Indexing Guidelines for Automatic Indexing – BASICS for the beginning – DRAFT for leading discussion at the Wednesday, May 4, 2011, Indexer Meeting.
May 2, 2011

Mindset: Change art of indexing to science of indexing where we have predictability and enforcement in the quality of indexing.

Mantra: "An indexing guideline that is not configurable or enforceable or controllable, is not a indexing guideline".

Work plan: Indexing Guidelines can be broken down into three sections. "A" is receiving work now, but will be further vetted during the Acceptance Testing period. "B" will be a focus after we go to production. "C" needs to be done now while we are developing the golden set for the Acceptance test.

A. Configurations of AI software
    a. Max / Min # of terms assigned
    b. Numerical rank
    c. Weighting of terms
        - Foods and products
        - Organism names – scientific and common
        - Organizations
    d. Stop word list
    e. Hierarchical specificity
    f. Use of cross references
    g. Assignment of NALT terms / Suggestion of terms not in NALT but are given a high rank by the software (assuming that higher rank is going to give higher importance of that subject to the searcher of documents).
B. Subject indexing guidelines for specific topics
    a. Select a few journals in one topic (thinking we want to keep these topics small, but need to experiment)
    b. Select an indexer that is knowledgeable about indexing in that topic
    c. Indexer writes down the conventions and patterns in indexing practice
        - E.g., Assignment of role terms
        - E.g., disease and causal agent
    d. Vet these conventions in the wider indexer audience.
    e. Consider the creation of knowledge bases, cross references in thesaurus, or rules that can enforce these rules.
C. General indexing guidelines
    a. Geographics
    b. Methodology
    c. Organisms/product/common name
    d. Vet these conventions in the wider indexer audience.

e. Consider the creation of knowledge bases, cross references in thesaurus, or rules that can drive indexing consistency and accuracy.

Resources:

1) Should look at the Table of Contents and Index for the NLM indexing guidelines for ideas. Can we devise a checklist/form of facets, somewhat like NLM does? (Clinical, gender, age, host, causal agent, physiological function, anatomical part, agricultural practice…) This may be handy for some specific subject areas..thinking Nutrition may lend itself well to this. May be able to use some of this for general guidelines as well.
2) Ray's email of Friday, March 25, subject: Revision of Guide to Subject Indexing.
3) S:drive\nal\tsd\indexing\guidelines, policies and procedures\Guide to Subject Indexing (revised 2010).doc. See next section for Summaries that will drive our early discussions.
4) Other?

Summaries for guidelines on the assignment of descriptors (650), geographical descriptors (651) and identifiers (653) as taken from the 2010 Indexing Guidelines document:

Assignment of descriptors - Summary

1) Indexers will only use *NALT* descriptors in field 650.

Comment: AI will assign NALT terms (but we will consider that highly ranked, non-NALT terms could be used in the identifier or cause some other action.)

2) Indexers will assign *NALT* descriptors in a manner that is consistent with their broader terms (BT).

Comment: AI will assign NALT terms to represent the concept in the thesaurus. For terms that may be misunderstood, we will write rules to disambiguate. If rules cannot be written then the term must be re-evaluated or we will flag for human intervention.

3) "Use" instructions, definitions, and scope notes in the *NALT* will be strictly observed.

Comment: AI will easily use the cross references and will do it consistently and without fail. Once again, terms that may be misapplied will need rules to disambiguate. If rules cannot be written then the term must be re-evaluated or we will flag for human intervention.

4) Concepts will be expressed using *NALT* descriptors at the most specific level of specificity used by the author of the document at hand. If, however, an indexer is using more than three narrower terms (NT) in the same hierarchy at the same level, he/she may choose to use the broader term (BT) to represent this concept.

Comment: Configure AI to use the most specific term in the hierarchy. We need a discuss ion with management on the use of upposting, which would be done programmatically OR use of query explode for searchers.

5) Two descriptors may not be used to represent a single concept when a precoordinated descriptor for the concept exists in the *NALT*.

Comment: Keep this. Will be a basic guideline to implement in the rule making process.

6) In circumstances where a) concepts are not represented at the required level of specificity, b) concepts are not covered or covered adequately in the *NALT*, or c) the *NALT* word block erroneously restricts the use of one or more terms, indexers will use guidelines A through C in Section 4.2 - Assignment of MARC tag 653 identifiers to assign terms.

Comment: Question was raised at last meeting…will we use 653 in an AI environment? I believe that if there is a noun or noun phrase that is not an NALT term but achieves a high rank (through frequency or presence in a Skill Cartridge – Biological entities or chemicals SC), then these could be assigned to 653. Thesaurus staff can use these as suggestions for additions to the thesaurus as descriptors or non-descriptors.

7) Indexers will assign taxonomic and common names of organisms and products according to Use of Taxonomic Names and in a manner that is consistent with all other standards in Section 4.1.

Comment: Discuss.

8) Indexers will assign descriptors for diseases (human, animal and plant) and experimental animal models according to the sections on Use of descriptors for diseases and experimental animal models and in a manner that is consistent with all other standards in Section 4.1.

Comment: "animal models" is a complex concept and not sure if we can consistently apply a rule. However, I believe that there are certain journals in which animal models are frequently seen. In those instances we may be able to write a rule that can be applied consistently, and with good results.

9) Indexers should examine word blocks carefully for each descriptor considered for assignment.

Comment: Indexers must use the

10) Indexers may use the broader term (BT) of any assigned descriptors if they feel the document at hand should be retrieved on this broader concept.

Comment: I feel that we need to limit to most specific as the future upposting OR term explode will negate the need for this. If the BT is a "discipline", then I would not. If the BT is more of a complex concept term, such as plant nutrition, then I would consider valuable.

11) A descriptor and its broader term (BT) may be used if the concepts are addressed in the article at both levels.

Comment: Keep to the specific for reasons given earlier.

12) It may be necessary to use two descriptors to represent a single concept.

Comment: Rules will be written to uphold this guideline. This is the example of "red meat quality" USA "red meat" And "meat quality".

13) The “Indexing Selection Scope and Coverage Guidelines”, Supplement B, can help indexers assign terms to convey the relevance of a document to AGRICOLA.

Comment: Supplement B is about selection of articles. A selection program is being written to select articles from a journal. This guideline will be used for that program.

14) The order of descriptors is not critical to information retrieval; however, indexers may choose to consider proximity and order when arranging descriptors.

Comment: Order of descriptors is defunct. We are in an online world – it does not matter.

15) Consult the facet “(publications by type).” Use these *NALT* terms when the document is one of these types.

Comment: Mixing genre and subject. This has never been a good idea but was forced upon us. Need to change how this is handled in the AGRICOLA record. A quality issue.

16) Indexers may use as many descriptors as necessary to convey all concepts identified for indexing. However, when the number of descriptors for one document exceeds 20, indexers are encouraged to re-evaluate their choices. Subsuming some narrower terms (NT) under their broader terms (BT) may provide AGRICOLA users with a more meaningful record.

Comment: Longer abstracts will usually inspire more terms than shorter abstracts.

17) Problems with the *NALT*, including missing terminology, restrictive hierarchies, definitions and scope notes, or other problems, must be brought to the attention of the NAL Thesaurus staff.

Assignment of geographical descriptors - Summary

1) All of the instructions provided in Sections 4.1.1.a through 4.1.1.c are to be used by indexers when applying geographical descriptors.

Comment: General considerations, specificity and exhaustivity.

2) Geographical terms may not be assigned solely based upon the appearance of geographical terms in the nomenclature of strains of bacteria and other organisms**.**

Comment: This may be hard to get AI to do.

3) For articles in which subjects (animals, plants, humans) originating from one country/area are studied in a country to which they have migrated or been imported, apply the geographical descriptor for the country in which the subjects were studied and indicate their geographical or ethnic origin in some other way. This applies only in the case of living subjects studied in a natural environment.

Comment: This may be hard to get AI to do.

4) Geographical terms may not be assigned based solely upon the affiliation of the author(s).

Comment: AI will probably not do this unless it is mentioned in the abstract.

5) Generally, if the work is conducted in a laboratory or a greenhouse, use of a geographical descriptor is not warranted.

Comment: Possibly a rule can be written, but probably not.

6) The geographical origin of a strain or line may sometimes be of significance in a way that warrants use of a geographical descriptor even if the work is done in vitro.

Comment: When provenance is used as a descriptor, we may allow geographical term.

7) There is no limit to the number of geographical descriptors that may be used for a given article. However, in most cases where more than five or six are applicable, it will probably be more useful to apply terms for broader geographical units that subsume the others

Comment: AI will assign all. System may uppost to BT or we will use term explode for retrieval.

8) Problems with the *NALT*, including missing terminology, restrictive hierarchies, definitions and scope notes, or other problems, must be brought to the attention of the NAL Thesaurus staff.

Assignment of identifiers - Summary

1) Abbreviations, acronyms, initialisms, etc. will be spelled out in the 653 field when the appropriate term is not available as a *NALT* descriptor.

Comment: Do not use acronyms, abbreviations and initialisms in the 653. Keep this rule.

2) Indexers may not use descriptors which appear in the *NALT* as identifiers, even if the thesaurus descriptor represents a concept other than the concept associated with the term in the document at hand.

Comment: Keep, easy.

3) Indexers may not use as identifiers terms which appear in the *NALT* as nondescriptors.

Comment: Keep, easy.

4) Indexers should examine the *NALT* thoroughly before using the 653 field.

Comment: Keep, easy. But AI may put synonyms in the 653. Synonyms would be removed by thesaurus staff.

5) Problems with the *NALT*, including missing terminology, restrictive hierarchies, definitions and scope notes, or other problems must be brought to the attention of the NAL Thesaurus staff.

Acceptance Testing Indexing Guidelines:

Procedure:

You will be given your list of bib id numbers that you are to edit. Only 128 of the 5000 are missing indexing and will require original indexing.

Search each bib id number in Citation server and make sure you are at the correct item.

Read the title and abstract.

Edit the 650s, 651s and 653s in the record.

Save the record.

Go to next item, repeat.

Checklist of Indexing:

WHO - the players, such as organisms, products, organizations, social groups, communities, cultivars, hybrids, genes. Other agents, such as chemicals, pesticides and drugs. Roles of the who, when significant.

WHAT - actions such as treatments, agricultural practices, and phenomenon - breeding, genetic transformation, brewing

PROPERTIES – acidity,

RESULTS or EFFECT - weed control, pest control

WHEN - time of year, preharvest treatment

WHERE - geographical 651

HOW - methodology - only index when subject of investigation.

What types of changes to make:

1) Add terms that need to be added to convey major points of the article.

a) Add complex concept terms that are missing, e.g. bacterial diseases of plants, fire ecology, weed biology, crop weed competition.

b) Add geographics if they are missing and significant

c) Add role terms if they are clearly a focus of the article

d) Add missing major concept terms.

2) Delete terms that are

a) Delete "stop words" e.g., duration, research, diameter, measurements

b) Delete role terms that are extrapolated.

c) Delete methodology that is not the subject of investigation.

d) Delete terms that are used outside the term meaning (misapplied terms)

e) Delete "tertiary" terms that are very minor in use in the article.