Guiding Principles for Thesaurus Construction of the NAL Agricultural Thesaurus (NALT)

February 24, 1999; Revised: March 26, 2014; Revised by Sujata Suri: April 17, 2018

Table of Contents

I. Purpose of the NAL Thesaurus

The NAL Thesaurus may serve one to many purposes in a subject and retrieval information system. The thesaurus is primarily meant to be used as a controlled vocabulary for the subject indexing of AGRICOLA (NAL Catalog), PubAg and Agricultural Research Services (ARS) projects. NALT is also used by ARIS, Ag Data Commons (beta), FSRIO, AgNIC, US Forest Service, etc., sometimes through NALT annotator, an indexing tool. NALT is bilingual (English & Spanish), Spanish version of NALT is used by IICA (Inter-American Institute for Cooperation on Agriculture).

Additionally, the thesaurus may be used as a means to enhance the search of the information system, either automatically or manually. Integration of the thesaurus with the search provides a means of enhancing recall and precision for the searcher without the searcher's knowledge of the thesaurus. NALT is hierarchically arranged and constructed with the intent of providing an inclusive search to enhance search queries. Searches may be further enhanced by the use of the equivalence, Hidden labels, in the thesaurus. Search terms which have equivalence and OR Hidden labels in the thesaurus have those synonymous and Hidden labels added to the search query.

II. Upper structure and Categories in the NAL Thesaurus

NALT has an upper structure consisting of 17 “discipline-oriented” categories. The categories are designated by SC (Subject Category). The categories were selected for web presentation, not for a paper-based thesaurus. The intent of the categories is for users of the thesaurus to find the terms in their subject area of interest despite their viewpoint. For example, a user who is concerned with administration of forest resources may intuitively search under the "Government, Law and Regulations"category for the concept of "forest policy." However, a user who is a forester may intuitively search under the "Forest Science and Forest Products" category to find the concept of "Forest policy." The categories should not be confused with any intent other than the means for providing subject-oriented access to the thousands of terms in the thesaurus.

Subject categories are allowed as descriptors. The top categories were entered in MultiTes as not only descriptors but kept as subject categories so that different reports could be run based on that relationship. The 2010 edition was the first edition allowing indexers to use the top categories as descriptors.

III. Subject Scope of the NAL Thesaurus

The subject scope and content of AGRICOLA, which defines agriculture broadly, is the premise for the subject scope in the NAL Agricultural Thesaurus. It includes agriculture and allied disciplines, including animal and veterinary sciences, entomology, plant sciences, forestry, aquaculture and fisheries, farming and farming systems, sustainable agricultural systems, agricultural economics, extension and education, food and human nutrition, food safety, earth and environmental sciences, etc. The scope is continuously evolving with the evolution of journals, subjects and literature.

IV. Descriptors, Nondescriptors and Hidden Labels

a. Descriptors or Preferred Terms (DE)

A descriptor is the preferred term to express a concept for indexing and retrieval. For accurate, clear, and consistent indexing, only one term should be used to represent a concept. A descriptor is the allowed term that can be used by indexers to describe the subject content of abstracts or titles, etc. In the example below, "Schizaphis graminum" is the descriptor and "greenbug" and "Toxoptera graminum" are the non-preferred terms, or lead-in terminology, that direct users to the appropriate term for indexing and retrieval.

greenbug USE Schizaphis graminum

Schizaphis graminum
Used for: greenbug
Used for: Toxoptera graminum

Toxoptera graminum USE Schizaphis graminum

a. *Non-descriptor or Non-preferred term or Used for or Lead-in terminology*

A non-descriptor is not a valid term for assigning to a document. Non-descriptors direct users to the preferred terms for indexing and retrieval. Non-descriptors appear in italics and it makes easier for indexers to distinguish it from descriptors.

b. *Hidden labels (HL)*

A hidden label is a term that is seen in NALT draft but is not in the web version of NALT. Hidden labels are removed from web version of the NALT as these are meant to assign a descriptor that we can’t say as a synonym, (singular or plural terms, hyphenated terms, phrases, etc.). Hidden labels appear in italics in NALT draft version or in MultiTes. It is denoted by MR (Machine Read To).
For example: bio-gas MR (Machine Read To) biogas, where biogas is a descriptor.

V. Relationships between Terms

NALT includes hierarchical, equivalence and associative relationships among concepts. The following relationships are used in the thesaurus:

USE/USE FOR	USE/UF
Hidden Label/Machine Read To	HL/MR
Broader term/Narrower term	BT/NT
Related terms	RT/RT
AND type USE/USE FOR	USA/UFA

a. *Hierarchical relationships*

Hierarchical relationships are indicated by "Broader Term/BT" and "Narrower Term/NT" designations in the thesaurus. The hierarchical relationship is a distinguishing feature of the thesaurus in contrast to a simple list of alphabetically ordered terms. Superordinate "Broader Terms" represent more general concepts than subordinate "Narrower Terms":

intercropping
Broader terms: cropping systems
Narrower terms: alley cropping

b. *Equivalence relationships*

Equivalence relationships are designated by USE (Use cross reference) and UF (Used For cross reference). An equivalence is made between terms when the two terms represent the same or nearly the same concept, e.g., synonymous terms, quasi-synonyms, common names of organisms and their scientific equivalent, spelling variants, usage variants, acronyms, e.g.,

c. *Associative relationships*

Associative relationships are designated by Related Terms (RT) reciprocal relationships. An associative relationship is made between terms that are conceptually related but are neither hierarchical or equivalence relationships in nature. Associative relationships serve to alert indexers and searchers that there are other terms in the thesaurus that may be of interest to them,
e.g.

photosynthesis	thylakoids
Related Terms: thylakoids	Related Terms: photosynthesis

The process of “photosynthesis" is a related concept to "thylakoids" because thylakoids are the site of photosynthesis.

d. *AND type USE/USE FOR relationship:*

"And type" cross references are used occasionally and are designated by USA (Use AND type) and UFA (Use For AND type), e.g.

alkaline water
Use AND type: alkalinity
Use AND type: water quality

which instructs the indexer to assign the terms "alkalinity" and "water quality" for the concept of "alkaline water". The reciprocal relationship, UFA appears at the terms "alkalinity" and "water quality" such as:

alkalinity	water quality
Use For AND type: alkaline water	Use For AND type: alkaline water

(Note, we are moving away from this relationship as SKOS can’t include it. We will be using specific descriptors, for example, “alkaline water” rather than assigning two descriptors “alkalinity” and “water quality”).

VI. Note Fields for Terms

a. *Scope Note (SN)*

A Scope Note serves to clarify the meaning and application of the term in relation to other terms in the NALT, e.g. anorexia

SN: Use for the uncontrolled lack or loss of the appetite for food; for the eating disorder characterized by the misperception of body image USE anorexia nervosa.
The first part of the Scope Note gives guidance on the concept that the term represents and the second part of the Scope Notes gives guidance on which term to use for a related concept. This format is generally followed for most Scope Notes, however, there are exceptions to this format.

b. *Definition field (DF)*

This field contains the descriptor's definition, which best represents the concept that is in the thesaurus.

1. Not all descriptors have definitions.
2. Hierarchical position aids users of the thesaurus in determining the meaning of the term, however, an ambiguous term (e.g., sediment pollution) requires a scope note or definition.
3. Federal government sources of definitions can be used verbatim.
4. Verbatim definitions from standard dictionaries and resources, (e.g., Dorland's, Stedman's, Academic Press Dictionary of Science and Technology), may not be used.
5. Definitions must be rewritten in order to be included in the thesaurus.
6. Definitions must be “politically neutral”, that is, without viewpoint or bias. Definitions should be factual and clearly stated as possible without erroneously focusing the concept to one discipline. Usually definitions should not include statements about scientific significance; however, some examples may be given in order to better understand the concept and its importance to agriculture.

Exceptions to these guidelines:

a. Some definitions for terms are hard to restate without using similar terms in standard dictionaries and references. If the term is commonly found having the same definition in many different resources, then it may be used verbatim without credit to source.

c. Journal articles that give a definition of a term can be used with the proper full bibliographic citation. A source that is used several times may be added to the bibliography that is available on the thesaurus and glossary web site.

d. *Definition Source (DS)*

1. It tells about the source of definition.
2. “NAL Thesaurus Staff” in DS field indicates that this was a definition written by staff.

VII. Source of Terms (SO)

The authorities and sources used for the various subject areas are listed in the About the NAL Agricultural Thesaurus at https://agclass.nal.usda.gov/bib.shtml Listed below are specific notes on sources of terms:

1. AGRICOLA and PubAg is the major resource for literary warrant. However, number of hits in "google scholar" is also checked. (Note: Taxonomic names are included on the bases of their importance also. For example, if a name has few hits and is medicinally or economically important, we take it. We also check if those hits are recent or old.
2. We follow "IUBMB Enzyme Nomenclature" for classifying enzymes.
3. For plants, GRIN (Germplasm Resource Information Network) is used for lower taxonomy (Genera and species etc.) and for upper taxonomy, we follow "Angiosperm Phylogeny Group".
4. For bacterial taxonomy, use "bacterio.net" also known as Bergey's Manual of Systematic Bacteriology.
5. For viruses, ICTV (International Classification and Taxonomy of Viruses)
6. For Fish, Fishbase and for Algae "Algaebase" is used.
7. Other taxonomic database, we use or consult are: ITIS (International Taxonomic Information System); NCBI;Mycology;Encyclopedia of Life; GBIF; Uniprot etc. (Note: Many times, a term is checked in multiple resources or even in recent publications to justify its inclusion as a preferred or non-preferred term). If a term is replaced from a descriptor to a non-descriptor, we add it into "List of replaced descriptors".

VIII. Form and Selection of Terms

1. Use noun and noun phrases when selecting terms. Use of adjectives alone is not allowed.
2. Use of long phrases that contain multiple concepts is not allowed, e.g. biological control of insect pests. Though it can be added as a hidden label for the purpose of indexing.
3. Pre-coordination (compound terms) will be warranted when the specificity is needed to describe a concept, e.g. insect anatomy. Pre-coordination is especially noteworthy when there are animal/human concepts or higher animal/arthropod concepts that need to be distinguished to facilitate retrieval.
4. Hyphenation
    a. Always check literary warrant as this should be the deciding factor when there are hyphenated and non-hyphenated spelling variants. If literary warrant is close to 50/50, use the non-hyphenated form as the descriptor and add the hyphenated form as a non-descriptor to the chosen term.

   b. Trend is to do away with hyphens, e.g. agribusiness, not agri-business, hyphenated terms can be added as hidden labels. Though in the past, some hyphenated terms have been added as non-descriptors.

5. Numbers and special characters are allowed, but use of special characters should be limited as much as possible. Greek letters are written, e.g. alpha, beta, and gamma. Many chemical names contain numbers, commas, hyphens, parentheses. The use of the + symbol should be avoided or must be limited to non-descriptors.
6. Drop the possessive from the term for diseases and disorders (e.g., Down syndrome, not Down's syndrome). This is following MeSH policy. The possessive form of terms is allowed under certain circumstances (e.g., farmers' attitudes). Users must be aware that punctuation is handled differently by the various database providers.
7. Capitalization follows Chicago Manual of Style. Technical terms follow accepted or proper usage of capitalization, e.g., pH. Capitalization is used for proper names as well as in other instances.

   Instances wherein capitalization will be used:

   - named geographic regions
   - disease names using proper names (e.g., Alzheimer disease)
   - named organizations
   - breed names
   - Taxonomic names
   - technical terms that require capitalization for the understanding of term meaning, e.g., pH, photosystem I, phospholipase A1, etc.
   - chemical symbols used in terms, e.g., NPK fertilizers

8. The MultiTes software controls alphabetization order. Spaces are recognized so that alphabetization is done by word, not by letter. For example, these terms are alphabetized in MultiTes in this order: plankton, plant anatomy, plant health, plant taxonomy, plant-insect relations, Plantago, planting, plants.
9. Use American forms of terms (American spelling and usage) for descriptors. Conside British forms as non-descriptors.
   Example of variant spelling: color, uf colour
   Example of variant usage: eggplants, uf aubergines.
10. Only use direct entry for descriptors. Indirect entry was stopped in 2017 and most of indirect entries (~2000) were removed as non-descriptors from the NALT. Example of direct entry: crop production
    Example of indirect entry: production, crop
11. Plural versus singular forms of terms:
    a. Use plural form of term for "count" nouns. (Count nouns are names of objects or concepts that are subject to the question “How many” but not “how much”).
    b. Use singular form of term for "abstract" nouns. (e.g., Emotions, activities, etc.).
    c. Consider adding the plural form as a Hidden label for singular form descriptors.

   d. Consider adding the singular form as a hidden label for plural form descriptors. Decision making on the addition of these types of non-descriptors rests on the added value obtained when the plural/singular form is used by searchers. If not particularly helpful for people, but of more value to machines, add as a Hidden Label.
   e. When plural and singular forms are irregular (not simply -s for plural) add as non-descriptors, e.g., mycelium/mycelia and hypha/hyphae.

12. Special note for Anatomical terms - MeSH uses singular for anatomical parts, but this is not compatible with CABT and AGROVOC. Use singular for body parts such as nose, liver, mouth (one organism usually has one) and plural for body parts such as eyes, ears, kidneys (one organism usually has more than one).
   a. Avoid abbreviations and acronyms unless it is highly accepted, (e.g., DNA, RNA). General Rule: Acronyms must be cross-references, not descriptors. However, we have exceptions to this rule for HACCP, USDA, USSR, pH, DNA and RNA).
   b. Full form is preferred in most cases.
   c. Capitalize acronyms when used.
   d. Limit punctuation by NOT inserting periods (.) between letters, (e.g., ELISA not E.L.I.S.A.)
   e. An acronym can be used as a descriptor when it has a parenthetical qualifier, e.g., NADPH (coenzyme)
   f. If an acronym is used in a cross-reference, it must be qualified with a parenthetical qualifier, e.g. EEG (electroencephalography) USE electroencephalography; SIT (sterile insect technique) USE sterile insect technique; PAGE (electrophoresis) USE polyacrylamide gel electrophoresis.

13. For pesticides and other chemicals, we prefer the common chemical name, e.g. alachlor, benomyl, carbaryl. References used for checking common chemical name include:
   a. PubChem/MESH and CheBi
   b. the PAN Pesticides Database (http://www.pesticideinfo.org/Index.html)
   c. Alan Wood's pesticide database http://www.alanwood.net/pesticides), which tells if name is approved by WSSA or other organization; gives IUPAC name, gives CAS name, ISO standard name, and non-ISO common name. They don't always agree on one common name, so sometimes you have to base your decision on database hits, your experience or the expertise of others. We try not to use the full chemical name, but sometimes that is all you have. We were surprised to learn that the full chemical name received more hits in the literature than the acronym, TCDD. Best advice is to use all resources available (human and information) in order to make a decision.
   d. If there is no common chemical name and an acronym is accepted by an organization (such as WSSA – Weed Science Society of America) then we use the acronym with a parenthetical qualifier, such as MSMA (herbicide), CDEA (herbicide).
   e. When you suspect that a chemical will have many uses, use (pesticide) as the parenthetical qualifier, such as DNOC (pesticide), DDT (pesticide).

(Note: usually herbicides do not have many uses, but any insecticide may also be acaricides, rodenticides, molluscicides, etc., so most of the time assign (pesticide) unless it is a herbicide.)

14. For agency names in USDA, use the full name of the agency for the preferred term. Use the acronym with parenthetical qualifier for cross references, e.g., FSIS (USDA) USE Food Safety and Inspection Service.
15. For organization names, use the full name of the organization for the preferred term. Use the acronym with the full name in the parenthetical qualifier, e.g. WHO (World Health Organization) USE World Health Organization.
16. If a term is a combination of acronym and other words AND it is not easily confused, this is acceptable form for a descriptor (or non-descriptor), e.g., PSE meat ; AGNPS model; GLEAMS model; SOS induction .
17. For US acronym standing for "United States" use U.S. to be consistent with other entries. Check to see what others have done in the file.
18. Literary warrant must be established for terms. Databases that can be searched include AGRICOLA, PubAg, Google Scholar, etc. but other information systems in the specific subject area are acceptable.

IX. Disambiguation of Homographs

Homographs are terms which are spelled the same but represent two or more different concepts. It is desirable to use the spelling of the term which is the most desirable to the users of the thesaurus and substantiated by literary warrant; however, it is also necessary to abide by the basic principle that each term in a thesaurus must represent one concept. The meaning of homographs is clarified for users by the use of a parenthetical qualifier that follows the term, e.g.

Togo (Africa)
Togo (Heteroptera)

This practice is in accordance with the NISO Z39.19 standard for Guidelines for the Construction, Format, and Management of Monolingual Thesauri.

X. Construction of relationships between terms

1. OR type Use/Use For is not allowed.
2. NISO thesaurus standard (Z39.19) is the general guideline for thesaurus construction; however, it is paramount to consider the indexer and searcher of the databases when constructing hierarchies, choosing terms and building relationships. Review the purposes of the thesaurus at the beginning of this document. The inclusive search feature and look up features of the categories must be considered when building hierarchies.
3. In the case of multiple uses of an organism, its "always true" hierarchy shall be included, e.g., rabbits in the "animals" hierarchy. Use the RT relationship to relate organisms to their "uses" or "roles," e.g., rabbits; RT meat animals, RT fur-bearing animals, RT pets, RT laboratory animals, RT vertebrate pests.
4. In the case of multiple uses of a chemical, its "always true" chemical hierarchy shall be included. However, chemicals may be included in a "use" or "role"

hierarchy as well, e.g. lindane; BT acaricides; BT chlorinated compounds.

5. For domestic animals, the common name is the preferred term, e.g., dogs, UF Canis familiaris. For plants, the scientific name is the preferred term, e.g., Triticum aestivum. Plant products, such as wheat, will be placed in plant product hierarchy with RT to the scientific name of the plant, e.g. wheat, RT Triticum aestivum. Taxonomic classifications will only include scientific names.
6. Polyhierarchy, when more than one BT is "always true," is allowed. Note the exception made for chemicals in #4 above.
7. Consider when a term concept belongs under two subject categories, such in the case of "forest policy." See discussion in section "Upper Structure and Categories in the Thesaurus."
8. Node labels or facet indicators may not have related terms. (Since node labels were eliminated, this is an old guideline).

XI. Specific Guiding Principles for Specific Subject Areas

1. *Geographical Terms*

a. Divisions are made on physical location rather than political affiliation.
b. U.S. Africa, Europe, and Asia are divided by compass direction and/or "geographically defined areas" such as Pacific States, Midwestern United States, Southern Europe, Sub-Saharan Africa
c. Regions used to express areas of economic, cultural, agricultural, etc. homogeneity fall under <named regions> and have "RT" relationships to the appropriate states, countries, etc. Food example, Corn Belt, Balkans, Great Lakes.
d. "St" and "St." are spelled out as "Saint"
e. Sources of terms: MeSH, Webster's New Geographical Dictionary and CIA Website

2. *Plant Diseases*

a. Specific plant disease names of economic importance are added as descriptors, e.g. Chestnut blight.
b. Causal agents will be RT relationship to the plant disease name, e.g. chestnut blight, RT Cryphonectria parasitica.
c. Authority for disease names and causal agents is: American Phytopathological Society "Common Names of Plant Diseases" located at: https://www.apsnet.org/publications/commonnames/Pages/default.aspx
d. Care should be taken when using this information, as with any authority, as the fungal names and forms (anamorph and teleomorph) are not always current with fungal nomenclature.
e. Specific plant disease names are classified according to "by etiology." If the disease is known to infect a specific part or host, polyhierarchy under "by part" or "by host" is acceptable. However, one must make sure that the relationship is always true to make these BT/NT relationships.

f. Host names should be added as RT relationship in the case where a disease is specific to a specific crop, e.g. chestnut blight, RT Castanea.
g. A "plant diseases with unknown etiology" is added to Common Names of Plant Diseases.

3. *Fungal nomenclature*

a. In 2009 edition, the higher-level classification of the Fungi was changed in accordance with Hibbett, et al. 2007. A higher-level phylogenetic classification of the Fungi. Mycological Research 111:509-547.
b. In the past, anamorphs and teleomorphs were handled as following:
   a) Teleomorphs (sexual) are the preferred descriptor.
   b) Anamorphs (asexual) are non-descriptors with a notation in the Scope Note field that reads "anamorph." Example:
   Bipolaris oryzae, USE Cochliobolus myiabeanus SN: anamorph
c. If a non-descriptor for a fungal name did not have the scope note, it was assumed a synonymous name.
d. After, Nov. 2013, the scientific literature is the most definitive source of fungal nomenclature. Other fungal databases consulted for fungal nomenclature are, Mycobank, NCBI, UniProt, etc.

XII. Proposals for the NALT

Proposal workflow (established 2013):
Access database for proposals.

1. Indexers or stakeholders submit new terms or propose changes to the thesaurus using an excel “proposal database”. Indexers provide justification for change. Stakeholders also propose terms for NALT through e-mails to NALT coordinator.
2. Proposals are assigned to thesaurus coordinator by default, however, certain items that do not require research (or limited research) can be given to Technician for input into Multites working file.
3. Thesaurus coordinator researches term and inputs changes in MultiTes or assigns to Technician for input.
4. Those items that are difficult are researched using Thesaurus Research Form (as used previously), research is distributed and discussed with larger thesaurus team for resolution. Difficult terms are discussed at weekly meeting or can be discussed at broader meeting with additional subject experts, if needed.
5. When items are completed, they are marked completed in the database.
6. Replaced descriptors (English as well as Spanish), are noted in the “Replaced Descriptors” file in the S:\NAL\TSD\THES\20XX file folder.

Additional documents:

a. Multilingual Thesaurus Construction Principles

See separate documents "Multilingual Thesaurus Construction Principles"

b. Monthly Data Integrity Checklist

See separate documentation “Monthly Data Integrity Checks”

c. Publishing a new edition of thesaurus

See separate documentation “Procedures for publishing a new edition of thesaurus and glossary”