Information Management Resource Kit

Module on Management of Electronic Documents

UNIT 3. METADATA STANDARDS AND SUBJECT INDEXING

LESSON 4. WHAT IS 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:

- understand the **purpose of subject indexing**;
- understand **the way it is used** for retrieval of documents;
- have an overview of **classification systems**; and
- be able to know the **future directions** of subject indexing.

Why is subject indexing important?

I want to find documents on "water with a high concentration of salt". How could I go about this?

It is not enough for someone merely to collect documents and put them together. It is just as important to arrange these documents in such a way that **people can find them** later.

There can be many **different types of people** interested in the documents you collect: they may be experts in the field, students, or people who are simply interested in a topic. These people may also be from other cultures, speak different languages, etc...
How can all of these people find documents on topics that interest them?

The purpose of subject indexing is to **allow a searcher to find the material in a collection that is on a specific subject**.

The role of subject indexing

Accession Number:	97-153734
Title:	Water supply reliability as influenced by natural salt pollution.
Publication Year:	1997
Subject Category:	Water resources and management;
Author:	Wurbs, R.A.
Availability:	NAL, USDA, Beltsville, Md. 20705 - USA. E-mail: gmccone@nal.usda.gov (DNAL TD201.U61).
Bibliographic Source:	references. In the special issue: Integrated water management / edited by W. R. Jordan. Water resources update (USA). (Win 1997). (no. 106) p.116-126.
AGROVOC keywords	SUBJECTS
English:	texas; new mexico; oklahoma; kansas; **saline water**; water quality; surface water; groundwater;
French:	texas; nouveau mexique; oklahoma; kansas; [illegible] saline; qualite de l'eau; eau superficielle; eau souterraine;
Spanish:	texas; nuevo mexico; oklahoma; kansas; agua salina; calidad del agua; agua superficial; aguas subterraneas;

LABELS

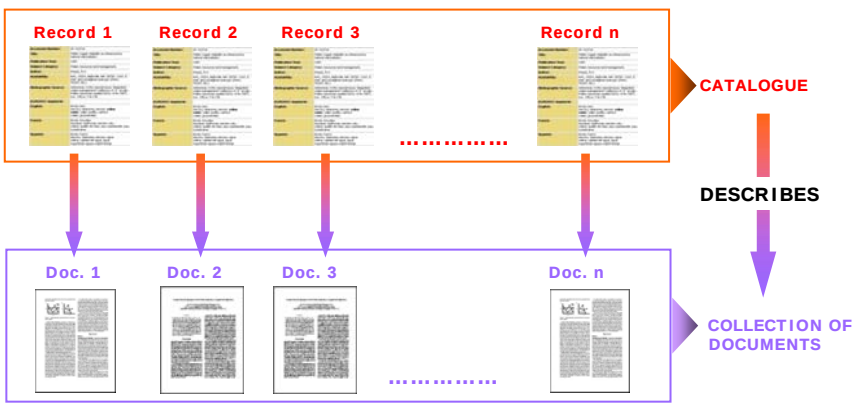
Each document of a collection is described by a **metadata record**, which consists of the title, author, date of publication and other information on the document.

"**Subject**" is a section of the entire metadata record.

Subject content is part of the record related to subject indexing and consists of a list of **keywords** or **labels** (e.g. texas, saline water, etc.).

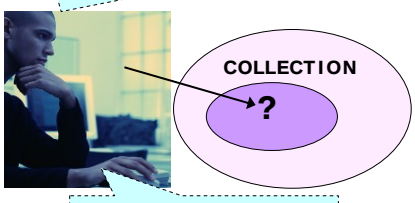
The role of subject indexing

Each record describes a document and all the metadata records referring to a collection of documents are stored in a **catalogue**. A catalogue can be in any format: paper, cards, electronic.



The role of Subject Indexing

I want to find documents on "water with a high concentration of salt".
How could I go about this?



Which labels are used for describing this concept?

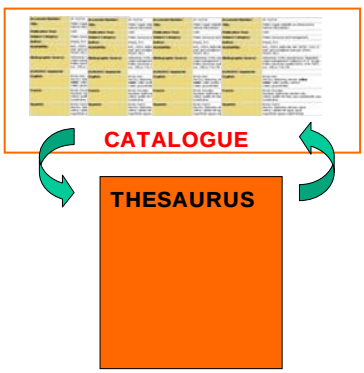
Let's continue with our example.

The **searcher** is actually searching in a collection for a few documents (e.g. subset) about the concept "water with a high concentration of salt". First, he must decide upon some **words for this concept**.

When no subject indexing is used, the searcher must search for all possible words that describe the concept "water with a high concentration of salt".

But, if subject indexing is used, the searcher can find just the **subset** of documents he wants by finding the correct labels.
The trick is that you must discover which **labels** are used for the subset in the **catalogue**.

The role of Subject Indexing



To find the **specific label** used to describe the subject the searcher can use a **thesaurus**: a thesaurus is a list of words that provides **subject terms and instructions** (or references) needed for information retrieval. A thesaurus relates the terms using **Narrower Term**, **Broader Term**, **Related Term** and **Use For**.

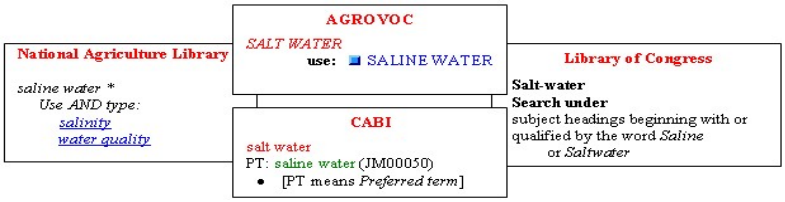
In **multi-lingual** thesauri, such as **AGROVOC**, searchers can also find documents in different languages. In this way, the same subset of documents can have **multiple language versions** for the labels. The series of **instructions or references** that allow users to find the specific labels are provided in the thesaurus itself.

In addition, the user can also search keywords in the catalogue to find records that describe similar subjects.

The role of Subject Indexing

There are many different thesauri. Some of the most popular in the field of agriculture are:

- **AGROVOC** from the Food and Agriculture Organization of the United Nations,
- **Library of Congress Subject Headings** from the United States Library of Congress,
- **CAB Thesaurus** from CAB International (CABI), and
- **NAL Agricultural Thesaurus** from the United States National Agriculture Library.



This diagram shows how a thesaurus is designed to help people find just the labels they need. This illustrates a search for "salt water", and how the user finds the instructions in how to make a correct search.
If someone is searching AGROVOC or CABI, and they look for "salt water", the instructions are to search "Saline water", but in the Library of Congress, there are different instructions. In the National Agriculture Library, the searcher finds nothing for "salt water", but if people search for "Saline water", they will get the correct searching instructions.

According to the thesaurus selected, the subset of items on "**salt water**" have **different labels**. Although the **concept** is the same, the searcher **must know which labels** that thesaurus has chosen to use in order to find documents about it in a specific collection.

The role of Subject Indexing

Subject indexing allows a searcher to find a specific **item** in a catalogue when the subject is known, or to discover **all items** a collection has on a given subject.

Once the searcher has found the correct label(s), he searches those terms and assumes that he has found **everything in the collection** on the subject "water with a high concentration of salt". There is less of a need for the searcher to also search "Salt water", "saltwater", "saline water", "salinity", "water quality", "agua salina", "eau saline" etc., since all these terms have already been grouped under the same label.
Thus, the searcher's time is saved!

Let's now have a look at the steps a subject indexer has to follow to get this result …

The role of Subject Indexing

This document is about saltwater…

Which records will I find when I use “saltwater” as a keyword?

Which subject terms are used in these records?

Can I use these terms to describe my document?

What is in the thesaurus?

The task of the subject indexer is to ensure that the record he makes will be found by users, when they search for specific subjects in the catalogue. Therefore, he must assign the **same labels** that he finds in **similar records in the catalogue**.

Therefore, the **subject indexer** has to:

• determine **what** the document is **about**,

• find which **subject terms** have been assigned for each topic, and

• **assign** those terms **to the new record**.

The role of Subject Indexing

In our example, if the indexer was working with a catalogue using **AGROVOC**, the keyword to use would be “saline water”, but using the **Library of Congress Subject Headings**, it would be “saltwater”.

AGROVOC keywords	
English:	texas; new mexico; oklahoma; kansas; **saline water**; water quality; surface water; groundwater;
French:	texas; nouveau mexique; oklahoma; kansas; eau saline; qualite de l'eau; eau superficielle; eau souterraine;
Spanish:	texas; nuevo mexico; oklahoma; kansas; agua salina; calidad del agua; agua superficial; aguas subterraneas;

The job of the indexer is extremely important: if he neglected, for example, to add the subject “saline water” to the record, this item **would not be retrieved** when the searcher did the correct search in AGROVOC.

If we keep the goals of the searcher in mind, we begin to understand the importance of one of the key ideas in subject indexing: **consistency**…

subject indexing Quality

The **key ideas** in subject indexing are Exhaustivity, Specificity and Consistency.

Exhaustivity — The indexer must ensure that the record will **cover all the topics** discussed in the document.

Specificity — Each subject word chosen by the indexer must match the **scope of the topic**.

Consistency — **All the indexers** should still strive to achieve an **uniform level** of exhaustivity and specificity for the documents treated.

Let's see some examples…

Subject Indexing Quality

Exhaustivity

Production and market for Kamloops trout reared in salt water

Exhaustive
- Production
- Marketing
- Kamloops trout
- Salt Water

Not Exhaustive
- Kamloops trout
- Salt Water

Let's imagine that we have to index a document entitled: "*Production and market for Kamloops trout reared in salt water*".
We have already mentioned that if the indexer ignored the concept of "salt water", the record would be **lacking essential information** and the user would not find this document.

In other words, the record would not be sufficiently **exhaustive**.

But after examining just the title of this item, we see that there is information not only about "salt water", but also about a "species of trout", along with information about "production" and "marketing". This is obvious even before we look into the text of the document, which will give us even more topics.

Subject indexing Quality

Production and market for Kamloops trout reared in salt water

Specific

- Fish Production
- Fish Marketing
- Kamloops trout
- Salt Water

Not Specific

- Production
- Marketing
- Fish
- Water

Along with the concern of exhaustivity, there is a related problem of **specificity**.

To use the same example, to describe "Kamloops trout", a suitable term must be found.

The concept of "salt water" must be paired with an appropriately **specific term used in the thesaurus**.

For example, in AGROVOC, both terms "water" and "saline water" can be used. In this case, the term "saline water" **must** be used since the term "water" is not sufficiently specific.

Subject indexing Quality

"Kamloops trout" is not accepted in AGROVOC.
I can either use the term "trout" or propose a new term…

The problems of **specificity** and **exhaustivity** are related to the terms **available** in the thesaurus you happen to be using.

In the case of "Kamloops trout", there is no accepted label for this specific concept in AGROVOC, and the indexer **would either have to use a less specific term**, e.g. "Trout", **or propose a new term** in AGROVOC. In the **Library of Congress Subject Headings**, there is the more specific term "Kamloops trout".
Different thesauri deal with exhaustivity and specificity in various ways.
Whether a specific term like "Kamloops trout" should be included in the subject indexing varies from collection to collection, depending on **general practice**.

Subject Indexing Quality

The last key concept is **consistency**.

If one indexer treats documents **more exhaustively** than the others, search results become unbalanced. Although absolute consistency is impossible to achieve in reality, all indexers should still strive to achieve a uniform level of exhaustivity and specificity.

Correct Specifity and Exhaustivity for each Thesaurus

Production and market for
Kamloops trout reared in salt water

AGROVOC	CABI	Library of Congress	National Agriculture Library
Trout Fishery production Fish processing Marketing Saline water	Trout Fish production Fish processing Food marketing Saline water	Kamloops trout Fish trade Fish-culture Saline waters	Trout Fish production Fish processing Food marketing Salinity Water quality

From the preceding discussion it should not be surprising that correct subject indexing **can vary significantly** from collection to collection based on differences in specificity, exhaustivity and consistency, and on the available terms.

Overview of Tools for Subject Analysis

How is the indexer or even the searcher supposed to know that there is such a term as e.g. "Saline water" in **AGROVOC**? This is where the thesaurus and subject indexing work together. The user should search at some point for "**water**" in the thesaurus and receive the following result:

- **WATER**

UF: *HYDROMETRY*
HYDROSORPTION
MAGNETIC WATER
MAGNETIZED WATER

NT:
- DISTILLED WATER
- DRINKING WATER
- FRESHWATER
- ICE
- IRRIGATION WATER
- RAINWATER
- SALINE WATER
- WATER VAPOUR

RT:
- BODY WATER
- GROUNDWATER
- HYDROLOGICAL CYCLE

The searcher sees other terms:

- **NT** is **Narrower Term**, and thus is more specific. "saline water" is a specific type of "water".
- **RT** is **Related Term** and shows the searcher other terms that may be of interest.
- **UF** is **Use For**, or **synonyms** for "water".

Therefore, if the subject indexer has an item about Hydrometry, the indexer learns here that he **must assign** the term "water".
Additionally, for those interested in Hydrometry, they learn through this arrangement that they **must search** under "water".

Overview of Tools for Subject Analysis

When we examine the **narrower term**, "**saline water**", we see the following information:

- **SALINE WATER**

UF: *SALT WATER*
SALTWATER

BT: WATER

NT: BRACKISH WATER
SEA WATER

RT: OSMOTIC STRESS
SALINITY
SALT TOLERANCE
WATER DESALTING

Hierarchical Arrangement of Subject Terms

WATER
↳ **SALINE WATER**
↳ **BRACKISH WATER**
SEA WATER

We see here that "saline water" has its own set of **Narrower Terms**, **Related Terms** and **Use For** terms but it has gained a *Broader Term* (BT), i.e. *Water*.
This **hierarchical structure** is followed throughout a thesaurus.

Classification

Imagine a collection of books.
They must be arranged in some way: in chronological order, or by size, or even by colour, etc.

All of these methods are types of **classification**.

Classification is the assignment of a code (number) to each item that belongs to a certain class, the type of code depending on the classification scheme used.

An important method of classification, very popular in library settings, is **by subject**.

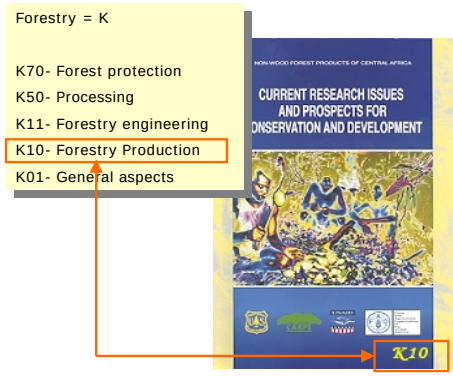
Classification

For example, in AGRIS Category Code, **K10** means **Forestry Production**: this allows the librarians to arrange their materials more easily than if they wrote "Forestry Production" on documents.

The major purpose of subject classification is in fact to **arrange items for browsing**.

Classifications often assign **only one classification number** to each item. This is because a physical item can only be in one place on a shelf. In this case, the classifier must determine the **most important subject** and assign the corresponding number.

They may also be useful for retrieval.



Classification

There are many classification schemes, for example: the **Universal Decimal Classification**, **Library of Congress Classification**, **AGRIS/CARIS Classification, CAB International.** Classification schemes vary considerably in their level of specificity just as with subject thesauri. In the AGRIS/CARIS Classification, Forestry is represented by **K**, in the Universal Decimal by **630**, in the Library of Congress by **SD**.

Some classifications are **universal**, that is, they attempt to classify everything in the universe. The **Universal Decimal and Library of Congress** classifications do this.

Other classifications are more **specialized**, such as the AGRIS/CARIS classification, which deals only with food and agriculture.

Future directions

Currently, there are many experiments to use **computers** to **index documents automatically**. Most of these attempts analyse documents **by counting the words** in a text and relating those words in various ways.

The forest of information: beating path through the jungle

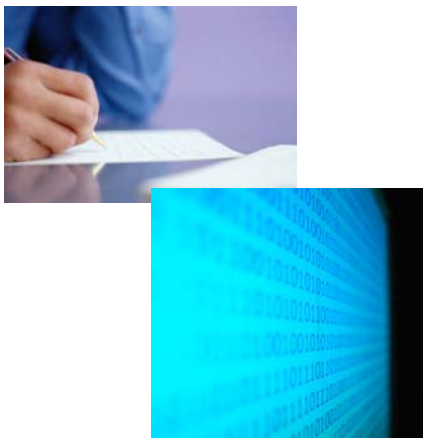
Risto Päivinen is Director of the European Forest Institute, Joensuu, Finland and Coordinator of the International Union of Forestry Research Organizations (IUFRO) Task Force on Global Forest Information Service.
Roger Mills is Head of the Library and Information Service, Oxford Forestry Institute

Trees grow more slowly than crops - during the last cen this simple fact dictated a path of largely separate development for forest-related information retrieval with the broader field of agricultural, environmental and biological information. Specialist services have emerge those seeking data on trees and forests, in recognition the long "half-life" of literature on this subject and the ne

Normally, there is the consideration that the number of times a word is used in a document, the more relevant it is. Thus, a document that uses the word "**forest**" many times in a document will most probably be about forests.

It is also important **how a word is used**. In this example, the word "literature" toward the bottom of the text is less important than the word "forest" in the title. A problem is: in this example, a computer may consider the word "jungle" to be just as important as the word "forest", although this document is not about jungles.
One of the challenges of automatic indexing is to avoid such errors.

Future directions

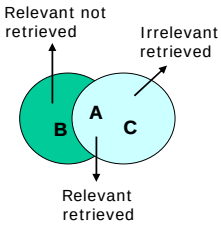


With human indexing, the searcher can be sure that if a human indexer has assigned the AGROVOC term "saline water" to a record, then that document is about the concept "water with a high concentration of salt". With automatic indexing, if the word "saltwater" occurs in a document, there can only be a greater or lesser **probability** that the item truly is about "water with a high concentration of salt".

As we have seen, with human indexing there are problems of everyone achieving the correct specificity and exhaustivity; in automatic indexing there are different problems of **precision** and **recall**. These are the **basic measures** for **evaluating** automated indexing.

Let's have a look at the details…

Future directions



RECALL

$$\frac{A}{A+B} \times 100\%$$

PRECISION

$$\frac{A}{A+C} \times 100\%$$

Both precision and recall are **ratios** determined by the items **retrieved** in a search vs. the total number of items that are **relevant**.
Let us consider that a search for "salt water" has retrieved a total of **70** items. Upon inspection, **30 (A) are relevant** to your search, while **40 (C)** only had the words "salt water" in the text and **are not interesting**. With further analysis, you discover that there are **25 (B) documents** about salt water that you did **not find** for some reason.

Recall is the ratio of the number of relevant records retrieved to the total number of **relevant records** in the collection. This means: did you retrieve all the documents on your subject? This is usually expressed as a percentage (e.g. **54%**).

Precision is the ratio of the number of relevant records retrieved to the total number of **records retrieved** in your search (e.g.**43%**). This means: how many records do you have to go through to find the correct ones?

Future directions

RECALL

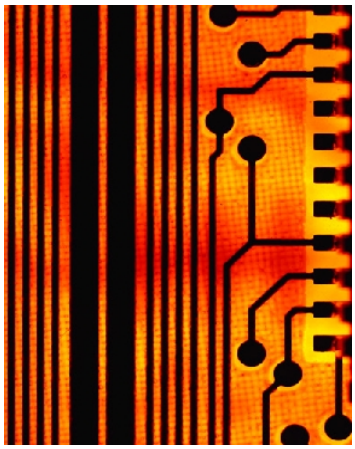
PRECISION

What would you think is the relationship between recall and precision?

- ○ As recall goes up, precision goes down, and as precision goes up, recall goes down (**inverse relationship**).
- ○ As recall goes up, precision goes up, and as precision goes down, recall goes down (**direct relationship**).

Click on the answer of your choice

Future directions



An interesting trend in automatic indexing is the current attempt to link **thesauri** to the **keywords** of the document.

That is, the computer will be able to “know” that the concept “water with a high concentration of salt”, no matter how it is phrased, **will automatically receive the correct subject term**, e.g. “Saltwater” in the Library of Congress Subject Headings.

These are new attempts that are still under development.

Future directions

Another interesting trend is the development of **ontologies**.

What an ontology attempts to do is relate the subject terms together in **more complex ways** than is now done with a thesaurus.

Different scientific communities can create their own ontologies and someday they should all be able to interact.

For example, researchers interested in water from the viewpoint of fisheries development may require different relationships than someone in water management.

The task of an ontology is to clarify these relationships more specifically than Narrower Terms, Broader Terms and Related Terms to enable more precise searching.

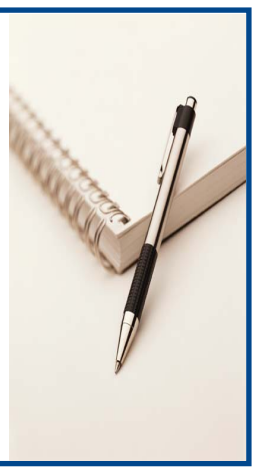
Future directions

Let us reexamine the previous example.
The Narrower Terms and Related Terms become **more specific** in an ontology.
For example, “ice” and “water vapour” are physical forms of water, while “distilled water”, “drinking water”, “freshwater”, and “saline water” have to do with water quality, but “irrigation water” deals with uses of water. “Body water”, “rainwater”, and “groundwater” have a different relationship than the other terms, and finally “hydrological cycle” refers to a function of how water is recycled.

Thesaurus Arrangement	Ontology Arrangement
WATER	**WATER**
UF: *HYDROMETRY* *HYDROSORPTION* *MAGNETIC WATER* *MAGNETIZED WATER*	**UF:** *HYDROMETRY* *HYDROSORPTION* *MAGNETIC WATER* *MAGNETIZED WATER*
NT: DISTILLED WATER DRINKING WATER FRESHWATER ICE IRRIGATION WATER RAINWATER SALINE WATER WATER VAPOUR	**Form:** ICE WATER VAPOUR **Quality:** DISTILLED WATER DRINKING WATER FRESHWATER SALINE WATER **Uses:** IRRIGATION WATER **Location:** BODY WATER GROUNDWATER RAINWATER **Origin:** HYDROLOGICAL CYCLE
RT: BODY WATER GROUNDWATER HYDROLOGICAL CYCLE	

Summary

- The **purpose** of subject indexing is to allow a searcher to find the material in a collection that is on a specific subject.
- To index a new document, the subject indexer must 1) determine what the **document is about**, 2) find **which subject terms** have been assigned for each topic, using catalogue and thesaurus, and 3) **assign** those **terms** to the new record.
- The **key ideas** in subject indexing are Exhaustivity, Specificity and Consistency.
- The task of **classification** is the same as that of subject indexing: to bring similar items together for later retrieval.
- **Precision** and **recall** are the main problems of automatic indexing.
- An interesting trend for subject indexing is the development of **ontologies**.



Exercises

The following five exercises will help you test your understanding of the concepts that were covered in the lesson and provide you with feedback.

Good luck!

Exercise 1

Wildlife
Forest products
Food security
Central Siberia

Let's consider the title:

"Role of wildlife and other non-wood forest products in food security in central Siberia"

Based only on the title, are the concepts listed on the left sufficiently **exhaustive** of the topics of this document?

- ○ Yes
- ○ No
- ○ It depends on which thesaurus is used for subject indexing.

Click on the answer of your choice

Exercise 2

Based only on the title, which of these options is correct?

"Role of wildlife and other non-wood forest products in food security in central Siberia".

- ○ The term "Forest products" is sufficiently specific to the concept "non-wood forest products".
- ○ The term "Forest products" is not sufficiently specific to the concept "non-wood forest products".
- ○ It depends on which thesaurus is used for subject indexing.

Click on the answer of your choice

Exercise 3

human immunodeficiency virus infections
Use:
HIV infections

This reference is from the NAL thesaurus.
What does it mean?

- ○ HIV infections is a more general term in the NAL **thesaurus**.
- ○ HIV infections is a more specific term in the NAL **thesaurus**.
- ○ HIV infections is another term that may interest the searcher in the NAL **thesaurus**.
- ○ HIV infections is the correct term to search in the NAL **thesaurus**.

Click on the answer of your choice

Exercise 4

What are the determiners of quality for Automatic Indexing?

- ○ Recall/Precision
- ○ Consistency / Specificity / Exhaustivity

Click on the answer of your choice

Exercise 5

An ontology differs from a regular thesaurus because it contains more terms.

- ○ True
- ○ False

Click on the answer of your choice

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
Indexing and abstracting in theory and practice / F.W. Lancaster. 2nd ed. 1998.
Indexing from A to Z / Hans H. Wellisch. 2nd ed. 1995
Subject analysis : principles and procedures / D.W. Langridge.

Automatic Indexing
Automatic Indexing and Abstracting / Glenda Browne, *Online Currents*, the AusSI *Newsletter* 20(6):4-9, July 1996 and LASIE 27(3):58-65.
http://www.aussi.org/conferences/conferencepapers/browneg.htm

Classification

Beyond Bookmarks: Schemes for Organizing the Web/

http://www.public.iastate.edu/%7ECYBERSTACKS/CTW.htm

Elements of Library Classification / S.R. Ranganathan.

The Organization of Information / Arlene Taylor.

Ontology
Ontology: Philosophical and Computational / Barry Smith
http://ontology.buffalo.edu/smith//articles/ontologies.htm