From Classification to Thesaurus ... and Back? Subject Indexing Tools at the Library of the Afrika-Studiecentrum Leiden

Marlene van Doorn and Katrien Polman

Afrika-Studiecentrum, P.O. Box 9555, 2300 RB Leiden, The Netherlands,
<doorn@ascleiden.nl>, <polman@ascleiden.nl>

Marlene van Doorn works as information specialist at the Library, Documentation and Information Department of the African Studies Centre in Leiden. Her work focuses on facilitating access to African Studies resources, in particular through the library's online catalogue, the abstracts journal African Studies Abstracts Online and thematic Web dossiers. She is responsible for the development and maintenance of the African Studies Thesaurus.

Katrien Polman works as information specialist at the Library, Information and Documentation Department of the African Studies Centre in Leiden. Her work focuses on facilitating access to African Studies resources and meeting the information needs of ASC library users. Her tasks include selecting, indexing, and abstracting publications, including articles from periodicals, for the library's online catalogue and its abstracts journal, African Studies Abstracts Online. She evaluates and selects digital resources for the web service Connecting Africa and is involved in the development and maintenance of the ASC's African Studies Thesaurus. She is also the editor of the ASC library's Web Dossiers.

Van Doorn, Marlene, and Polman, Katrien. **From Classification to Thesaurus ... and Back? Subject Indexing Tools at the Library of the Afrika-Studiecentrum Leiden.** *Knowledge Organization, 37(3)*, 203-208. 16 references.

ABSTRACT: An African Studies Thesaurus was constructed for the purpose of subject indexing and retrieval in the Library of the African Studies Centre (ASC) in Leiden in 2001-2006. A word-based system was considered a more user-friendly alternative to the Universal Decimal Classification (UDC) codes which were used for subject access in the ASC catalogue at the time. In the process of thesaurus construction UDC codes were used as a starting point. In addition, when constructing the thesaurus, each descriptor was also assigned a UDC code from the recent edition of the UDC Master Reference File (MRF), thus replacing many of the old UDC codes used by then, some of which dated from the 1952 French edition. The presence of the UDC codes in the thesaurus leaves open the possibility of linking the thesaurus to different language versions of the UDC MRF in the future. In a parallel but separate operation each UDC code which had been assigned to an item in the library's catalogue was subsequently converted into one or more thesaurus descriptors.

1.0 Introduction

The African Studies Centre (ASC) Leiden is an independent foundation established in 1948. Its central aims are to undertake research on Africa in the social sciences, to maintain a specialist library and documentation department, and to facilitate the dissemination of information on Africa. The ASC library collection is the only collection in the Netherlands focusing entirely and exclusively on Africa. It is also one of the largest Africana collections in Europe. Current holdings comprise approximately 75,000 books and pamphlets, 2000 periodicals, of which almost 600 are current subscriptions, 25,000 microfiches of development plans and African newspapers, and about 1,000 documentaries and feature films on DVD. The library collection is a broad-based collection in the field of the social sciences and humanities, focusing on socio-economic and political developments, government, law and constitutional devel-

opment, history, religion, anthropology, women's and gender studies, education, and literature. Roughly half of the collection is in English, about a third in French, and the remainder is divided mostly between German, Dutch, Afrikaans, Portuguese and Spanish.

In recent years, the ASC Library, Documentation and Information Department has put considerable effort into the development of a digital library, providing access to various types of electronic resources. This has been done both by linking to electronic publications within the existing library catalogue and in the form of new services such as Connecting Africa, a web portal which contains metadata records of digital resources with links to full text materials, and AfricaBib, a collection of bibliographical databases hosted since 2008.

With the aim of improving the accessibility of its information resources, the ASC library has developed its catalogue into a comprehensive bibliographical tool in the field of African studies. Between 2000 and 2006, the ASC library carried out a project to improve subject access by building an African Studies thesaurus and converting all subject codes used up to then into thesaurus descriptors.

2.0 From a classification to a word-based indexing system

The ASC catalogue has been available online since 1997. Unlike most catalogues, it contains not only entries for books, journals and DVDs, but also for journal articles and chapters from edited works. Of a total of some 150,000 entries, over half are articles, many with an abstract. Subject access to the collection is through the online catalogue, which runs under OCLC software. Until 2006, subject access was provided by numerical codes based on the UDC. We were using a homegrown version of the UDC, based on the 10th Dutch edition of 1970 and the French edition of class 3, published in 1952. To take into account developments in African studies we had built many numbers for our own use. An alphabetical card index of key-words in Dutch, located in the library, guided library visitors and staff to the UDC codes used in the catalogue. Effective use of the system was not self-evident and frequently required the assistance of the library staff, particularly in the case of users who did not master Dutch. With the development of automated library services in the 1990s there was a growing number of remote users who had to do on their own.

The idea grew to switch to a more user-friendly word-based indexing system and to convert all the UDC codes we had used up to then to the new system so that the entire collection could be searched with one subject language. Fortunately there would be no consequences for shelving, since materials are arranged in the stacks according to date of acquisition.

2.1 Connecting with an existing word-based system

Initially we had hoped to be able to switch to an existing thesaurus or word-based system, preferably one that was also used by other libraries with African Studies collections. The language should be English, to facilitate accessibility of the catalogue to foreigners and remote users, and if possible also Dutch, since many of our library users are Dutch speakers. However, after looking at the indexing systems used in other African Studies libraries we reluctantly concluded that there was no really suitable ready-made word system available. The *Library of Congress Subject Headings*, for example, is too large and complex for our use. The OECD Macrothesaurus is not specifically focused on Africa and its coverage of a number of key subject areas of the ASC collection is insufficiently detailed. Moreover, its future at the time was uncertain and the OECD subsequently discontinued its maintenance. Short of constructing our own word-based indexing system from scratch, the only other option was to start with what we already had, the UDC.

2.2 Using the UDC to develop a word-based system

The "UDC option" could be realized in two ways, either by making an index to the UDC codes we used or by extracting a thesaurus from them. In both cases, the end user would be able to search the online catalogue using words. In other respects, however, the two options are fundamentally different. In the first case, the UDC is made accessible through the creation of an alphabetical subject index, which can be used to find the right code. The classification used by the University Library of Kiel is accessible in this manner (http://www.ub.uni-kiel.de/fach/systematik/Einleitung.html) (Erdei 1999). When we were conducting the preliminary research for our project in 1999, Leiden University Library also had plans to facilitate the use of its classification system through a word-based retrieval system by linking each code to a word string consisting of a description of the code followed by descriptions of concepts higher up in

the classification (Huisman 1999). In the second case, the UDC ordering is used to group concepts into fields of study or subject areas as the starting point for constructing a thesaurus (Riesthuis and Bliedung 1991; Frâncu 2000, 2004). Each UDC code is transformed into a descriptor or descriptors which best convey the concept represented by the UDC code. The descriptor is embedded in a thesaurus structure, with cross-references (non-preferred terms), hierarchical (BT/NT) and associative relationships (RT). An example is the ETH-Bibliothek Zürich, which has used the UDC to develop thesauri in German, English, and French (www.nebis.ch) (Loth 1996a, 1996b; Schwaninger 1996, 1997).

2.3 Choosing between the options provided by the UDC

Whichever option we chose, digitizing the Dutch alphabetical card index to the UDC codes used in the library catalogue seemed a good first step. The resulting text file contained a total of 18,089 records, of which 8,880 were distinct UDC codes, though not all were unique in the strict sense of the term. About 2,000, for example, were a combination of one of 18 codes and a person's name, such as our use of the code 92 followed by the name of a person to classify biographies.

Subsequently, we explored the first option, that of making an index to the UDC codes. We envisaged using the UDC MRF. This would provide us with descriptions in English to which we could add the Dutch descriptions from our own card file, as well as descriptions from the French and German versions of the MRF which were in the process of being developed (McIlwaine 2000). The codes we used would have to be matched with the codes in the UDC MRF, a process which could be automated. The overlap with the UDC MRF was estimated at maybe 60 percent. This option, potentially very attractive, proved rather more complicated than we anticipated. The terminology of the alphabetical index to the MRF would require a considerable amount of editing to turn it into a satisfactory retrieval language. At the least, it would be necessary to replace descriptions or phrases with more precise terms and to deal with problems such as polysemy. In addition, a separate procedure would have to be developed for the codes we used which did not match an MRF code. A further consideration was the fact that this option entails continued use of codes and a classification as the indexing language, contrary to the wish of the ASC information specialists to switch to a word-based system. For these reasons we finally decided in mid 2000 to choose the alternative option of constructing a thesaurus using the ASC UDC codes as starting point.

3.0 Challenges in constructing the thesaurus

Constructing the thesaurus was a challenge in multitasking and in bridging the differences between a classification and a word-based system. Some of the main challenges were conceptual, arising from semantic differences between English and Dutch, and the confusion created by the fact that activities that were analytically distinct were inextricably linked in practice and undertaken simultaneously. Other challenges more directly concerned issues of thesaurus construction, such as specificity of vocabulary, vocabulary control, and the organization of country related descriptors.

3.1 Conceptual challenges

The working language was Dutch and we were starting with terms or descriptions in Dutch, linked to concepts represented by a code derived from a classification. These had to be transformed into a descriptor or descriptors in English, the first language of the thesaurus. Moreover, since we wanted to convert all the UDC codes assigned to titles in the online catalogue into descriptors once the new thesaurus was completed, we were at the same time also setting up a concordance between the thesaurus descriptors and the UDC codes from which they had been derived. Often there was no one-to-one match between a descriptor and a code. Inconsistent classification practice over the years and shifts in the meaning and coverage of codes further complicated both thesaurus construction and the mapping of codes and descriptors. In practice, this meant we were also reviewing ASC classifying practice in an endeavour to improve future recall and precision rates in information retrieval.

3.2 Specificity of vocabulary

Specificity of vocabulary and level of precoordination, i.e., when to use a compound term and when to split it into simpler single terms, is generally acknowledged as one of the most difficult problems in thesaurus construction (Aitchison et al. 1997). An apt illustration is our decision not to include descrip-

tors for specific instances of ethnic literature. This would have meant including a descriptor for every combination of ethnic group and literary genre, which we felt would have led to a disproportionate number of descriptors, as well as requiring more attention in maintenance. Instead, publications about Hausa literature, Yoruba drama, Shona prose, etc., are indexed with the descriptor for the ethnic group together with the descriptor for literature or for the specific genre in question. The descriptors literature and ethnic literature have a user tip to this effect.

In some cases, a decision not to use a compound term also had repercussions for the conversion of the UDC codes. The UDC codes for trade, export and import, for example, were often linked by a colon with the UDC code for a particular product. In the thesaurus, most of these compound codes have been split into their component concepts and occur as single terms. An exception was made in a few cases, such as slave trade, agricultural exports and food imports. These compound terms are so commonly used that we felt that splitting them into their constituent elements would have been counterproductive.

We made fairly frequent use of upward posting, treating specific terms as if they are equivalent to their broader terms, especially in subject areas which are not the core of the library collection and where a low level of indexing specificity is sufficient. An example is felines, with three non-preferred terms, cats, leopards, lions.

3.3 Vocabulary control

Vocabulary control is inherent to a thesaurus. Control extends to the form of the term, such as the use of singular or plural, spelling, abbreviations and acronyms, the choice of preferred and non-preferred terms, the inclusion of certain types of terms, such as proper names, and the definition of the meaning of a term by adding scope notes or using qualifiers.

In a classification, vocabulary control is of secondary importance. A concept acquires its meaning by virtue of its place in the classification. This is particularly evident in the case of polysemy. We dealt with this on a case by case basis. In the case of different ethnic groups or languages having the same name, the name of the country was added as a qualifier, such as Ndebele language (South Africa) and Ndebele language (Zimbabwe). Sometimes we were able to find alternative terms. For example, to distinguish between drugs of abuse and drugs used as medication, corresponding to codes from two different subclasses, namely 66, chemical and related industries, and 61, medical sciences, we used two separate descriptors, drugs and medicinal drugs. Conversely, in other cases it was possible to combine codes from different classes of the UDC when the concept which they represented was essentially unambiguous. The descriptor parliament, for example, corresponds to codes from the subclass 32, politics, and 34, law.

Almost a third of the terms in the African Studies Thesaurus are the names of languages and peoples or ethnic groups. Amongst the sources we used are the *African Studies Thesaurus* (Otchere 1992), *Ethnologue* (Lewis 2009) and *African Ethnonyms* (Biebuyck et al. 1996). In principle the preferred term is the name without a prefix. There are cross-references to alternative names, variant spellings and names with prefixes. For example in the case of Fulfulde, a language spoken in West Africa, there are eight such non-preferred terms. By comparison, the UDC MRF includes only two of the non-preferred terms, while the term Fulfulde itself does not appear. To find a term for indexing or retrieval in a thesaurus, searching for that term will often be sufficient. In a classification, browsing in the hierarchy is more likely to lead to the desired code.

3.4 The organization of country related descriptors

We felt that an especially appropriate and relevant feature of an area studies thesaurus would be a country presentation of peoples, languages, polities and political parties. Initially we thought we could realize this by creating an RT relationship between a country and its various objects. However, in most African countries, this led to a long alphabetical list of RTs in which the peoples, polities, languages and political parties were all mixed up, something which we felt was not particularly useful. The intercalation of node labels would have led to a more helpful order. Unfortunately, this was not an option because the software program we were using to construct the thesaurus treated these labels as though they were descriptors for the purpose of constructing hierarchies. This would have created incorrect hierarchies between a country and its objects. We eventually dealt with this problem by introducing broad geographical headings for peoples, languages, polities and political parties, one set for each African country, in the manner of Ghanaian languages, Ghanaian peoples, Ghanaian polities and Ghanaian political parties. These broad geographical headings have an RT relationship with the country in question and a BT/NT relationship

with the descriptors they group. We have dubbed them "artificial" descriptors because they are not used for indexing. They serve exclusively to group the objects of the country in question.

4.0 Completing the African Studies Thesaurus

In early 2001, work on constructing the thesaurus began when the digitized card file of UDC codes used in the ASC library catalogue was imported into MultiTes, a software program specifically designed for thesaurus construction and chosen amongst others because it supported a multilingual thesaurus. Five years later, the African Studies Thesaurus was ready for use. It contains a total of 12,319 terms, of which 5,257 are descriptors or preferred terms and 7,062 are non-preferred terms. The thesaurus contains some 1,600 fewer descriptors than the UDC codes which served as the source. The difference can be largely accounted for by upward posting, the splitting of compound codes into component concepts and the combining of codes from different classes, as described above.

Early on in the project, it became obvious that the complexity of the task prevented us from realizing one of our goals: the construction of a multilingual Dutch-English thesaurus. In addition, the systematic display of descriptors in broad subject categories which we had envisaged as an integral part of the thesaurus has yet to be realized.

4.1 A multilingual thesaurus?

In a multilanguage thesaurus-building exercise, all languages have to be treated equally, relational structures have to be developed for each of the languages, and numerous crosslingual and intralingual issues are bound to emerge. Under the circumstances, all our effort went into the construction of an English-language thesaurus, the preferred language for our new indexing system. However, we did retain the original Dutch terms from the alphabetical index to the UDC. These were imported into the thesaurus database and were helpful for the native Dutch speakers involved in the 'translation' process. We also hoped to be able to use them later on to develop a full-blown Dutch version of the African Studies thesaurus.

4.2. Broad subject ordering: back to a classification?

When constructing the thesaurus, each descriptor was assigned a UDC MRF code, intended as a starting point for developing a systematic display of descriptors in broad subject categories. Unfortunately this was not achieved in the course of the project. Now, however, there are compelling reasons for completing this, amongst others the advantages of a classification for subject browsing, and the potential use of categories in mapping terms from multiple databases (Clarke 2001) in order to provide integrated subject access to the ASC digital library.

Preliminary work indicates that it is possible to present the African Studies thesaurus as a UDC classification. This provides a framework for the development of a broad subject ordering. However, for reasons inherent both to the UDC itself and to our use of it in the thesaurus, it is unlikely that it will be possible to generate a broad subject ordering automatically.

Firstly, not all subclasses of the UDC are hierarchically related to their higher classes in the strict sense of the term. Unless the structure of the UDC notation reflects correct semantic relations right truncation of UDC codes to create broader categories is not feasible. This problem is compounded by the selective nature of the UDC classification produced on the basis of the African Studies thesaurus. Not all subclasses are represented and the number of levels of division included varies considerably between classes. In classes where there are relatively few descriptors, such as class 5, the level of division, and consequently right truncation of UDC codes, will be higher than in classes with many descriptors, such as class 3.

Secondly, while constructing the thesaurus we assigned UDC MRF codes to the descriptors without having a preconceived broad subject ordering in mind. Undoubtedly we will need to shift some descriptors from one (sub)class to another when determining subject categories. Once the broad subject ordering is completed, it will be necessary to choose a notation to represent the concepts and their filing order, and for use in online subject searches. Connecting with the UDC notation would seem a logical choice.

5.0 Conclusion

Now that we have been using the thesaurus for almost three years, the respective advantages and disadvantages of a thesaurus and a classification are becoming increasingly obvious. As a word-based system, the thesaurus is far more user-friendly than a classification. Moreover, it expresses multihierarchical and associative relations not present in a classification, facilitating the identification of relevant in-

dexing terminology and improving subject access. The African Studies thesaurus is available on the web at http://www.ascleiden.nl/Library/Thesaurus/. Web statistics indicate a steady and growing use.

On the other hand, a classification expresses relations not present in a thesaurus, such as subject categories or fields of study. A classification makes it easier to monitor developments in the library's collection profile. A classification can be used to search for all publications in a particular category, as was the case when we used the UDC. Illustrative in this respect is the search query used to retrieve titles on religion for the *Journal of Religion in Africa*. When the UDC was used as the indexing language, one search query in the online catalogue using the truncated UDC code for class 2, religion, was sufficient to retrieve all relevant titles. Following the introduction of the thesaurus, the search statement became a string of 100 descriptors, which had to be split up into a series of search queries in the online catalogue.

Our experience to date underlines the complementarity of a thesaurus and a classification as subject languages for indexing and retrieval (Bland and Stoffan 2008). With the development of a systematic display in broad subject categories of the descriptors in the African Studies thesaurus, we hope to further enhance the value of the thesaurus as a tool for managing African Studies information resources. A systematic subject ordering may further enhance the search and retrieval options of the library's online catalogue as well as facilitate the mapping of terms from different word-based subject indexing systems.

References

Aitchison, J., Gilchrist, A. and Bawden, D. 1997. *Thesaurus construction and use: a practical manual.* 3rd ed. London: Aslib.

Biebuyck, D.P., Kelliher, S. and McRae, S. 1996. *African ethnonyms: index to art-producing peoples of Africa*. New York: G.K. Hall.

Bland, R.N. and Stoffan, M.A., 2008. Returning classification to the catalog. *Information technology and libraries* 27: 55-60. Available http://www.ala.org/ala/mgrps/divs/lita/ital/272008/2703sep/bland_html.cfm

Clarke, S.G.D., 2001. Thesaural relationships. In: Bean, Carol A. and Green, Rebecca, eds. *Relationships in the organization of knowledge*. Dordrecht: Kluwer, pp. 37-52.

Erdei, K., 1999. Systematik im Online-Katalog der UB Kiel: ein Werkstattbericht. *Bibliothekdienst* 33: 302-10.

Frâncu, V., 2000. Harmonizing a universal classification system with an interdisciplinary multilingual thesaurus: advantages and limitations. In Beghtol, Clare, Howarth, Lynne and Williamson, Nancy, eds., *Dynamism and stability in knowledge organization. Proceedings of the Sixth International ISKO Conference, 10-13 July 2000, Toronto, Canada*. Advances in knowledge organization, 7. Würzburg: Ergon, pp. 200-05.

Frâncu, V., 2004. UDC-based thesauri and multilingual access to information. *Extensions and corrections to the UDC* 26: 48-57.

Huisman, F., 1999. Anders zoeken met een classificatie. *Informatie professional* 3n10: 49-53.

Lewis, M. P. (ed.) 2009. *Ethnologue: languages of the world.* 16th ed. Dallas, TX.: SIL International. Available http://www.ethnologue.com

Loth, K., 1996a. Überlegungen zu einer computergerechten Reorganisation der UDK. *Extensions and corrections to the UDC* 18: 9-13.

Loth, K., 1996b. Wissensorganisation durch ein neues Notationssystem: eine konstruktive Kritik der UDK. *ABI-Technik* 16n1: 17-28.

McIlwaine, Ia. 2000. Personal communication 28 July.

Otchere, F. E., 1992. *African studies thesaurus* : *subject headings for library users.* Westport, CT : Greenwood Press.

Riesthuis, G. J. A. and Bliedung, S. 1991. Thesaurification of the UDC. In Fugman, Robert, ed., *Tools for knowledge organization and the human interface. Proceedings of the 1st International ISKO Conference, Darmstadt, 14.-17 August, 1990*. Advances in knowledge organization 2. Frankfurt: Indeks, pp. 109-17.

Schwaninger, L. 1996. Hierarchiebildung bei numerischer Indexierung: schnellerer Zugang zum Wissen mit einer online-abfragbaren DK. In Neubauer, W.; Schmidt, R. eds. *Information ohne Grenzen. Wissensvermittlung im Zeitalter der Datennetze. 18. Online-Tagung der DGD: Proceedings*. Frankfurt am Main: Deutsche Gesellschaft für Dokumentation, pp. 75-83.

Schwaninger, L., 1997. Mehrsprachigkeit und Begriffshierarchie bei der Literaturrecherche an der ETH-Bibliothek Zürich. In Ockenfeld, M.; Schmidt, R. eds. *Die Zukunft der Recherche: Rechte, Ressourcen und Referenzen.19.Online-Tagung der DGD: Proceedings*. Frankfurt am Main: Deutsche Gesellsc haft fur Dokumentation, pp. 81-92.