68th IFLA Council and General Conference
August 18-24, 2002

Code Number:	**007-122-E**
Division Number:	**IV**
Professional Group:	**Classification and Indexing**
Joint Meeting with:	**-**
Meeting Number:	**122**
Simultaneous Interpretation:	**-**

Subject retrieval in distributed resources: a short review of recent developments

Martin Kunz
Die Deutsche Bibliothek
Frankfurt am Main,
Germany

Abstract:

Subject-based retrieval in distributed resources is a current problem in online searches for bibliographic references. Building portals to similar resources is only the first step, the subsequent navigation via different search interfaces presents certain difficulties. To make retrieval easier it is necessary to adapt these different resources.
Potential approaches (standardisation as opposed to "cross-walks") and methods (automated as opposed to intellectual effort) will be discussed. This includes a brief appraisal of the future of work with multilingual terminology:

- *The "classical" approach (Multilingual Thesauri),*
- *The "Internet" approach (linking)*

Recent developments in mono- and multilingual environments will be presented (MACS, CARMEN, Economics Crosswalk).

Summary:
Subject searching across distributed resources is a current challenge when carrying out online searches for bibliographic data. The construction of portals for comparable sources is only the first step, the subsequent navigation of disparate search interfaces still presents problems. If retrieval is to be improved, there must be some adaptation of these differing resources.
Potential approaches (standardisation versus crosswalks) and methods (automated versus intellectual) will be discussed. This includes a brief appraisal of future work on multilingual terminology:

- the "classical" approach (multilingual thesauri)
- the "internet" approach (links)

New developments in monolingual and multilingual environments will be described (MACS, CARMEN, Economics cross-concordance).

1. Starting point

Let us begin with a brief glance back at what the world of librarianship looked like not so long ago. The German-speaking sector in particular had been known for several years for the sheer inexhaustible creativeness of its multiplicity of variant procedures and codes as regarded subject description, both in classified and in subject heading format. It is to the latter that we shall turn our attention in what follows. Many of these procedures were intended purely for local application and each of them was naturally better than any of the others. Time passed. The great majority of scholarly and public libraries in all the German-speaking countries use a unified scheme for subject heading, namely the RSWK (Rules for the subject word catalogue). It seems to me that this scheme , which has now been in use for over 15 years, established itself as the basis for the SWD (subject word authority file). The SWD has developed into the indexing basis of **the** universal thesaurus for the German-speaking countries, and as such has long since expanded beyond the limits of its original audience of university and public libraries. Scholarly research libraries have also had great success working with the SWD, if not always according to a strict interpretation of the RSWK.

The vocabulary of the subject descriptors of the SWD was aimed primarily at the monograph, which continues to be the most common form of publication. On the other hand, the different missions of our libraries leads us to the description of different types of publication. Thus, almost unnoticed by the great world of librarianship, periodical articles have long been described for regional bibliographies using the vocabulary of the SWD and the SWD's descriptors have been employed not only to document print publications but also for other media types such as museum artefacts or television broadcasts. This has demonstrated how wise those responsible for the development of the RSWK and SWD were, in concerning themselves with questions of terminology, above all in this case with the topics of pre- and post-co-ordination and the structured presentation of the language of documentation, to hold to the relevant standards of professionals from librarianship and documentation. This made possible the development of a language of documentation which corresponded to the state of the art in structure and content, and which could therefore necessarily be expected to attract interest outside the bounds of the world of librarianship. The vocabulary of the SWD should in principal be applicable to all types of documentation, it goes beyond provincial restrictions and in its universality transcends the profile of any individual library.

2. The problem of heterogeneity

What is now known globally as the information landscape makes it possible in theory to conduct searches from anywhere in any chosen bibliographic database, and for all types of document to be included in databases. These documents are also accessible according to the varying processes of the relevant subject description. The SWD has certainly the widest application in matters of linguistic description, but no monopoly. Specialist subject thesauri have the right to exist alongside it, having been built up over the years mainly by documentation centres of national repute. Even if the ideal situation were to have one single universal thesaurus used by all institutions creating subject listings within a given linguistic arena, you cannot expect renowned documentation centres to re-index all their stock. We should be content that libraries at least can all make use of the same language of documentation.

Let us have a quick look at such a specialised thesaurus and compare it to the SWD. As might be expected, we find numerous instances of agreement but also some discrepancies. Overall however, it can be said that there is practically no concept that cannot be described by one indexing vocabulary or another. The favoured expressions of any documentary language can therefore related to those of another either by congruence, 1:1 equivalence or 1:many equivalence, which includes "AND" or "OR " relationships.

BISP descriptors	SWD	
(Bundesinstitut fuer Sportwissenschaft = Federal Sports Institute)		
Abwehrtaktik	Abwehrtaktik	congruence

(defence tactics	defence tactics)	
Alltagsobjekt	Alltagsgegenstand	equivalence
(everyday object	everyday item)	
Abwehrschulung	Abwehr + Training	logical AND
Abwehr		
Abwehr		logical OR
Abwehrspiel		
(defensive play)		

It is also no surprise that there is no equivalent in a number of cases, the SWD is after all a universal thesaurus and a specific vocabulary is to be expected from a specialist thesaurus, even if there is nothing to hinder the SWD incorporating these descriptors providing certain conditions, which might be characterised as terminological compatibility, are met. On the other hand, the projects I should like to present in the following show that it is rather the SWD that contains the more specific and comprehensive subject vocabulary. Relating such different terminologies to each other was practically impossible only a few years ago, but today we have at our disposal techniques with which you may be fairly familiar through the internet, but which also have their problems and limitations. Whether you describe this desired process as a cross-concordance or a crosswalk is irrelevant in the end, to put it simply it is about creating links between equivalent terms describing similar concepts in the two thesauri, it is about the affiliation (I can think of no better word for it) of documentary languages. There needs to be a system interposed between the two thesauri, which manages the extant links, directs the amendment or introduction of new links and at the same times supports users in navigating between the separate data collections.

3. The projects

I should now like to sketch three projects, which have similar structures, two of which are monolingual (CARMEN, Economics cross-concordance), the third is multilingual (MACS).

3.1. CARMEN

One part of the CARMEN Project concerns itself with the association of the thesaurus of the Informationszentrum Sozialwissenschaften (IZT – Information Centre for Social Sciences) with the SWD. The method by which equivalencies are determined and links created is charming in its simplicity: starting from alphabetical lists which contain the keyword material from a specific subject area, the relationships between the two thesauri are determined intellectually. Not very long ago such a process would have seemed antiquated and everyone would have accepted that it was necessary to develop a new search engine, one better than any previous search engine, the" mother of all search engines". But the period of boundless euphoria about the capacity of such products seems to be past, , as is also the phase which might be described in the words of 1 Moses 10 as "so let us now go and build a search engine, so that we may make a name for ourselves and be not scattered". "The widely held belief that the lightning-fast progress of technology will, any minute now, present us with "search engines" that will at one stroke make all that horrible controlled language and standardisation totally unnecessary, belongs in the realm of fantasy or the language of salesmanship" (Christoph Wolters: Ist die Schlagwortnormdatei für die Objektdokumentation im Museum geeignet? [= Is subject heading authority control suited to documenting object in a museum] In: AKMB-News 1/1998). Within the parameters of this project we have also rejected the idea of generating automated comparisons by means of a process still to be developed, for whatever such a method might have been able to achieve, every descriptor that might for example have been identified mechanically as being exactly the same would still have required intellectual inspection, to confirm that the same letters actually have the same significance. Instead of this we have therefore set ourselves energetically to the intellectual work and in six months working with a half-time scholarly colleague have been able to process 3,500 keyword entries so that equivalent relationships between the thesaurus of the Informationszentrum Sozialwissenschaften, the thesaurus of the Deutsches Institut fuer Paedagogische Forschung (German Institute for Educational Research) and the SWD have been established and then recorded in a link management system.

There follows a sample page from the concordance between the Informationszentrum Sozialwissenschaften thesaurus and the SWD:

IZT thesaurus		SWD
Rollenwandel	<g	Rolle
Segregation	=h	Segregation
soziale Entwicklung	=+ h	Gesellschaft + Entwicklung
soziale Integration	= h	Soziale Integration
soziale Stabilität	< g	Stabilität
sozialer Prozess	= h	Sozialer Prozess
sozialer Wandel	= h	Sozialer Wandel
soziokulturelle Entwicklung	= h	Soziokultureller Wandel
sozioökonomische Entwicklung	= h	Sozioökonomischer Wandel
Systemveränderung	=o h	Systemveränderung
Systemveränderung	=o h	Politischer Wandel
Transformation	<o m	Politischer Wandel
Transformation	<o m	Sozialer Wandel
Unterentwicklung	= h	Unterentwicklung
Wertwandel	= h	Wertwandel
wissenschaftlicher Fortschritt	= h	Wissenschaftlicher Fortschritt

Even if you are not familiar with German, you should be able to see that true congruencies predominate. Since we are operating here in a monolingual environment, that is quite natural. Alternative forms to the different descriptors for the same concepts are given, however the subject specialists mainly choose the same common nomenclature. Overall the comparison of the thesauri found about 85% equivalence, of which 75% was congruent. It was also noticeable that the SWD clearly had a more comprehensive subject vocabulary that either of the two sets of subject terminologies.

3.2 Economics cross-concordance

As with Project CARMEN, the objective here is to establish links between descriptors in the Economics standard thesaurus and the SWD which are acknowledged to be equivalent. In contrast to CARMEN, however, here an automated check for duplicates will be carried out first. We are not sure how sensible this process will turn out to be, if it will spare us intellectual effort or if it will potentially lead to greater confusion. In any case it is important for us here to gain objective experience for subsequent projects instead of relying on subjective theories.

3.3 MACS

The year 1997 saw turning-point and a qualitative step forward in the ten-year history of the SWD. Recent years saw a repeated demand that the SWD be widened to include terms from other languages. In areas in which academic language and also the language of publication in German-speaking areas was largely determined by English, the English expressions had always been cross-referenced in the SWD and in the same way in matters of cultural study, it was usual to work with common Italian or French expressions. The inclusion of such cross-references was always intended to meet the needs of a German language SWD. So far no account was taken of the requirements of literature in foreign languages or the international exchange of subject description data in a global network of library information. The first efforts of the Subject Cataloguing section of the DDB to fill this information gap date back to the beginning of the previous decade. Thanks to a recent initiative of the Schweizer Landesbibliothek a start could be made to the MACS project (**M**ultilingual **Ac**cess to **S**ubject Headings). Working on this project alongside the Schweizer Landesbibliothek are the Bibliotheque Nationale of France, the British Library and the Deutsche Bibliothek. The aim of MACS is to study the links between the three extensive subject heading authority files – LCSH, RAMEAU and SWD.

The immediate objective of the project is to indicate in each authority file the equivalent preferred descriptors of the other authority files for a few chosen subject areas. For the time being the work is concentrating on the areas of sport and theatre, and on a selection of the most commonly used descriptors. The equivalencies between descriptors are being determined intellectually, as in the CARMEN project, with using automatic processes. We did not see it as part of our task to find a definitive solution to the problem of automatic translation by attempting to develop a programme that would identify the one-to-one equivalent expression for a particular word in another language. “Machine translation, does it exist? The only honest answer to this question is NO” (S. Krauwer: Machine translation: state of the art, trends and user perspective. In TELRI Proceedings of the first European Seminar “Language resources for language technology”. Budapest 1996.

sh85147274 Women	frBN001838184 Femmes	040182029 Frau
sh85147294 Women –Employment	frBN001521764 Femmes – Travail	*0401820292+ 04069349X Frau+ Berufstätigkeit*
sh85147456 Women authors	frBN001614243 Femmes écrivains	040533115 Schriftstellerin
sh85147587 Women in literature	frBN002484062 Femmes -- Dans la littérature	041136179 Frau <Motiv>
sh85148133 Work	frBN002059353 Travail	040025675 Arbeit
sh85148146 Work environment	frBN001764645 Conditions de travail	040026418 Arbeitsbedingungen
sh85073639 Working class	frBN001578199 Classe ouvrière	040687996 Arbeiterklasse
sh85148201 World history	frBN001555147 Histoire universelle	040791580 Weltgeschichte
sh85148236 World War, 1914-1918	frBN001617257 Guerre mondiale (1914-1918)	040791636 Weltkrieg <1914-1918>
sh85148273 World War, 1939-1945	frBN002351158 Guerre mondiale (1939-1945)	04079167X Weltkrieg <1939-1945>
sh85148515 World War, 1939-1945 Underground movements	frBN001634192 Guerre mondiale (1939-1945) -- Mouvements de résistance	*04079167X + 040792625 Weltkrieg <1939-1945> + Widerstand*
sh85149310 Youth	frBN001555624 Jeunesse	040288595 Jugend

It is not envisaged that a complete multilingual thesaurus will be created. That would require isomorphism, that is that equivalence would have to exist not just between preferred terms but also structurally, so that for each term in the hierarchy there would have to be a corresponding equivalent in each of the languages involved. For this the structures of the national authority files would have to be adapted to each other and any gaps there might be would have to be filled in. The complete final product would in the end be capable of replacing the national files, in that they were subsumed in MACS. This classical approach to a multilingual thesaurus has no chance of ever being created. Based on the practical experience of the MACS project and looking at the multilingual thesauri which already exist, one has to be greatly sceptical as to whether multilingual thesauri in the purest form, following the guidelines of the specialist national and international standards, are still meaningful, if the considerable effort that such a method of documentation would demand is still worth it.
The ultimate objective of the MACS venture is of course the inclusion of all subject areas. The next logical step after that would be to open it up to other languages.
The process being developed for MACS will not affect the structure of the individual national authority files, anymore than the IZT had to be modified during CARMEN. That does not exclude the possibility that during the process there may be alterations to the individual

thesauri, in so far as they concern matters of correction or increased precision, which would have happened sooner or later anyway and the qualitative improvement of individual pieces of data comes as a by-product of sorts. The user conducts the search in his native language or in one with which he is familiar and is then enabled to use the indicated equivalents to continue his search in affiliated databases. The technology required does not have to be invented from scratch, it is well known to us from use on the internet, the interface already exists (Z39.50) and all this can be put to use for the link between authority files and title data – provided that is available online.

4. User interfaces

Regardless of the type of document which is recorded in a database, whether it is a conventional or an electronic publication, if it is films, museum objects or archive material that is being dealt with, if it is recorded in a "classical" title entry or with Dublin Core, if it is available as full-text or only as excerpts, the problem facing the user is the same. The situation is still that there are two contrasting ways of organising knowledge. Between the user on the one hand and the bibliographic database on the other, a thesaurus functions as a verbal link between these two systems and along with the order in which the database is organised, fulfils two functions:

1. support of user queries
2. structured navigation in the data elements.

MACS uses intellectually-determined equivalencies to link the content of bibliographic databases which use a controlled vocabulary to describe their content and present in an ordered, structured way. I do not therefore see it as particularly helpful, perhaps even as a retrograde step, if such data is overlaid by an automated process using electronic dictionaries – which might be one possible alternative to what we are doing. For MACS, that would imply that the clear, comprehensible classification which is present in each of the systems would be dissolved, and the user would therefore have to navigate from ambiguity to ambiguity.

Zug -> English: groove, pull, course, train, convoy drift, flight, tendency, draught, draft
Draft -> German: Tratte, Wechsel, Abteilung, Aufgebot, Aushebung, Nachschub, Entwurf, Skizze, Konzept, Zug
Wechsel -> French: échange, retraite ...
Retraite -> German: Einsiedelei, Rückzug, Ruhestand, Wechsel, Zapfenstreich ...

If the user has a perfect understanding of the subject terminology in the foreign language, then he has no need of such a product, and if he does not know it well, or is unsure, it will not help him. What could be of use is a link from the preferred term of one thesaurus to its equivalent in another. The following is a deliberately simple example demonstrating step by step a search using SWD via MACS to search the data of the Bibliotheque Nationale de France: let us assume that the user is looking for information about cycling in general, not being sure of the terminology he begins with the keyword Radfahren (Cycling), which leads him to the descriptor Radsport (cycle racing), he would likewise have got there if he had begun his search in the SWD with the word Fahrradsport (bicycle racing) or just Fahrrad (bicycle) (search terms are boldened).

Fahrrad
Q: M
SYS 31.7
BF Velo
OB Zweirad
VB **Radfahren**
VB Velomobil
Radsport
Q: Sport-B

SYS 34.3
BF **Fahrradsport**
VB **Radfahren**
VB Radsportler

After the user has looked at the titles in the DDB, he finds the link to MACS, where I assume that he will receive a short, easily understandable communication about what it is about and what to expect. So with curiosity he presses the button and moves to the following page, which would be presented to each user in his own language:

Suchwort eingeben (enter search term)
radsport
Bemerkung: Es werden nur Schlagwörter gesucht (Note: only subject heading are searched)
Gewählte Sprache /Sprache auswählen)Chosen language/select language)
Deutsch (SWD)
Gewählte Bibliothek(en) /Bibliothek(en) auswählen (Chosen library/select libraries)
Swiss National Library
Bibliothèque nationale de France
Die Deutsche Bibliothek
The British Library
Sie können eine oder mehrere Bibliotheken auswählen (you may choose one or more libraries)
Search **Suche in den ausgewählten Bibliotheken**
Browse **Zeige die verknüpften Übersetzungen an**

As our cycle racing fan knows how important this kind of sport is in France, he is particularly interested in what the BNdF has to offer on this subject:

Ihre Suchergebnisse in den Bibliotheken...(Results of your search from libraries …)
Bibliothèque nationale de France: 31 hits.

1. Il était une fois le cyclisme à Corbeil-Essonnes; un siècle d'histoire; Guy Caput, Roland Oberle;
2. Guide du cyclisme; manuel pratique et conseils de santé; docteur Gérard Porte;
3. Le livre d'or du cyclisme; 1995; Jean-Luc Gatellier; préf. de Laurent Jalabert;
4. L'année du cyclisme, 1996; Pierre Chany et Claude Droussent;
5. Le livre d'or du cyclisme; 1996; Jean-Luc Gatellier; préf. de Richard Virenque;
6. Cyclisme; 56 exercices et programmes; C. Carmichael, E. R. Burke; [trad. par GwénaÈel Hubert];
7. Cyclisme; les conquérants de l'Arc-en-Ciel, 1927-1996; les champions du monde; par Claude Dassonville;
8. L'année du cyclisme 1997; Claude Droussent;
9. La fabuleuse histoire du cyclisme; Pierre Chany;
10. Le livre d'or du cyclisme 1997; Jean-FranÐcois Quénet; préf. de Marc Madiot; photos, Graham Watson;

Granted that this a simple example, but since our experience so far shows that in over 60% of all cases there is indeed a 1:1 equivalent, I do not think the example is so far removed from reality.

A structured search using the preferred terms of a familiar system is always easier to manage than browsing in a free text environment. The following example illustrates the browsing facility in MACS. The user would like to see all the subject headings in which the (polyvalent) word “Training” occurs.

Type in the subject
training
In the chosen language

Français (RAMEAU)
In the chosen Libraries
Swiss National Library
Note: Only subject headings are included in the search
Bibliothèque nationale de France
Die Deutsche Bibliothek
The British Library
Tip: You can select or deselect the libraries you want to search
Search **To search the selected libraries**
Browse **To view possible translations**
English (LCSH) Deutsch (SWD) Français (RAMEAU)
All terrain cycling -- Training
Geländeradsport + Training
Vélo tout terrain + Entraînement
Cycling -- Training
Radsport + Training
Cyclisme + Entraînement
Employees -- Training of
Mitarbeiterschulung
Personnel --Formation

By using the “Browse” command he receives the list given in excerpted form above, it is generated automatically, whereas the links it contains are decided intellectually. The list demonstrates the problems that an inter-thesaurus must deal with, namely how to handle 1:many equivalencies which arise from the differing terminological practice of the documentation languages involved. The more the terminology is based on lexicographical principles, the easier it is to make the links. Since in the case of LCSH and to a lesser extent in that of RAMEAU, we are dealing very much with a pre-co-ordinated vocabulary, there is frequently the need to use several SWD descriptors along with AND or OR logical operators in order to establish an equivalent. This is also one way to improve situations where there only a rough correspondence exists.

If more languages are to be included, then managing the links will become commensurately more complicated – which is equally true for the linking of German-language thesauri to the SWD. In order not to fall victim to an explosion of combinations, we might consider associating new documentation languages only with the one language of their choice, preferably the language which is most similar to them as far as pre- and post-co-ordination is concerned.

MACS is based on the assumption that the user accesses the results of intellectually assigned subject descriptions via a thesaurus. We are dealing with bibliographic databases of considerable size. Neat hierarchies, such as you may come across in Yahoo, say, cannot cope with this mass of information, thesauri can distinguish to a better and more comprehensive degree between material to be indexed than a method based on syntactical indexing. If the number of documents available grows even more than the number of affiliated projects dealing with heterogeneous material, it should be straightforward to deal with, since our method of indexing already envisages a structured user interface as an inherent part of our service.

5. A brief summary

The three projects show that it is possible to create links between heterogeneous data sets with manageable demands on intellectual resource and without great technological demands. We have seen how sets of content described using different documentation languages can be related to each other. The prerequisite for such a process is that the data sets are extensive and significant in scope with a fully elaborated documentation language. The purpose of our endeavour is not to support the private terminology of small institutions or to contribute to the construction of parallel developments.