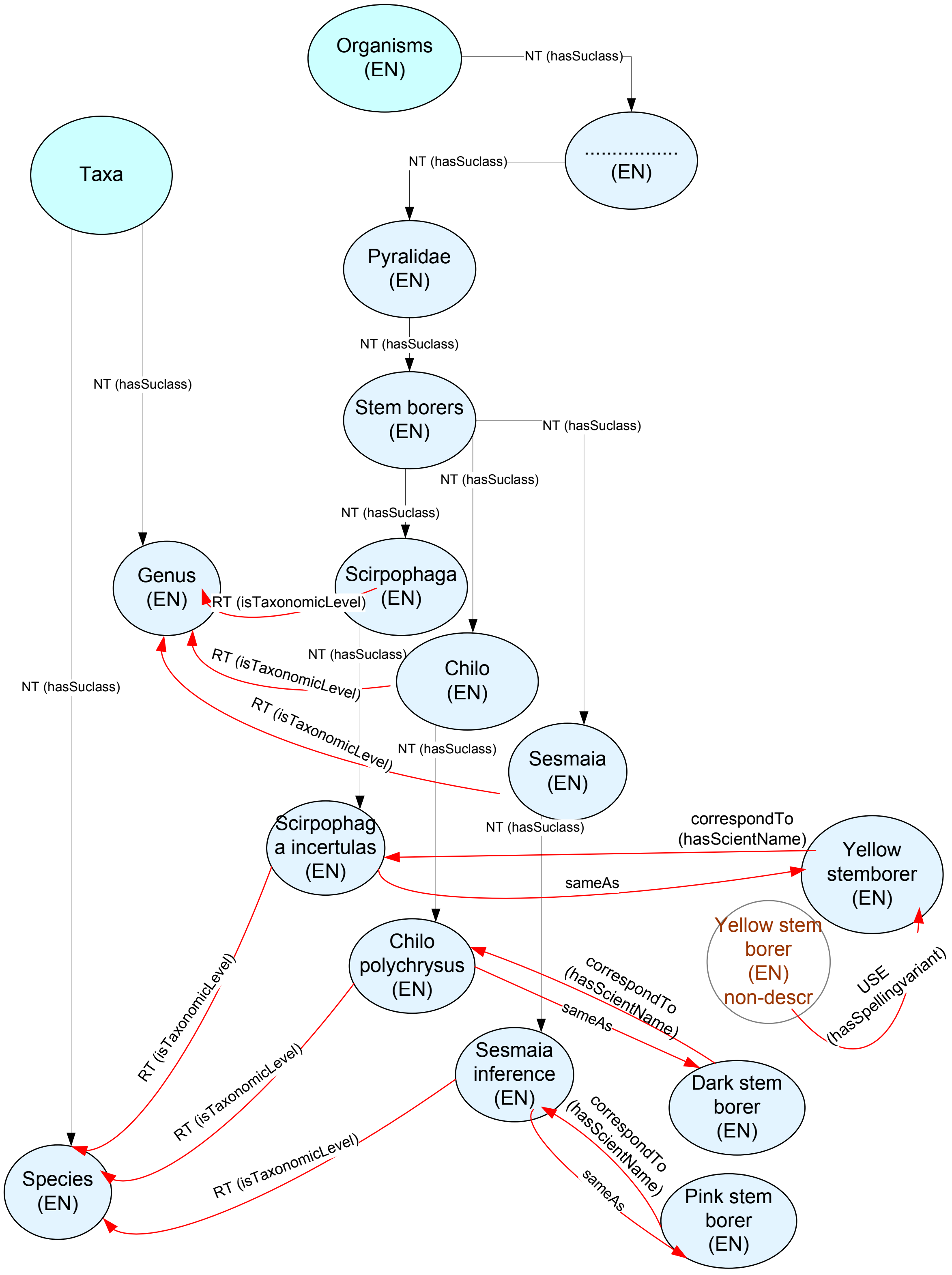


Figure 9

Changes to perform in AGROVOC

We need to perform the following changes in AGROVOC:

1. add a term “Kingdom” and put BT taxa;
2. add a term “Class (taxa)” and put BT taxa;
3. add a term “Order (taxa)” and put BT taxa;
4. add a term for “Family (taxa)” and put BT taxa.

Add new relationships in the linktape table:

link typeid	language code	linkdesc	Linkabr	linkdescription	createdate	rlinkcode	parent link typeid	link level
1	EN	sameAs	sameAs	Used for descriptors of the same concept	2008-11-06	2		TR
2	EN	correspondsTo	correspondsTo	To link an alternative descriptor to the descriptor that has the relationships	2008-11-06	1		TR

Add rows in the scope table to accommodate new scopes for common name plants, common name animals, etc. Codes will be:

CA	Common name for animals
CF	Common name for fungi
CP	Common name for plants
CB	Common name for bacteria
CV	Common name for virus

Once these points are done, the task n. 4 of the LoA can begin by following this procedure:

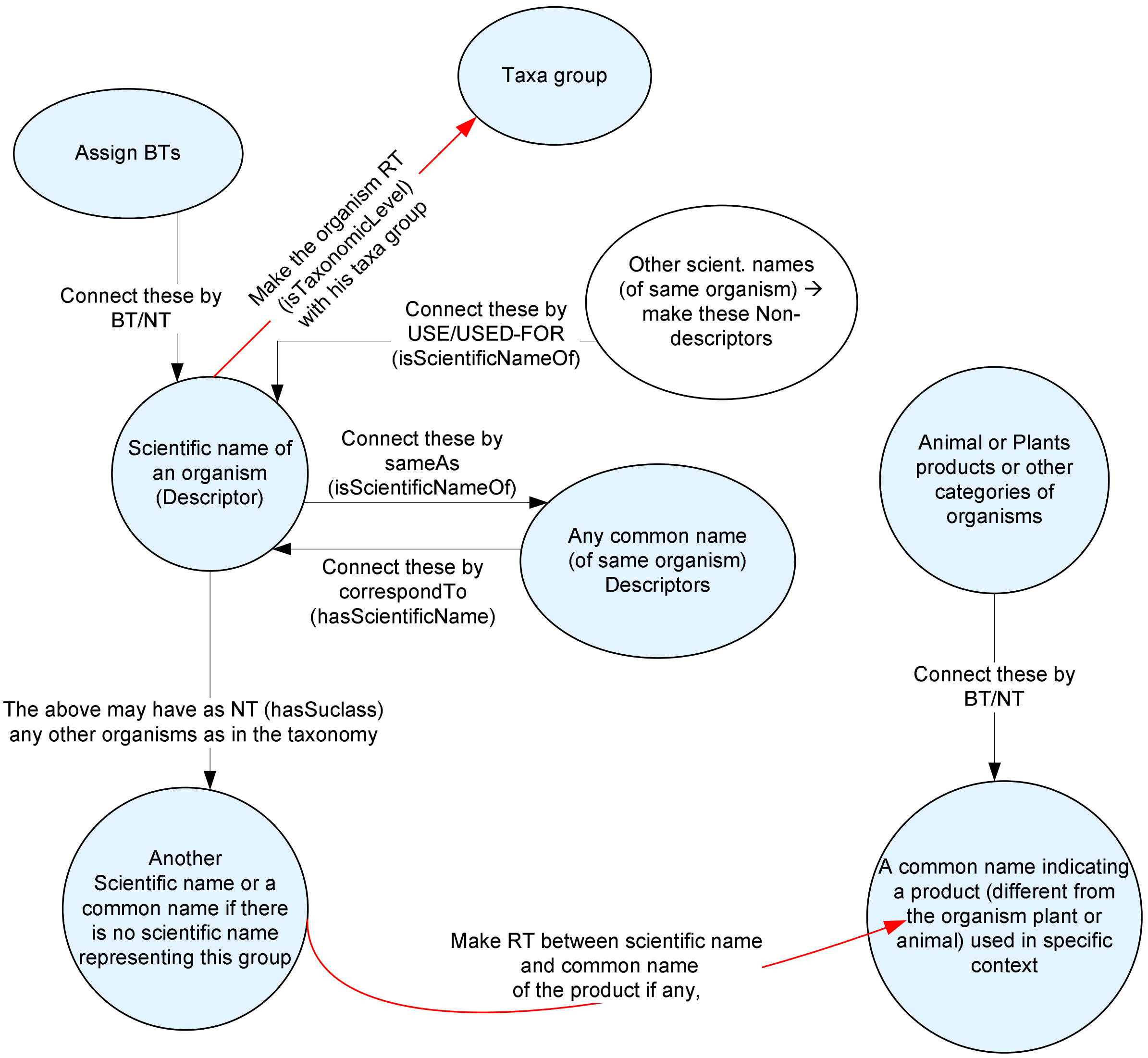

Taxa group
Assign BTs
Make the organism RT (isTaxonomicLevel) with his taxa group
Connect these by BT/NT
Connect these by USE/USED-FOR (isScientificNameOf)
Other scient. names (of same organism) → make these Non-descriptors
Scientific name of an organism (Descriptor)
Connect these by sameAs (isScientificNameOf)
Any common name (of same organism) Descriptors
Connect these by correspondTo (hasScientificName)
Animal or Plants products or other categories of organisms
Connect these by BT/NT
The above may have as NT (hasSuclass) any other organisms as in the taxonomy
Another Scientific name or a common name if there is no scientific name representing this group
Make RT between scientific name and common name of the product if any,
A common name indicating a product (different from the organism plant or animal) used in specific context

Figure 10

Conclusions

The proposed solution has been analyzed together with other alternatives. It was recognized to be acceptable for the current AGROVOC thesaurus and the future concept server, as well as the creation of sub-domain specific ontologies.

The organisms will be identified with a unique descriptor. This will facilitate the conversion to an ontology.

The scientific relationships between taxonomic groups are ensured by the identification of taxonomic level (class, family, group, etc).

Different hierarchies of organisms can be built using specific functionalities of the AGROVOC tools, by either creating new categories or reorganizing existing concepts.