NATIONAL OPEN UNIVERSITY OF NIGERIA

DAM 207

Indexing and Classification Theory

Module 4

DAM 207 (Indexing and Classification Theory) Module 4

Course Developer/Writer
Edeama O. Onwuchekwa, National Open University of Nigeria

Programme Leader
Dr. B. Abiola, National Open University of Nigeria

Course Coordinator
Vivian Nwaocha, National Open University of Nigeria

Credits of cover-photo: Henry Ude, National Open University of Nigeria

National Open University of Nigeria - 91, Cadastral Zone, Nnamdi Azikwe Express Way, Jabi, Abuja, Nigeria

www.nou.edu.ng centralinfo@nou.edu.ng
oer.nou.edu.ng oerunit@noun.edu.ng OER repository

Published in 2011, 2014, 2021 by the National Open University of Nigeria

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of National Open University of Nigeria concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. The ideas and opinions expressed in this publication are those of the authors; they are not necessarily those of National Open University of Nigeria and do not commit the organization.

How to re-use and attribute this contents

Unit 1 Indexing

1.0 Introduction

In this unit, you will learn about indexing and the different types indexing languages.

2.0 Objectives

At the end of this unit, you should be able to:

- explain the concept of indexing
- discuss the levels in the process of indexing
- explain the theories of indexing process
- identify the differences between the natural, free and controlled indexing language.

3.0 Main Content

3.1 Indexing

The technique of producing an index is called indexing. Indexing is the process of providing a guide to the intellectual content of a document or a collection of documents. The result of this process is an index, which serves as a pointer to the intellectual content in a document. This role is performed through the descriptors that are used in describing the intellectual content of documents. The reader who is interested in a document will use the descriptors assigned to the document by the indexer. The ultimate objective of the index is to reduce the efforts a user expends in accessing a topic of interest in a particular document or a set of documents which have been stored in a collection.

An index is an important tool for retrieving information contained in documents stored in the library, documentation or information centre. It provides a means of locating the information relevant to a request. Some researchers have postulated that there are five levels in the process of indexing:

The concordance: This consists of all references to all words in the original text arranged in alphabetical order.

The information theoretic level: This calculates the likelihood of a word being chosen for indexing based on its frequency of occurrence in a given text document.

The linguistic level: This attempt to explain how meaningful words are extracted from large units of text. (Some Indexers have proposed that opening paragraphs, chapters etc are good sources for choosing indexing terms).

The textual or the skeletal framework: Here the text is prepared by the author in an organized manner and held together by a skeletal structure. The onus therefore, lies on the indexer to identify the skeleton and markers that will determine the content of the given text.

Indexing theory: This is the inferential level. An indexer should be able to make inferences about the relationships between words and phrases by understanding the sentence structure.

3.2 Indexing Language

Indexing language is made up of words or descriptors that are used in the intellectual contents of documents. These terms are expected to be used by the searcher in order to search for documents in a collection.

There is a need for an artificial language to be used by the indexer and the searcher to describe a document since the terms or concepts identified in a book are represented by words or phrases. The function of this type of language is to ensure that the indexer and the searcher operate at the same level by using the same language. This is to facilitate the retrieval of relevant information from the collection of the library.

This language appears in a variety of forms. It could be **Natural Indexing Language** in which the indexer uses the exact words and phrases used by the author of the document. This is easy to use by the indexer and the searcher but the major problem is that there is no discrimination between synonyms, semantics, homographs, singular and plurals. This type of indexing tends to scatter documents on the same subject, where the authors have used different terms. Natural indexing language is used mainly in the back of book index and computerized indexes such as Keyword in Context (KWIC) and Key Word out of Context (KWOC) indexes.

Another type of indexing language is the **Free Indexing Language.** In this type of indexing language, there is no restriction as to the words or phrases used by the authors or some other words. Both types of indexing languages are very suitable for computer produced indexes. Perhaps the most important type of indexing language is the **Controlled Indexing Language**.

In this type of language, the indexer exercises some control over the terms that are to be used as index terms because the indexer assigns only terms that have been listed as possible index terms. There is generally a preconceived standard list of terms to be used for a particular system.

Thus, when and indexer has identified terms that represent the document, he/she will consult this standard list to ensure that the terms used are consistent. There are two types of this standard list. This list is sometimes called an authority list. The first type is the alphabetical controlled list in which the terms are arranged alphabetically. The two common examples are subject headings list and thesauri. The second type is the classification scheme which assigns notation to subject terms.

3.3 Theories of Indexing

Some theories for explaining the process of indexing does exist although information scientists differ in accepting some of these views. Different researches have but Fugmann (1993) proposed a theory of indexing based on five general axioms.

The axiom of definability proposes that compiling information relevant to a topic can only be accomplished to the degree to which a topic can be defined.

The axiom of order suggests that any compilation of information relevant to a topic is an order creation process.

The axiom of sufficient degree of order posits that the demands made on the degree of order increases as the size of a collection and frequency of searches increase.

The axiom of predictability says that the success of any directed search for relevant information hinges on how readily predictable are the modes of expression for concepts and statements in the search file.

The axiom of fidelity equates the success of any directed search for relevant information with the fidelity with which concepts and statements are expressed in the search file.

Self-Assessment Exercise

What are the theories of indexing?

4.0 Conclusion

In this unit you have learnt about the concept of indexing, levels and theories of the indexing processes and the different types of indexing languages.

5.0 Summary

Indexing is the process of providing a guide to the intellectual content of a document or a collection of documents and the theories of indexing are:

- The axiom of definability
- The axiom of order
- The axiom of sufficient degree of order
- The axiom of predictability
- The axiom of fidelity.

Indexing language is made up of words or descriptors that are used in the intellectual contents of documents. These terms are expected to be used by the searcher in order to search for documents in collection. Examples of these indexing languages are natural, free and controlled languages.

6.0 Self-Assessment Exercise

Define the concept of indexing and describe the different types of indexing languages.

7.0 References/Further Reading

Aina, L.O. (2004). *Library and Information Science Text for Africa.* Nigeria: Third World Information Services Ltd.

Chowdhury, G.G. (1999). *Introduction to Modern Information Retrieval*. London: Facet Publishing.

Fugmann, R. (1993). *Subject Analysis and Indexing: Theoretical Foundation and Practical Advice*. Frankfurt : Indeks Verlag.

Lancaster, F. W. (1991). *Indexing and Abstracting: Theory and Practice.* London: Library Association.

Langridge, D.W. (1989). London. London: Bowker-Saur.

Wynar, B. S. (1992). *Introduction to Cataloging and Classification.* (8th ed.). Englewood.

Unit 2 Indexing System

1.0 Introduction

What you will learn in this unit concerns the types of indexing system, the techniques for indexing a document will equally be discussed.

2.0 Objectives

At the end of this unit, you should be able to:

- describe the types of indexing system
- discuss the subject indexes and its type
- explain techniques of indexing a document.

3.0 Main Content

3.1 Pre-Coordinate and Post Coordinate Indexing System

There are two major types of indexing systems. These are pre-coordinate indexing and post-coordinate indexing.

In pre-coordinate indexing, a subject terms is chosen to represent a document which will serve as the lead term to that document. The document may contain one or more subject terms. One of the terms will be the lead term and the others will be coordinated with the lead term. Because the coordination is done before searching by the user, such type of indexing is called "pre-coordinate indexing." Post-coordinate indexing, on the other hand, involves breaking down a multi-concept subject into single concepts and the searcher would then combine the terms that represent the document required. There is no lead term. Each term is independent and can be combined to suit the interest of the searcher. The co-ordination is done at the time of searching.

A document entitled "The Influence of Britain in the Education of Librarians in Africa" is a multi-concept subject. It can be broken down into the following single concepts:

- Librarians
- Education
- Britain
- Africa.

In pre-coordinate indexing, the lead term would be librarians because that is the main focus of the study. Thus pre-coordinate indexing can be done as follows:

1. Librarians – Education, Africa: Britain
2. But in post-coordinate indexing, the various terms can be combined independently as:

- Librarians and Education

- Africa and Britain

Any document in which these four terms are present would be retrieved. In pre-coordinate indexing, there is always a citation order, that is a prescribed order must be followed, whereas in post-coordinate indexing there is no need for a citation order: Printed indexes such as book indexes, printed indexes and abstracts, national bibliographies, subject catalogues of libraries, etc., are examples of pre-coordinate indexing.

Self-Assessment Exercise

Differentiate between a pre-coordinate indexing system and a post-coordinate indexing system.

3.2 Techniques for Indexing Documents

Indexing is an art that involves a number of stages. The first stage in indexing a document is to have a general idea of the document by going through the title, preface, foreword, content pages and possibly introduction. One can also flip through the text and make some spot reading. This will give the indexer sufficient familiarization with the document; hence this stage is called the **familiarization** stage. The indexer wants to know what the document is about by identifying concepts that are conveyed by words and phrases in the document, examining the title, abstract, preface, introduction, chapter headings, major headings, sub-headings, etc. It is important that the indexer takes into account the needs of the users.

The next stage, which is the **analysis**, involves the indexer using his intellectual judgment by identifying the concepts the book has treated. Sometimes, the indexer may use the exact term used by the author or he might formulate an appropriate term. These terms are intended to accurately describe the whole document. The indexer at this stage is doing what is referred to as subject analysis or concept analysis.

This is where the subject background of the indexer comes into play, especially if he has a sufficient subject background of that document. At this stage, the terms identified by the indexer are what he judges to be the terms that represent the totality of the document. In a situation where the use of terminology is controlled, the indexer cannot use these terms directly as index terms or access points. Rather, terms identified have to be translated into an indexing language used by the system which is the language used by both the indexer and the searcher in an information storage and retrieval process. This language exercises some control over what terms to be used as index terms.

During this stage, the indexer assigns subject descriptors chosen from the controlled language that the users of the discipline are familiar with. This stage is called the **translation** stage. However, in a setting where there is no need to exercise control over the terminology of the system, such as the block of a book index or computerised indexes, this last stage may not be necessary.

4.0 Conclusion

In this unit, you have learnt about pre and post-coordinate systems of indexing. You have also been introduced to the techniques for indexing a document.

5.0 Summary

What you have learnt in this unit include the types of indexing system and the techniques for indexing. In the next unit you shall learn about the evaluation of an index.

6.0 Self-Assessment Exercise

1. Explain the pre and post coordinate systems of indexing.
2. Describe the techniques in indexing a document.

7.0 References/Further Reading

Aina, L.O. (2004). *Library and Information Science Text for Africa.* Nigeria: Third World Information Services Ltd.

Lancaster, F. W. (1991). *Indexing and Abstracting: Theory and Practice*. London: Library Association.

Langridge, D.W. (1989). *Subject Analysis: Principles and Procedures.* London: Bowker-Saur.

Wynar, B. S. (1992). *Introduction to Cataloging and Classification*. (8th ed.). Englewood.

Unit 3 Evaluation of an Index

1.0 Introduction

Whatever the type of index produced, there are needs for evaluation in order to determine how effective the indexing has been in relation to how many documents that contain a particular term can be retrieved from the system. Also to be determined is how many of the documents retrieved from the system can be said to be relevant to the user who is interested in that term. Thus, a good index has a number of parameters by which it can be judged if it is good or not.

This unit describes the quality of an index and discusses the parameters for index evaluation.

2.0 Objectives

At the end of this unit, you should be able to:

- state the qualities of an index
- explain the parameters of evaluating an index.

3.0 Main Content

3.1 Quality of an Index

If the indexing process is done properly, it would be expected that only relevant and specific documents would be retrieved within the shortest possible time, in which case the index could be said to be good. If, however, in the process of searching for relevant documents, a lot of difficulties are encountered, then we say the index is bad.

However, in determining the policies, certain features of indexing have to be explained. These include depth of indexing, specificity, exhaustively and weighting, etc.

Depth of Indexing involves selecting as a large number of topics from a document, that is making as many important topics as are treated in a document as index terms for the document. **Specificity** involves selecting only terms that are specific to the document, which is a term that entirely covers the document.

Exhaustively on the other hand, is related to the number of concepts covered in a document that would be selected for indexing a document. Thus, the more topics selected from a document the more exhaustive the indexing is.

Weighting is another important device employed by indexing agencies. This involves assigning weights to the various terms selected from the document, thereby showing their relative importance and then ranking the terms. Thus, if there are 10 terms selected from a document, and the indexing agency as a policy does not include more than five terms, it would be easy to select the five terms based on a weighting scale. The frequency of occurrence of words in a title or text of a document is a good way of weighting terms.

Self-Assessment Exercise

What are the qualities that make a good index?

3.2 Evaluation of an Index

An index has to be evaluated in terms of its efficiency and effectiveness. Also, there are principal measures for evaluating the effectiveness of an index. These measures are:

- **Recall Ratio** is a quantitative ratio of the number of relevant documents retrieved to the total number of relevant documents present in a collection. This is a quantitative measure used to determine the ability of an index as an aid to retrieving documents containing information on a particular request from a collection of documents present in a library of an information centre.
- **Precision Ratio** is a quantitative ratio of the number of relevant documents retrieved to the total number of documents retrieved.

$$\text{Precision Ration} = \frac{\text{No. of relevant documents retrieved}}{\text{No. of documents retrieved}} \times 100$$

Thus, if out of the 100 documents retrieved in the system, using the index prepared by Indexer D, only 35 are relevant to the user, the precision ratio of index is:

$$\frac{35}{20} \times 100 = 75\% \text{ precision.}$$

It is therefore obvious that there is an inverse relationship between recall ratio and precision ratio. The higher the recall ratio, the lower the precision ratio and vice versa. The more documents that are recalled, the less precise the indexing system would be, and the less documents that are recalled, the more precise the indexing system is. Thus, the indexer must ensure a fair balance of recall ratio and precision ratio. We therefore, expect about 70% recall ratio and 60% precision ratio.

It should be noted that specificity and exhaustively have influence on recall and precision ratios. When the indexing policy of a library or an indexing agency is to support exhaustively, then it would result in a high recall of documents and a low precision that is most of the documents recalled would not be relevant. On the other hand, when an indexing agency supports specificity, then the recall of documents would be low, but the precision would be high as only documents that are relevant to the user would have been recalled.

Specificity promotes low recall and high precision while exhaustively promotes high recall and low precision.

Time: The main function of an index is to reduce the time it would take a user to retrieve documents in a collection. Thus, a good index is that which takes a minimum time to retrieve documents that are relevant and precise to the query. However, the time it takes a user to retrieve relevant documents does not depend solely on the index alone, the ability of the user to precisely select terms that appropriately describe the query is a factor in quickly retrieving relevant documents in a collection. Thus, if an index is good but the user has not used the appropriate descriptors, the user would take a longer time to retrieve relevant documents; but all things being equal, a good index should enable a reader to use a short time to get relevant documents from the collection.

Cost: A good index should be able to serve its purpose with minimum cost. Thus, a good index should be affordable to an average library. No matter how efficient an index, if it is costly, then it might not be available to an average library. No matter how efficient an index, if it is costly, then it might not be available to an average library, archives or information centre. When it is available a reader might have to subsidize the cost, which many readers might not be able to afford.

4.0 Conclusion

The aim of any index is to enable the user to use the index with minimum difficulties. In addition, a good index should be able to retrieve all documents needed for a particular query within the shortest possible period. If the indexer has chosen appropriate index terms to describe the documents in a collection, it would be possible for a user to retrieve all the relevant documents needed instantly.

5.0 Summary

In this unit, you have learnt about the evaluation of an indexing system.

6.0 Self-Assessment Exercise

1. What are the qualities of an effective index?
2. Explain the parameters to be considered when evaluating an index.

7.0 References/Further Reading

Aina, L. O. (2004). *Library and Information Science Text for Africa.* Nigeria: Third World Information Services Ltd.

Chowdhury (1999). *Introduction to Modern Information Retrieval.* London: Facet Publishing.

Unit 4 Automatic Indexing and Classification

1.0 Introduction

Now that you have learnt the concept of classification and indexing in the conventional way, it is time to learn about automatic classification and indexing processes. You will equally learn about classification and indexing beyond the traditional library.

2.0 Objectives

At the end of this unit, you should be able to:

- explain what is meant by automatic system of classification and indexing
- describe the automated method of classifying information
- explain the concept of classification and indexing beyond the conventional setting.

3.0 Main Content

3.1 Classification and Indexing beyond the Traditional Library

Subject access tools have been developed to organize and manage electronic resources. They are known by various names such as subject guides, web guides, subject categories, subject directories, subject hierarchies, pathfinders, and so on. What many of these systems have in common is that they manifest the traditional classification principles of hierarchical structure, domain partition, subordination of the specific to the general, and array of related subjects.

Indexing and classification processes have been performed intellectually by human for quite a long time now. Automatic systems have been developed comparatively in the recent times, where classification and indexing are performed with the assistance of computers. Although the methods for representing the contents of document differ from system to system, the first task which is the analysis of the subject is the same in each case.

When the assignment of the content identifiers is carried out with the aid of modern computing equipment, the operation becomes **automatic indexing.** While term weighting lies in the heart of information retrieval, many techniques have continued to be of interest to the process of information retrieval. Automatic classification is a multivariate statistical technique that groups together similar objects in a multidimensional space.

Although classification was mainly designed for organizing bibliographic items on shelves, in some cases attempts have been made to adapt existing schemes such as the Dewey Decimal Classification (DDC), Library of Congress Classification (LCC), and the Universal Decimal Classification (UDC) to the web environment.

Many researchers have used library classification schemes for organizing information on the web. Typical examples are:

Bubl Link: The subject terms used in BUBL LINK were originally based on LCSH; its users can gain access to the digital resources by a classified list or through an alphabetical list of subjects.

Cyberdewey: This is an example in the use of the DDC in organizing digital information resources. Users can select a subclass or topic to get access to the list the general information resource

Cyberstacks: is a centralized integrated and unified collection of selected digital resources using the LC class scheme. It allows users to browse through virtual library stacks containing monograph or serial works, files, databases or search services to identify relevant information. Resources are categorized first within a broad classification then within narrower subclasses, and resources are listed under a specific class. Each document record comprises a number of field and subfields, each one of which contains a particular unit of information which could be authors' name, publisher's name, title, keywords, number class, ISBN etc. The document record may also contain an abstract or a full text of the document concerned. A text retrieval system is designed to provide fast access to records through any of the sought keys or access points.

In a way, the use of hierarchical or classificatory structure electronically is still relatively new. As information resources continue to grow, one may expect corresponding growth and refinement in ways to organize them. At this point, it is perhaps not too early to consider some of the functional requirements of information organizers. The desirable characteristics may be summarized as follows; a scheme designed for organizing digital resources should be:

- Intuitive, logical, and easy to use, with hierarchies and cross-references clearly displayed and with current and expressive captions
- Flexible, adjustable, and expandable, to reflect rapidly changing and diverse environments
- Useful in a wide range of settings, and applicable over a wide range of the number of sites to which it applies and,
- Relatively easy to maintain and revise.

4.0 Conclusion

In this unit, you have learnt about the automatic system of indexing and classification. You have also been able to identify the different methods of classifying information electronically.

5.0 Summary

What you have learnt in this unit concerns automatic indexing and classification

6.0 Self-Assessment Exercise

Discuss the concept and development in automatic indexing and classification.

7.0 References/Further Reading

Chowdhury (1999). *Introduction to Modern Information Retrieval* London: Facet Publishing.

Juris, Dilevko & Lisa, Gottlieb (2009) .The Relevance of Classification Theory to Textual Analysis. *Library & Information Science Research*, Volume 31(2), pp. 92-100.

Marcella Rita & Newton Robert. (1994). *A New Manual of Classification.* Aldershot: Gower Publishing limited.