Information Management Resource Kit

Module on Management of Electronic Documents

UNIT 3. METADATA STANDARDS AND SUBJECT INDEXING

LESSON 5. STEPS FOR SUBJECT INDEXING

NOTE

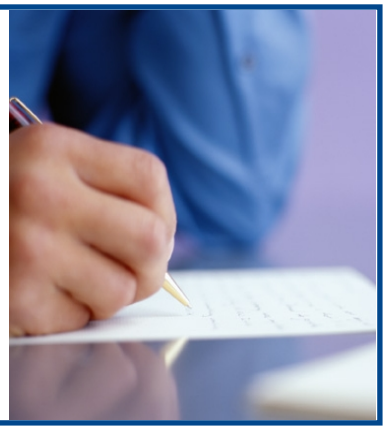
Please note that this PDF version does not have the interactive features offered through the IMARK courseware such as exercises with feedback, pop-ups, animations etc.

We recommend that you take the lesson using the interactive courseware environment, and use the PDF version for printing the lesson and to use as a reference after you have completed the course.

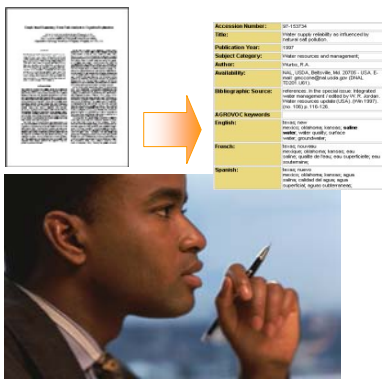
Objectives

At the end of this lesson, you will be able to:

- analyze **the topics** of a document for indexing
- identify **the key concepts** of the document, and
- apply the principles of **exhaustivity and specificity** when selecting the indexing terms.



Introduction



The task of the subject indexer is to assign to a document appropriate subject terms from the thesaurus, following a consistent level of **exhaustivity** and **specificity**.

This task requires an analytical effort and a thorough understanding of each document to be indexed.

There are some basic principles that can guide you when performing this task.

How to proceed

Essentially, indexing is a task based on these four steps:

1. UNDERSTAND THE CONCEPTS IN THE DOCUMENT

2. ANALYSE THE TOPICS IN THE DOCUMENT

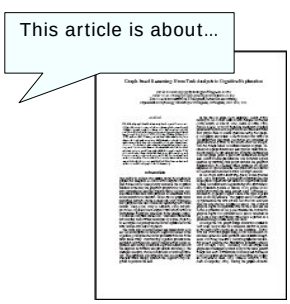
3. CHOOSE THE KEY CONCEPTS

4. CHOOSE THE CORRECT INDEXING TERMS

Now, let's see in detail how these tasks are performed…

How to proceed

1. UNDERSTAND THE CONCEPTS IN THE DOCUMENT



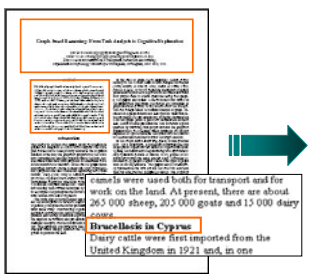
The objective of the first step is to get a **broad understanding** of the information contained in the document.

Indexers do this by concentrating on certain parts of a document, and glancing over the rest.

There is **no need to read the entire document.** In the rare cases when there is no abstract, table of contents, or chapter headings, it will be necessary to read a significant amount of the document to obtain an adequate understanding of the subject matter, but even then, **one should only glance over the text**.

How to proceed

2. ANALYSE THE TOPICS IN THE DOCUMENT



The important information can be gained from reading certain information, particularly in the following areas:

- title
- abstract or summary
- table of contents
- preface, introduction, etc.
- first paragraphs
- illustrative material and its captions
- words or groups of words that are underlined or printed in an unusual typeface
- concluding remarks
- index

How to proceed

Note down all the **relevant concepts.**

To identify all the relevant concepts, the indexer should keep the following **subject areas** in mind:

WHAT? What are some of the important topics mentioned in the document? E.g. specific plants or animals, products, etc.

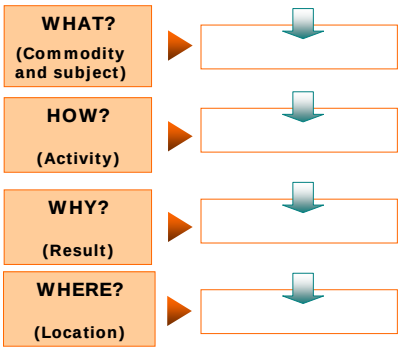
HOW? Are there specific activities taking place? E.g. buying or selling, processing food.

WHY? What were the basic reasons for the research? What were the results?

WHERE? Are any specific locations mentioned in the document?

How to proceed

3. CHOOSE THE KEY CONCEPTS



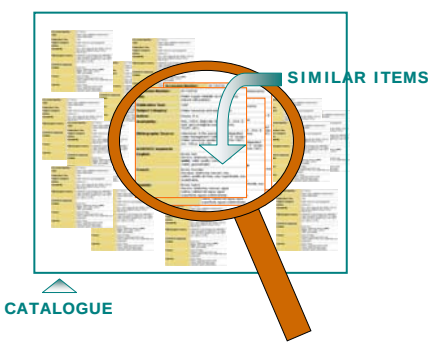
Documents will often have more than one concept for each of these subject areas, while others may not deal with all subject areas.

At this stage, the task of the indexer is to ensure completeness (**exhaustivity**) so that no important concept is left out.

At the same time, be careful of over-indexing. Keep in mind that the use of any term should mean that a user will find **substantial information** when they search under that term.
For example, if there are only two sentences on a trout in a 20 page document, this is not substantial information.

How to proceed

4. CHOOSE THE CORRECT INDEXING TERMS

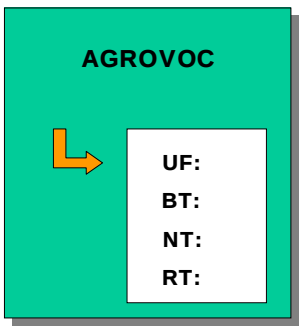


Once you have a clear and broad understanding of the concepts in the document, you can begin to **search for similar items** in the catalogue.

There are a number of ways of doing this.

The best way is to do **freetext keyword searches** of the concepts you have noted down, look for other records that appear to match your concept, and see what subject terms other indexers have used to describe each of your concepts.
In this way, you will often find terms that would never have occurred to you.
This is how indexers seek to achieve the same level of consistency.

How to proceed



Next, you can look up your terms in the **thesaurus** and see if there are any additional terms, or if any synonyms are mentioned.

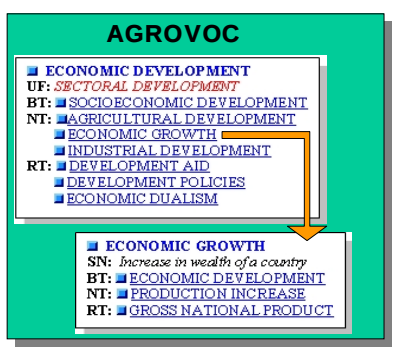
We shall use an example from the **AGROVOC** thesaurus.
By using the existing indexing records in conjunction with AGROVOC, where there are notes like **Use For**, **Broader**, **Narrower** and **Related Terms**, the indexer finds the correct level of **specificity**.

How to proceed

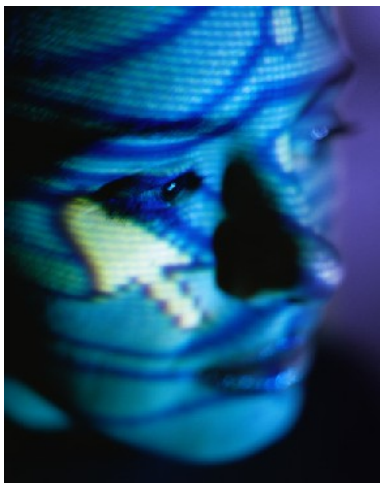
Also, the **precise meaning** of a term may be unclear, or you may find that it means something quite different in the thesaurus.

For example, in AGROVOC thesaurus, there are two terms: ECONOMIC DEVELOPMENT and ECONOMIC GROWTH. The differences may not be apparent until you look **into the thesaurus**.

We then discover that ECONOMIC GROWTH is actually a **Narrower Term** of ECONOMIC DEVELOPMENT, and we learn that it deals more with generating wealth, than the more general term of ECONOMIC DEVELOPMENT.



How to proceed

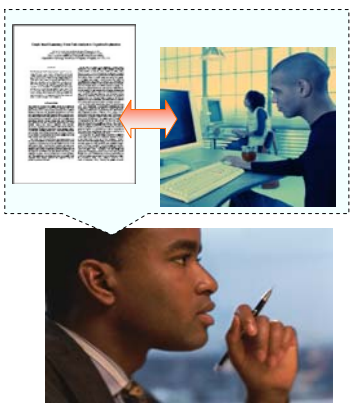


If all else fails, and you cannot find anything similar, you must be **creative**.

You can look up items by **the same author**, since people often write about the same things, or associated with the same project.

If you are positive that there is **no subject term** for your concept, you may wish to use a **more general term**, or to consider **proposing** it as an **additional** term in the thesaurus.

How to proceed



Finally, reexamine the subjects one last time and decide:

• if the terms you have chosen **describe** the document faithfully, and

• if the terms will **allow people** interested in the subject area to find the information in this document.

If we continue to keep in mind the primary goal of keeping **similar things together**, the entire process will become clearer. Precisely how this is done depends on the thesaurus chosen for the catalogue.

Example

Now, let's have a look at an example of subject indexing.

1. UNDERSTAND THE CONCEPTS IN THE DOCUMENT

Our first objective is the **broad understanding** of the information contained in the document.

Click on **DOCUMENT** to view the document.

It would be advisable to print the document and underline the relevant concepts.

Remember: you just need to read the title and the summary and quickly browse the text.

DOCUMENT

When you have finished, click on **HIGHLIGHTED DOCUMENT** to view which terms I have highlighted.

Do they match yours?

HIGHLIGHTED DOCUMENT

Example

Reading the title and the summary, and scanning the text, we get the following terms:

2. ANALYSE THE TOPICS IN THE DOCUMENT

BRUCELLOSIS CONTROL IN CYPRUS

Brucellosis
diseases
man,
economic losses
public health
world.
Mediterranean region
Sheep
Goat
Brucella melitensis
undulant fever
control

cattle
sheep
goats
vaccination
eradication
identification and elimination of infected animals
laboratory facilities
Elimination
a fairly large expenditure of funds

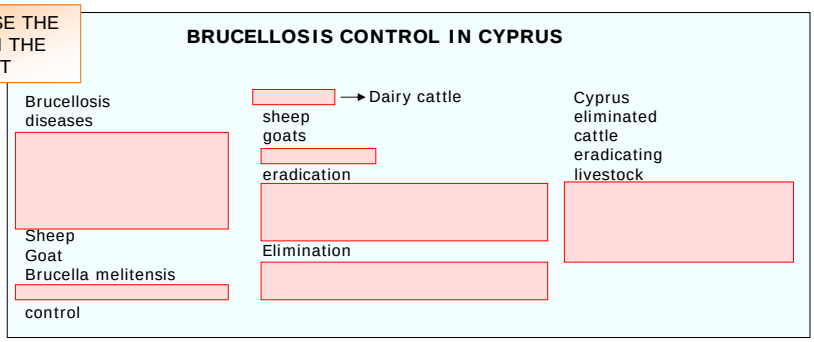
Cyprus
eliminated
cattle
eradicating
livestock
Cyprus goat
native fat-tailed sheep
Chios sheep
Awassi sheep

But not all of these terms are really **topics** of this document…

Example

After scanning the remainder of the text, we can eliminate most of the terms:

2. ANALYSE THE TOPICS IN THE DOCUMENT



For example, the document says nothing about **vaccination** or **identification and elimination of infected animals**. We can also make something more specific: **Dairy cattle** instead of **Cattle**. In the text, there is also **mention** of more specific types of goats and sheep, but **not enough** to merit more specific mention of them.

Example

At this point, we can begin to group the terms logically:

3. CHOOSE THE KEY CONCEPTS

WHAT?	HOW?	WHY?	WHERE?
Sheep Goats Dairy cattle Brucellosis Brucella melitensis Diseases Livestock	----------	Elimination of disease Control of disease Eradication of disease	Cyprus

Example

As we continue our analysis, we see that **Livestock** is more general than **Dairy cattle** and can be eliminated. There are several topics under **WHAT**, but there is no **HOW**. **Disease elimination**, **control**, and **eradication** are synonyms. We choose a form, although we still do not know which one will be used in the catalogue.

3. CHOOSE THE KEY CONCEPTS

WHAT?	HOW?	WHY?	WHERE?
Sheep Goats Dairy cattle Brucellosis Brucella melitensis	----------	Disease elimination	Cyprus

These are the concepts we should find in the catalogue.

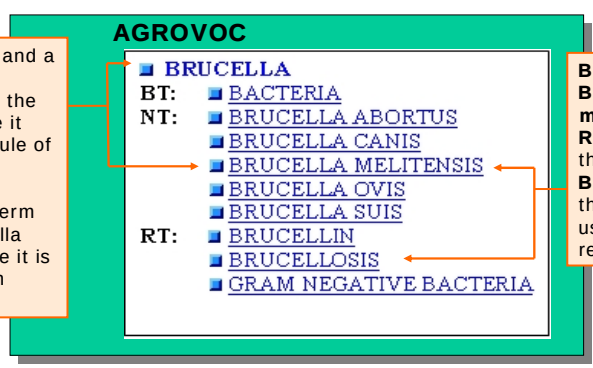
Example

4. CHOOSE THE CORRECT INDEXING TERMS

Searching the FAO online catalogue (which uses AGROVOC) for **sheep**, **goats**, **dairy cattle**, and **Cyprus**, we encounter no difficulties at all, since these are the terms that are used.
On the other hand, for the terms **Brucellosis** and **Brucella melitensis**, we have to consult the thesaurus:

A **Broader Term** and a **Narrower Term** cannot be used in the same record since it would break the rule of **specificity**.

In this case, the term to select is Brucella melitensis because it is more specific then Brucella.



Brucellosis and **Brucella melitensis** are **Related Terms** through **Brucella**, and therefore can be used in the same record.

Example

Now comes **Disease elimination**. A **freetext keyword search** of **"disease elimination"** in the catalogue shows that when this concept is found, the indexers consistently input **DISEASE CONTROL**. Here is a small selection of the search result:

Elimination of iodine deficiency disorders in South-East Asia; Report of a Regional Consultation, New Delhi, 24-26 February 1997 (English) WHO, New Delhi (India). Regional Office for South-East Asia , 1997 , 34 p.
Accession No: 374926, Report No: WHO--SEA/NUT/138 , Call No: PAM616.39 W891 (LIB)
Descriptors: DEFICIENCY DISEASES; IODINE**; DISEASE CONTROL** ; TRACE ELEMENT DEFICIENCIES
Geographic coverage: SOUTH EAST ASIA

Virtual **elimination** of vitamin A deficiency: obstacles and solutions for the year 2000. Report (English) *International Vitamin A Consultative Group. Meeting*, 17, Guatamala City (Guatamala), 18-22 Mar 1996 / International Vitamin A Consultative Group, Washington, DC (USA) , 1996 , 130 p.
Accession No: 357618, ISBN 0-944398-91-X , Call No: 616.39 In86 (LIB) Notes:Summary (En)
Descriptors: RETINOL; XEROPHTHALMIA; **DISEASE CONTROL** ; VITAMIN DEFICIENCIES; TRACE ELEMENT DEFICIENCIES; MALNUTRITION; DEVELOPING COUNTRIES; DEVELOPMENT AID; FOOD ENRICHMENT; MONITORING

Salt iodization for the **elimination** of iodine deficiency (English) Venkatesh Mannar, M.G., Dunn, J.T. / International Council for Control of Iodine Deficiency Disorders, Adelaide (Australia) , 1995 , 126 p.
Accession No: 354068, ISBN 90-70785-13-7 , Call No: 664.8 V55 (LIB)
Descriptors: COMMON SALT; IODINE; FOOD ENRICHMENT ; TECHNOLOGY; METHODS; PLANNING; DEFICIENCY DISEASES; **DISEASE CONTROL**

Normally, the indexer will look at many more records!

Example

It is also possible to use the **thesaurus**, where there is nothing under **disease elimination**, but if we search **disease eradication**, we find:

AGROVOC

DISEASE ERADICATION **use:** DISEASE CONTROL

RESULT (IN AGROVOC)

SHEEP ; GOATS; DAIRY CATTLE ; BRUCELLOSIS; BRUCELLA MELITENSIS ; DISEASE CONTROL ; CYPRUS

Finally, we can **examine records for similar documents** and see if we have missed something. In this case, we have not.

On the left you can see the correct subjects in **AGROVOC** for this document.

Example

When we follow a similar process in following thesauri, we find these subjects:

National Agriculture Library
sheep diseases
goat diseases
dairy cattle
cattle diseases
brucellosis
brucella melitensis
disease control
cyprus

Library of Congress Subject Headings (LCSH)
Brucellosis in animals—Cyprus—Prevention.
Sheep—Diseases—Cyprus.
Goats—Diseases—Cyprus.
Dairy cattle—Diseases—Cyprus.
Brucella melitensis—Cyprus—Prevention.

CAB Thesaurus
sheep diseases
goat diseases
dairy cattle
cattle diseases
brucellosis
brucella melitensis
disease control
cyprus

The **LCSH** are structured in a different way from the other thesauri. Terms are linked together in highly specific ways and in specific orders. Concerning the subject terms used in the **National Agriculture Library** and **CAB International** Thesauri, in this case they are exactly the same, but different from those used in **AGROVOC**.

Using other thesauri, we can also find additional subject areas.

More information about other subject areas

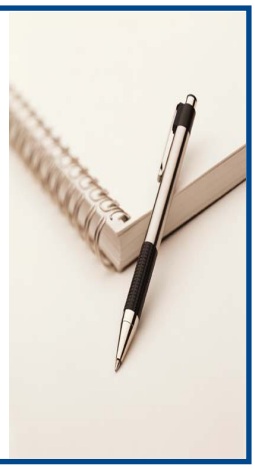
Summary

• The task of the subject indexer is to assign appropriate subject terms from the thesaurus, following a consistent level of **exhaustivity** and **specificity**.

• These are the steps for subject indexing:
1) Understand the concepts in the document.
2) Analyse the topics in the document.
3) Choose the key concepts.
4) Choose the correct indexing terms.

• The selection of appropriate terms is made by **searching for similar items** in the catalogue and by looking up the terms in the **thesaurus** to see if there are any additional terms, or if any synonyms are mentioned.

• If a **subject term** is not available for a specific concept, the indexer may wish to use a **more general** term, or consider **adding it** to the thesaurus.



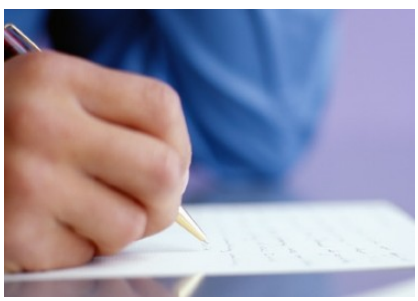
Exercise

The following exercises will allow you to put in practice the indexing principles learnt.

Good luck!

Exercise

Now, it is your turn to index a new document! To start the indexing:



1) **Print** the document.

2) **Read** the title and the summary, and browse the text.

3) **Underline** the relevant terms that describe the topics in the document.

4) Among the underlined terms, **delete** the synonyms and the terms which are not describing the topics in the document.

What you obtain is a list of terms that reflect the key concepts of your document.

YOUR DOCUMENT

Click on this icon to view and print the document

Exercise

What are the key concepts that you have identified?

planting date	north carolina
harvest date	a correlation observed between aflatoxin b1 and
irrigation effects	reduced yield.
infection	corn ear worm or european corn borer
aflatoxin b1	stress conditions that reduce yield
production	predisposing corn to infection
aspergillus flavus	European corn borer/Ostrinia nubilalis
field corn	corn arworm/Heliothis zea

VIEW ANSWER

COMMENT

Type your text in the box. Then, click on VIEW ANSWER to view our answer.
Click on COMMENT to have an explanation of our result.

Exercise

Can you arrange the key concepts into the subject areas?

WHAT?	HOW?	WHY?	WHERE?
field corn *aspergillus flavus* aflatoxin b1 production	planting date harvest date irrigation effects	Infection by A. flavus reduced yield stress European corn borer *or* Ostrinia nubilalis corn earworm *or* Heliothis zea	North Carolina

Click on KEY CONCEPTS to view the list of key concepts.

Type your text in the corresponding box. Then, click on VIEW ANSWER.

KEY CONCEPTS

VIEW ANSWER

Exercise

You now have to search for similar items for each subject area, both through the **thesaurus** and the **catalogue**.

For the following subject areas we encounter no difficulties at all:

WHAT?	HOW?	WHERE?
ZEA MAYS ASPERGILLUS FLAVUS AFLATOXINS	PLANTING DATE HARVESTING DATE IRRIGATION	NORTH CAROLINA
field corn *aspergillus flavus* aflatoxin b1 production	planting date harvest date irrigation effects	North Carolina

When searching **field corn** in the AGRIS catalogue, we discover that the term used consistently is: ZEA MAYS. For **aspergillus flavus** and **aflatoxin b1 production**, there are ASPERGILLUS FLAVUS and AFLATOXINS respectively.

For **planting date**, there is: PLANTING DATE, **harvest date** is HARVESTING DATE. **Irrigation effects** is handled simply as IRRIGATION.

North Carolina is handled as NORTH CAROLINA. Some thesauri deal with areas more comprehensively than others. AGROVOC does not include cities, but other thesauri do.

Exercise

At this point, the task becomes more complex.

WHY?

Infection by A. flavus
reduced yield
stress European corn borer
or Ostrinia nubilalis
corn earworm
or Heliothis zea

Let's now consider the concept "Infection by A. flavus"

As we already noted, this is not only an **infection**, it is an infection of a **plant** by a **specific organism**. But, if we look up "infection" in AGROVOC, we see the following note:

INFECTION
SN: *Process of becoming infected; for the resulting diseases use the appropriate descriptor(s)*

Our article is **not** about the **process** of become infected, but about the **disease**, so it directs us to index the "disease". Therefore, we consider that this is more specifically, a **plant disease**.

Exercise

PLANT DISEASES
UF:
FOLIAR DISEASES
LEAF DISEASES
PLANT DISORDERS
NT:

- BITTER PIT
- BLIGHT
- BLOTCHES
- CANKERS
- CHLOROSIS
- DIEBACK
- FRUIT CRACKING
- FUNGAL DISEASES
- GREENING
- GUMMOSIS
- LEAF CURLS
- PHYLLODY
- PLANT GALLS
- ROTS
- SCABS
- SCALD
- SCORCH
- SHOT HOLES
- SPOTS
- VITRIFICATION
- WILTS

In the thesaurus we can find **PLANT DISEASES**, but before you use it, you must examine the associated terms.

So, you discover that there are several specific types of Plant diseases (Narrower Terms).
You must determine **which specific type** this might be so that you achieve the correct level of specificity.
In this case, to select the relevant term, it would be advisable to search for additional information in the document itself.

Which term would you select?

*Type your text in the box. Then, click on **Confirm** buttom.*

Exercise

Another method would be to discover what sort of diseases ASPERGILLUS FLAVUS causes.

We can look this up in a specialized source, or we can find the same information **in the thesaurus**. Going through the levels of Broader Terms...

ASPERGILLUS FLAVUS
↳ASPERGILLUS
↳DEUTEROMYCOTINA
↳FUNGI

...we discover that ASPERGILLUS FLAVUS is a **fungus** and, consequently, the concept **infection by Aspergillus flavus** is indexed as: FUNGAL DISEASES.

Exercise

WHY?

FUNGAL DISEASES
CROP LOSSES
STRESS
OSTRINIA NUBILALIS
HELIOTHIS

Infection by A. flavus
reduced yield
stress European corn borer
or Ostrinia nubilalis
corn earworm
or Heliothis zea

Reduced yield is also rather difficult.

Although the term YIELD is used, our article is not about yields, but about **reductions** of yields. It is about a plant disease that causes a specific type of corn to be lost.

Imagination is needed and a bit of searching, but the indexer is expected to find the term: CROP LOSSES.
Experience and knowledge of the terms available in the thesaurus aids tremendously in this task.
This could also be a point to suggest that a Related Term reference be made from YIELD to CROP LOSSES.

The rest of the terms are simpler:

Stress conditions = STRESS
European corn borer/OSTRINIA NUBILALIS= OSTRINIA NUBILALIS
corn earworm/ Heliothis zea =HELIOTHIS (since there is no more specific term)

Exercise

The results of your indexing by AGROVOC subject terms is:

AGROVOC

ZEA MAYS
ASPERGILLUS FLAVUS
AFLATOXINS
PLANTING DATE
HARVESTING DATE
FUNGAL DISEASES
CROP LOSSES
STRESS
IRRIGATION
OSTRINIA NUBILALIS
HELIOTHIS
NORTH CAROLINA

Click on OTHER SYSTEMS to view the indexing of the other systems we have discussed.

OTHER SYSTEMS

Examine the differences in the levels of **specificity** of each record, and how it is reflects the range of words available in each thesaurus.

It should now be clear that all subject indexing is based on **relationships among different records in the database** and to the **terms available in the thesaurus**.

If you want to know more...

Subject indexing / General

AGRIS: Guide to Indexing http://www.fao.org/agris/download/agrefs-e.htm

Library of Congress Subject Headings - Principles of Structure and Policies for Application. http://www.tlcdelivers.com/tlc/crs/shed0014.htm

AGRICOLA -- Guide to Subject Indexing / Martha W. Hood
http://www.nal.usda.gov/indexing/subjguid.html

Theory of subject analysis : a sourcebook / edited by Lois Mai Chan, Phyllis A. Richmond, Elaine Svenonius.
What should catalogs do?/ Bernhard Eversberg
http://www.biblio.tu-bs.de/allegro/formate/tlcse.htm