Improving Subject Searching in Databases through a Combination of Descriptors and UDC

Mariàngels Granados[*] and Anna Nicolau[**]

[*]*mgranados2@gmail.com,* [**]*anicol@telefonica.net*
School of Library and Information Science, University of Barcelona, Spain,
http://www.ub.edu/biblio

Abstract
Problems with subject access to online catalogues and databases are not new. Studies on the use of OPACs have revealed two apparently endemic problems: on the one hand, the large number of searches with zero hits (failed searches) and on the other, the retrieval of an excessive amount of bibliographic records (information overload). In this paper we describe a new information retrieval technique based on the combination of descriptor weighting and the use of the Universal Decimal Classification (UDC) call numbers.

The use of classification call numbers in order to search the catalogue has traditionally been very restricted. In most catalogues, call numbers are used only as topographical indicators and are not searchable. The new system described here makes much fuller use of them.

The system is based on the hypothesis that a set of descriptors correspond to a UDC call number. Through the analysis of the frequency of distribution of descriptors and call numbers, we create a set of clusters that allow increasing precision and recall. At the same time, these clusters offer alternative search modes, making it possible to systematize the indexing process and increase its consistency. Here we present a case study of the use of the system with the ERIC database.

Introduction

The documentation centre where this study is based specializes in education sciences. The collection contains documents of various kinds, referring to primary, secondary and pre-university education in Spain. The centre serves university teaching and research staff, social scientists, primary school teachers, administrative and management staff, and library staff.

We chose education sciences as the object of our study because of the familiarity with the area we have gained during our professional careers. Though not experts in education sciences, we are well acquainted with indexing and classification in the discipline.

Indexing

Based on our experience in indexing with a variety of multidisciplinary databases and practically all types of documents, we propose a new indexing system which uses a language based on descriptors and the Universal Decimal Classification (UDC). The system works by turning the UDC into the equivalent or transliteration of a system of major descriptors assigned in a logical, regular fashion in order to represent the main content of a document.

This new system involves the use of what we call descriptor- and UDC-permuted indexes. These indexes facilitate comparison of the terms and call numbers inside the same concept according to their proximity, which represents the frequency of their use. With this approach we can establish whether a set of descriptors have been used together with the same UDC call number to represent the same concept. Indexers can then set up a pattern to follow when they wish to represent the same concept, or, in the case of an indexer working alone, he or she can be sure of acting consistently if the concept appears more than once. Indeed, consistency is one of the main problems of indexing. According to Lancaster[1], it is affected by the following factors:

- The number of terms assigned: in general, the degree of consistency is very high for the first two terms chosen. Indexers tend to agree on the main terms.
- The use of a controlled vocabulary *versus* a free indexing vocabulary: a controlled vocabulary provides more consistency, as long as the indexer is familiar with the subject matter and terminology being used.
- The size and specificity of the vocabulary used: in general, the larger and more specific a vocabulary is, the more consistent the result.
- The characteristics of the subject matter and its terminology: specific terms imply more consistency, and abstract terms less consistency. The type of document is also a factor (e.g. consistency is very low in the case of image indexing).
- Indexer-related factors: their training, experience, knowledge of the subject matter, and so on.
- The tools indexers can use: e.g. whether they share the same type of support materials.
- The length of the document to be indexed: consistency is higher when indexing short documents than long ones.

As indexers use the same terms to represent the content of documents that deal with the same concepts, our system also ensures uniformity and homogeneity. We also achieve a high degree of specificity, due to the correspondence between the set of descriptors and the UDC. No set of primary descriptors will be either more general or more specific than its UDC call number. Exhaustiveness is guaranteed because all concepts expressing a specific content will always be present. The degree of exhaustiveness will, of course, be determined by the policy of the centre.

The visualization of the document description follows a specific order and includes descriptors and identifiers in the same field. This is not the order in which they appear in the document, but the order of the document's main concepts, this is a nuclear order, which finds its logical equivalence in the decimal classification. Minor descriptors are assigned to a different field which is defined in the computer application, and appear in the same order according to the same criteria. Chronological identifiers appear at the end of the sequence. All terms are also indexed separately, so that they offer the same search potential.

Policy of the centre

The centre's indexing policy has the following main priority areas: primary education, educational policy, school administration, history of education, educational methods and theories, and teaching activities. These topics (and documents on Catalonia) are indexed in depth.

[1] Lancaster, W. & Pinto, M. (2001). *Procesamiento de la información científica.* Madrid: Arco Libros.

Subjects are indexed combining descriptors and identifiers:

1) Descriptors from the controlled languages.
2) Identifiers to represent and name people and organizations, geographical names, chronology and document types.

We chose a post-coordinated indexing system, taking advantage of its faceted structure and its flexibility in combining descriptors, which allows their logical ordering, as well as retrieval by each of their elements. Specialized controlled vocabularies are used and adapted to the library's materials and interests of the public. Those concepts in the documents which are dealt with partially or superficially will be recorded as minor descriptors if they are regarded as sufficiently significant. If not, they will not be omitted, to avoid information noise.

We index up to a third level of specificity at most. Depending on the length and depth of the treatment of a topic in the document, a more general or more specific term will be chosen (this is known as vertical specificity). In selecting terms for our thesaurus, priority is given to compound terms, e.g. 'Educational legislation' and not 'Legislation' and 'Education' (known as horizontal specificity). The descriptors of the indexing language chosen will be as specific as possible. Staff meetings will be held to validate and approve the descriptors.

In our evaluations of the indexing, user surveys on the effectiveness of the system help us to assess the extent to which they are able to retrieve relevant information. The accuracy rate of the information obtained is also evaluated. The thesaurus is managed via an analysis of the problems observed. For instance, we look out especially for any terms no longer used or descriptors that are used too often in searches, or with a meaning different to the one they have in the thesaurus. In case of doubt, professional experts and collaborators are consulted. The types of document users request more frequently are dealt with more fully: the table of contents is added and documents are summarized; these fields are both searchable by key word, title and document type.

The centre has trained support staff available to help visitors with little experience in consulting the database or those who do not have the time to carry out searches themselves. Users performing searches on their own are given maximum information on the use of the controlled language (including the online thesaurus); users may also consult the permuted indexes in the database. The indexing is based on the whole document, although efforts are always made to identify the most relevant paragraphs. This infrastructure allows the adoption of a mixed system in which indexing staff perform reference tasks and reference librarians take turns in performing indexing tasks. This helps to establish close links with users and adapts indexing to their needs. Descriptors are as specific as possible, and are validated and approved at staff meetings. Funds are available to provide in-service training for librarians in the areas dealt with at the centre.

At our education sciences library, we chose a combination of post-coordinated languages. The thesaurus is adapted from the ones already available in this field:

- Education Resources Information Center (2007). *Thesaurus*. Retrieved September 25, 2007 from http://www.geric.ed.gov.
- Houston, James E. (1990). *Thesaurus of ERIC descriptors*. 12th ed. Phoenix: Oryx.
- European Union Commission and Council of Europe (2003). *Tesaurus europeu de l'educació*. Catalan version. Retrieved September 25,2007 from: http://www.doredin.mec.es/documentos/TEECAT.pdf.

ERIC is an information system sponsored by the US Department of Education and the Institute of Education Science. This thesaurus is our reference point; we have translated it and adapted it to our context. The online version was consulted together with the printed version mentioned above.

The database used by this language structures information in the following way:

High School Foreign Language Programs: A renewed Challenge.
Descriptors: Secondary Education; *Second Language Instruction; *Cultural Awareness; Cultural Education; Critical Thinking; Worksheets; Cross Cultural Studies; Reading Instruction; German
Identifiers: Cultural Literacy
Pub Type: Book

The descriptors preceded by an asterisk are major descriptors and the rest minor descriptors. Identifiers, as we can see, are shown in a different field, as is the document type. The descriptors are jumbled. Our database arranges information in the following way:

Title
Major descriptors:* Educational Policy; *Spain
Minor descriptors: Educational Policy; France
Minor descriptors: Teaching; History; Spain
Pub Type: Journal article
CDU: 37.014(460)
(This document deals mainly with Spanish education policy; it also discusses French education policy, and offers historical information on education in Spain)

Our computer system allows searches for information according to its relevance. The visual representation of the document description follows an order and includes descriptors and identifiers in the same field. This is the order of the main concepts expressed in the documents, not the order in which they appear in the document, this is a nuclear order, and, as we can see, it finds its logical equivalence in the UDC. Minor descriptors are assigned to a different field and are represented in the same order following the same criteria.

Classification

Our arguments for the role of classifications in information retrieval using automated systems are the following:

- Classification systems can combine searches by key word and searches by index (browsing). They allow on-screen display of topic-related records. They can also be used as an interface to consult a subject catalogue.
- Searches can be contextualized within a semantic context. This facet is especially useful when combined with the search by key words, as it reduces the problems of ambiguity caused by the natural language.
- Searches for generic or specific terms can be widened or narrowed thanks to the browsing capacity implicit in the hierarchical structure. Using the code assigned to a known record, other more generic or more specific concepts can be identified, thus improving search strategies.
- Classification codes can be used as a bridge language to overcome the problems posed by very large databases with records in different languages.
- The organization of concepts is based only on subject relationships, not on the alphabet, so that the topics inside a discipline are not dispersed (as is the case with alphabetical classification).

The advantages of the UDC are:

- Simplification in the representation of concepts, since a syntax is applied to relate terms to each other.
- Construction of call numbers comparable to the chain of descriptors.
- Flexibility in the combination of elements and in the ordering of some sub-divisions to adapt to local needs.
- This universal classification system includes all subject areas, to varying degrees.

The major drawback that theorists mention with regard to the use of UDC (and one that we have also noted) is that the user performing the search needs to know the specific number of the concept, and needs to know the classification structure. A further difficulty is that there is no updated and developed online version[2]. But this problem can be averted if the first search is made by key word. In this case, the indexes will send us to the specific UDC number, which we will be able to use to widen our search to the sets of equivalent major descriptors.

The following considerations should be borne in mind when using the UDC:

- The order of elements within call numbers is the order established beforehand in the classification language.
- The classification system by subjects is be used as a topographical system.
- The sub-divisions of form related to electronic documents which do not appear in the UDC will be represented by a colon and the index 681.3, or in a more precise way if possible.
- It is proposed that the UDC should include new resources, and that a new sub-division should be created for them.

  For instance:
  Online electronic resources on education -> 37.01:681.3
- To group documents on school matters together, the call number begins with the element referring to the school subject.

  For instance:
  Document on natural sciences in primary education -> 502:373.3
- Sub-divisions of form and chronological sub-divisions are always placed at the end of the call number, in this order.

  For instance:
  History of education in Catalonia during the 19th century -> 37(467.1)(091)"18"
- For geographical sub-divisions of Catalan toponyms, we use the sub-divisions established by the Rubió i Balaguer classification system.

  For instance:
  Response to special education needs in the city of Barcelona
  -> 373.6(467.1 Barcelona)
- The document management system must allow retrieval via any one of its elements.
- The call number obtained by applying the UDC should correspond to major descriptors only.

[2] UDC Consortium (2006). *Outline of the UDC*. Retrieved September 25, 2007 from http://www.udcc.org/outline/outline.htm.

Document management system used

Cataloguing and indexing module

This module should:

- Allow export or import of records in MARC21.
- Accept the MARC21 format for bibliographical descriptions and display of results.
- Apply MARC control to copies: numbering with barcodes, topographic call number.
- Copy records to create new ones.
- Create different templates for data entry according to different document types.
- Automatically correct an authority in all associated bibliographic records.
- Index concepts that need more than one term to be represented (multiple-word descriptors), e.g. a noun-adjective combination with a different connotation to that of the two concepts taken separately.
- Allow weighting, so as to be able to define major and minor descriptors.
- Create links to related electronic documents: images, search support, etc.
- Work with different sub-catalogues of the same unit for different document locations or types (to reduce the scope of the search, or to speed up searches, or to avoid noise).
- Manage the thesaurus to establish references which relate different indexing terms without causing incompatibilities (*see*, *see also* and hierarchical references).
- Automatically create indexes for the following fields: author, title, subject (descriptors and identifiers) and key words.
- Carry out post-coordinated searches of the UDC, without the need to split up related call numbers.
- Enter summaries of collective works only, including specialized periodicals and compilations: journals, yearbooks, congresses, seminars and bulletins.
- Manage the UDC authority files, in new documentation centres and libraries, after an estimate of the cost.

Consultation and retrieval module (OPAC)

Requirements:

- The ability to add or modify the indexes after the system has been implemented.
- Integration of the automatic control of authority data with the appropriate hierarchical relations and reference sets.
- Users should be able to consult indexing/retrieval languages in order to guide their searches and should have access to the thesaurus and the subject/UDC permuted indexes defined by the system.
- All fields used in the data entry should be displayed, including all notes.
- Descriptions should have different output formats and versions: a form or ISBD (for users), a version in MARC (for library staff) and a reduced and a complete version.
- The catalogue should have internet access with an easy-to-use, parametrizable interface.
- The catalogue should be accessible through various searchable fields: author, title, subject, year, key word, call number…
- The catalogue should be searchable by free text, so that a list or index of key words can be created from the internal data; searches with Boolean operators are possible: “and”, “or” and “not”; truncation and proximity operators are possible.
- Retrieval of all specific terms by entering the generic term to which they are subordinate.
- The possibility of performing searches on previous search groups.
- The possibility of selecting records retrieved from a search for later use.

- Export of search results: print, diskette, CD-ROM, and electronic address.
- Self-protection against excessively long searches and the absence of a maximum session length.
- Avoidance of excessively general searches likely to retrieve a high number of records, in order to reduce the risk of system overload.
- No accents; no special requirements for upper/ lower case.

Example of a permuted index of an eight-document collection

Example of a UDC/ Descriptors permuted index

37.014.6(467.1 Alella)	Escola pública; Avaluació; Alella (Catalunya); 1998
37.057CEIP Fabra(467.1 Alella)	Escola pública; CEIP Fabra; Alella (Catalunya); 1998
371(467.1):681.3	Ensenyament; Noves tecnologies; Catalunya
371.014.1(467.1 Badalona) "1971"	Ensenyament primari; Igualtat d'oportunitats; Badalona (Catalunya); 1971
371.014.53(467.1 Badalona) "1971"	Ensenyament primari; Desigualtat social; Badalona; 1971
371.12(467.1):681.3	Professió docent; Xarxa telemàtica; Catalunya
371.263(467.1 Badalona) "1971"	Ensenyament primari; Avaluació inicial; Tests de diagnòstic; Badalona (Catalunya); 1971
373.3/.47(467.1)	Ensenyament primari; Catalunya
373.3(094.58)(460) "18"	Ensenyament primari; Sistema educatiu; Espanya; XIX
373.3.078(467.1)	Ensenyament primari; Escola pública; Administrador; Catalunya
511:371.014.5	Aritmètica; Ensenyament primari; Innovació curricular
511:373.43	Aritmètica; Ensenyament primari; Didàctica
78:371.3Mètode Willems i Chapui	Mètode Willems i Chapui; Ensenyament primari
78:373.3.02	Educació musical; Ensenyament primari
804.99:373.3(467.1 Badalona) "1971"	Llengua catalana; Ensenyament primari; Badalona (Catalunya); 1971

(The descriptors are ordered decimally, from 0 to 9)

Example of a Descriptors/UDC permuted index

Aritmètica; Ensenyament primari; Didàctica	511:373.43
Aritmètica; Ensenyament primari; Innovació curricular	511:371.014.5
Educació musical; Ensenyament primari	8:373.3.02
Ensenyament; Noves tecnologies; Catalunya	371(467.1):681.3
Ensenyament primari; Avaluació inicial; Tests de diagnòstic; Badalona (Catalunya);1971	371.263(467.1 Badalona) "1971"
Ensenyament primari; Catalunya	373.3/.47(467.1)
Ensenyament primari; Desigualtat social; Badalona; 1971	371.014.53(467.1 Badalona) "1971"
Ensenyament primari; Escola pública; Administrador; Catalunya	373.3.078(467.1)

Ensenyament primari; Igualtat d'oportunitats; Badalona (Catalunya); 1971	371.014.1(467.1 Badalona) "1971"
Ensenyament primari; Sistema educatiu; Espanya; XIX	373.3(094.58)(460)"18"
Escola pública; Avaluació; Alella (Catalunya); 1998	37.014.6(467.1 Alella)
Escola pública; CEIP Fabra; Alella (Catalunya); 1998	37.057CEIP Fabra(467.1 Alella)
Llengua catalana; Ensenyament primari; Badalona (Catalunya); 1971	804.99:373.3(467.1 Badalona) "1971"
Mètode Willems i Chapui; Ensenyament primari	78:371.3Mètode Willems i Chapui
Professió docent; Xarxa telemàtica; Catalunya	371.12(467.1):681.3

(The set of major descriptors in the document are ordered alphabetically)

Conclusions

We propose this methodology for application to metadata for the retrieval of Internet electronic resources, especially for the RDF semantic web.

References

DESIRE. (2004). The role of classification schemes in Internet resources description and discovery. Retrieved September 25, 2007 from: http://www.ukoln.ac.uk/metadata/desire/classification/classification.pdf .

Granados, M. & Dionís Orrit, A. (1991). *Informe presentat al Servei de Biblioteques i del Patrimoni Bibliogràfic de la Generalitat de Catalunya corresponent a la revisió del procés d'indexació d'aplicació a l'Hemeroteca Nacional de Catalunya.* Barcelona, 16 December 1991.

Larson, R. R. (1991). Classification clustering, probabilistic information retrieval, and the online catalog. *The Library Quarterly*, 61, 2, 133-173.

Mcllwaine, I. C. (2003). *Guía para el uso de la CDU.* Adapted by Rosa San Segundo Manuel. Madrid: AENOR.

San Segundo Manuel, R. (1999). Indización en cadena y su aplicación práctica. *La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de la Información. Actas del IV Congreso ISKO-España ECOCONSID'99*, Granada 22-24 April (pp. 53-59).

Slavic, A. (2004). UDC implementation: from library shelves to a structured indexing language. *International Cataloguing and Bibliographic Control*, 33, 3, 60-65. Retrieved September 25, 2007 from http://www.ifla.org/IV/ifla69/papers/032e-Slavic.pdf.