Change to Automatic Indexing (AI):
Subject Analysis Quality and the Development of Draft Guidelines

June 28, 2011 draft
July 18, 2011 last revision

Background
In 2011, the indexing process at NAL is drastically changing in order to better serve customers. In order to meet this goal, the indexing procedure is changing from a manual process to an automatic process. This reason for this change is many-fold:

a) To increase the number of indexed articles in AGRICOLA from 75,000/year to 500,000/year so the database content is relevant and up-to-date for users.
b) To ensure that all articles in AGRICOLA are indexed with a controlled vocabulary.
c) To ensure that all articles in AGRICOLA have been subjected to a quality assurance procedure.

This document is to establish "how" we index rather than "what" we index. The selection of journals and articles is done prior to the automatic indexing procedure and the policy governing that procedure is covered elsewhere. Even though AI will most likely be applied to other than AGRICOLA articles, the indexing of scholarly literature in agriculture is our emphasis at this time.

Mindset/Mantra: Change "art of indexing" to the "science of indexing", where we have predictability and enforcement in the quality of indexing. "An indexing guideline that is not configurable or enforceable is not an indexing guideline". A guideline or policy is:

- UNDERSTANDABLE (which means that it is easily understood by non-experts and the language is clear),
- EXECUTABLE (which means that the policy is not subject to individual interpretation), and is
- ENFORCEABLE (which means that we can implement a quality assurance procedure).

The Indexing Process
The purpose of indexing is to use terms to express the subject matter of items so that users can use these terms to locate items of interest in an information storage and retrieval system. Indexing is traditionally seen as a two step-process : 1) determine the subject matter of documents and 2) translate the subject matter into the terms of the controlled vocabulary. Indexing guidelines usually spend much time on the second step as these are mechanical and are easier to express (e.g., use the most specific term in the hierarchy). It is this first step, or the cognitive process of determining the "aboutness" of an item, that is usually vague and inadequate in indexing guidelines.

Previous NAL subject indexing guidelines
A 2010 version of the AGRICOLA indexing guidelines is available on the shared drive at s:\nal\tsd\indexing\guidelines and policies. However, much of this document is not useful for AI as it contains many "directions for humans":

- "consider"
- "carefully select"
- "select only information which warrants retrieval"
- "treat equally"
- "do not select concepts which are present only through implication or speculation"
- "do not select methodology or technique routinely"

- “represent the aboutness”

We need to realign our thinking to describing the indexing process in a MECHANIZED way so that it can be translated to computer commands. Are there patterns in the text or in our knowledge that inspire us to assign a particular thesaurus term?

Document-oriented indexing versus User-oriented indexing

Most books on indexing practice center on the “document-oriented” or “document-centered” approach. This approach focuses on the strict adherence and faithful description of the text at hand. The subject analysis is solely based on the item itself and the indexer is not to add viewpoint to the item. In “User-oriented” or “User-centered” indexing, the indexer considers what questions the user may ask to retrieve the item at hand and considers which terms to describe the item based on the user’s perceived viewpoint. Both of these approaches to indexing were used in the previous indexing guidelines for AGRICOLA. As we move to automatic indexing, both of these indexing approaches still apply. Even though the AI software only has the author’s text to process, our indexing rules can still take into account context and the perceived needs of users in that knowledge domain.

The NAL indexer has not been traditionally given data about users and their queries; however, the indexer has been trained to think of the user. One advantage that NAL indexers have is that they are professionally trained and usually hold an advanced degree in some subject domain of agriculture. Indexers usually index in an area in which they are familiar, but this is not always the case. An indexer indexing outside of their knowledge domain can still index the item, but it usually means that the text present in the document drives the index term selection.

Context-oriented and domain-oriented indexing

Indexers draw on their education and experience to understand the “aboutness” and context of the text. Why is kudzu an important topic to agriculture? Kudzu is a vicious weed that is a threat to agricultural crops and an invasive species that can choke out native plants. Kudzu is a plant, but what role is this plant in the context of agriculture? Why do we care about kudzu? Even if the text is simply a description of the biology of the kudzu plant, should the term “invasive species” be assigned? Should the term “weed biology” be assigned? The indexer has been trained to consider the user and ask the question for decision making: “Will the user be delighted or disappointed if they retrieved this item when searching on ‘invasive species’ or ‘weed biology’?” With human indexing, humans will differ in how they index this item. Inter-indexer and intra-indexer variation in the application of index terms is an accepted byproduct of human indexing. Some of this variation is inherent in the fact of being human. The selection of index terms relies on the Indexer’s knowledge of the subject as well as how they were trained to index that scenario. If the indexing guidelines are general or vague, the guidelines will fail to be interpreted consistently by indexers and trainers. Such inconsistencies would not exist in automatic indexing if a rule to always index “weeds” when “kudzu” is present. But beware Will Robinson! If this is a strict relationship, (“always true”), then this would be present in the thesaurus hierarchical relationship. When it is not present this way, then we can assume that it is not always true. Kudzu is always a plant, but is not in all situations an invasive species. If that relationship was always true, then kudzu would be a narrower term in the thesaurus. When developing automatic indexing rules, caution needs to be taken to not add information that is extraneous or implying a specific role that would not be true in all situations.

Context is also interpreted not only by the text in the item being indexed, but by the source journal for that article. If one is indexing about fish feeding, does one automatically assign “aquaculture” if it appears in the “Aquaculture” journal? Is this indexing for implication or is this simply providing context for the user? How often is this type of “general term” assigned by indexers?   A study of the indexing terms for specific journals and the typical articles contained in journals could support our development of mechanized indexing rules.

With automatic indexing, we can codify the knowledge of an educated indexer considering the user, and this knowledge can be incorporated into the AI system. Automatic indexing allows us to explicitly state our indexing practice for each scenario, to codify it in rules and have this request performed consistently by the AI software.

Patterns in Journals – “Journal Fingerprinting”

As experienced indexers, we know there are patterns in journals.   We know that there are some journals where we use the same category codes over and over. In some journals, we find ourselves repeatedly using the same NALT terms. Can we somehow find a way to “characterize” each journal? Is there a unique “fingerprint” for a journal and can we use this fingerprint to our advantage when it comes to automatic indexing?

See separate document “Journal Fingerprinting for Automatic Indexing” for details on how we can discover the patterns of content within the 95 journals and systematically characterize them so that we can better devise strategies to index them.

Automatic Indexing Guidelines, what we have as of July 18, 2011 (prior to work done on Journal fingerprinting)

Some assumptions/decisions that influence "how we index":

1. We will not be implementing up-posting in the AGRICOLA database.
2. We will be implementing hierarchy expansion in search in AGRICOLA. (This influences our specificity policy).
3. We will be implementing automatic inclusion of synonym rings for search.
4. We will no longer assign category codes; however, we may find another way to classify and may use this for refining automatic indexing or as a classification to present to the user. (This is separate document that is being developed).
5. The order of descriptors is not important.
6. We will no longer assign genre terms, such as "literature reviews" and "bibliographies". In fact, most of these terms under "publications" in the thesaurus will be removed from NALT.
7. Our old indexing rule: Do not apply geographics in certain situations (e.g., do not apply for greenhouse grown crops). This is so subjective that it will be hard to write a rule for this. Also, it is unknown how stringent human indexers applied this rule in practice.
8. We will continue to use NALT:
    a. Topical NALT terms will appear in the 650
    b. Geographical NALT terms will be used in the 651
    c. Non-descriptors will not appear in the record, but will be "translated" to their NALT descriptor and this term shall appear in the 650.
    d. We will continue to use the 653 for non-NALT terms, (e.g., such as highly ranking terms from skill cartridges).
    e. We would like to continue the policy of NOT using abbreviations, acronyms and initialisms in the 653, if possible. (Perhaps these can be captured some other way and translated to their complete term).

Automatic Indexing guidelines can be organized into three main sections:

1) Configuration of the AI software
2) Subject-specific guidelines
3) General Guidelines for all subject areas

Configurations of AI software. For those requiring actual numerical values, the actual numerical values are not given. These starting values will be determined with the aid of the vendor during the NAL scenario analysis.

a. Max / Min # of terms assigned
b. Numerical rank, minimum value
c. Weighting of terms by position in text (e.g. title)
d. Weighting of terms by subject

- More weight to:
    1) Foods and products
    2) Organism names – scientific and common
    3) Organizations (?)
    4) NALT descriptors and non-descriptors
- Less weight to:
    1) Methodology terms
    2) Chemical terms not in title (?)
    3) (Are there other interdisciplinary terms we want to give less weight to?)

e. Stop word list – need to see vendor stop list as well as supply our stop list on s: drive.

f. Hierarchical specificity – For now we are assigning most specific in hierarchy with the argument that a user's search on the broader term will retrieve the more specific in future AGRICOLA when the thesaurus is integrated into search.

g. Use of cross references in NALT to generate NALT term assignment
    - USE/UF
    - USE "and type"
    - Hidden labels
    - Typographical errors mapped to correct spellings, e.g. broccoli.

h. Assignment of NALT terms

i. Assignment of relevant terms not in NALT but are given a high rank by the software (such as that which is assigned by the Luxid Relevance Finder Skill Cartridge). Assuming that higher rank is going to give higher importance of that subject to the searcher of documents. Assess whether these are significant and if we want them added to MARC 653.

Subject indexing guidelines for specific topics

j. Select a few journals in one topic (see results from Journal Fingerprinting…journal "clusters". Thinking we want to keep these topics small, but need to experiment.

k. Select a group of indexers that are knowledgeable about indexing of that topic.

l. Indexers to review any relevant data from the Journal Fingerprinting idea.

m. Indexer writes down the conventions and patterns in indexing practice
    - E.g., Assignment of role terms
    - E.g., disease and causal agent

n. Vet these conventions in the wider indexer audience.

o. Consider the creation of knowledge bases, cross references in thesaurus, or rules that can enforce these rules.

General indexing guidelines

General rule-making guidelines:

1. Indexers will write rules for *NALT* descriptors in a manner that is consistent with their broader terms (BT).
2. Indexers will write rules for NALT descriptors in a manner that is consistent with their scope notes (SN), if applicable.
3. Indexers will use a pre-coordinated descriptor for a concept in lieu of using two or more descriptors.
4. For “role” type terms, Indexers will use caution to write rules that are consistent with an “always true” relationship. Role term rules should be researched extensively and discussed with all indexers to ensure conformation to the always true relationship.
5. In the case where a thesaurus entry can be made, use this option over creating a rule. For example, for the text of “red meat quality”, add cross reference to the thesaurus: USE “red meat” AND “meat quality”.
6. If the meaning of an NALT term is unclear or too restrictive, please bring these to the attention of the thesaurus group.
7. Do not write a rule for a “stop word”…that is, a term that is “tertiary” and not likely to be assigned consistently by all indexers and is on the stop word list. If you have a candidate term for adding to this list, send it forward.

These were the general guidelines given when indexing/editing the 5K acceptance test articles

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.

f) Delete terms where you do not see any evidence for them in the title and abstract.

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 the role is significant, such as “invasive species” or “plant parasitic nematodes”

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

PROPERTIES – acidity, disease resistance

RESULTS or EFFECT - weed control, pest control, fire scars

WHEN - time of year, preharvest treatment

WHERE – geographical terms are indicated in thesaurus and placed in 651

HOW - methodology - only index when subject of investigation is the methodology, not mere mention.

SPECIFICITY – index to the most specific. Ignore the “rule of 3”.

EXHAUSTIVITY – Index what you see in the title and abstract.