IFLA 2005 OSLO

World Library and Information Congress: 71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

June 07, 2005

Code Number: 044-E
Meeting: 139 SI - Classification and Indexing with Cataloguing

Multilingual Subject Access to Catalogues of National Libraries (MSAC) : Czech Republic's collaboration with Slovakia, Slovenia, Croatia, Macedonia, Lithuania and Latvia

Marie Balikova
National Library of the Czech Republic
Prague, Czech Republic

Abstract:
Czech authority file of topical terms is intended to form a base for multilingual controlled vocabulary. The aim of the proposal is to provide users of online library catalogues and internet services of cooperating institutions with an indexing and retrieval tool which enables multilingual and cross-domain searching ("one-stop" seamless searching). The goal of the project is to establish a multilingual subject approach to catalogues of participating libraries (Czechia, Croatia, Latvia, Lithuania, Macedonia, Slovakia, and Slovenia). In practice this means that a user in any of these countries would enter a query in his local language and receive hits from all the catalogues. The initiative is complying with the main goals currently defined by IFLA for the activity of Indexing and Classification Section, it means: Changing Roles of Subject Access Tools (Berlin), Implementation and Adaptation of Global Tools for Subject Access to Local Needs (Buenos Aires), and Cataloguing and Subject Tools for Global Access: International Partnerships (Oslo).

Czech authority file of topical terms is intended to form a base for multilingual controlled vocabulary. The aim of the proposal is to provide users of online library catalogues and internet services of cooperating institutions with an indexing and retrieval tool which enables multilingual and cross-domain searching ("one-stop" seamless searching).

The goal of the project is to establish a multilingual subject approach to catalogues of participating libraries (Czechia, Croatia, Latvia, Lithuania, Macedonia, Slovakia, and Slovenia). In practice this means that a user in any of these countries would enter a query in his local language and receive hits from all the catalogues.

The initiative is complying with the main goals currently defined by IFLA for the activity of Indexing and Classification Section, it means: Changing Roles of Subject Access Tools (Berlin), Implementation and Adaptation of Global Tools for Subject Access to Local Needs (Buenos Aires), and Cataloguing and Subject Tools for Global Access: International Partnerships (Oslo).

It is a cooperative venture of three large libraries in Czechia: National Library of the Czech Republic (NL CR), Moravian Library and Olomouc Research Library (Czech National Subject Authority File- CZENAS) and National Libraries of cooperating countries (Multilingual Subject Access to Catalogues of National Libraries (MSAC).

The aim of this initiative is to provide the users of online library catalogues and information gateways of cooperating libraries with a prototype for multilingual subject searching in online environment. Library collections of these libraries are large and without any doubt very valuable for researchers throughout Europe. What is needed is a standardized, authorized indexing and retrieval tool which would bring together all their catalogues and databases and enable multilingual subject searching.
At the beginning of the project, a number of factors affecting subject indexing in current environment and cross-searching for subjects have been identified.

These factors include

- standardization of subject retrieval process and indexing and classification tools
- subject retrieval methods
- possibility of interoperability among different indexing and classification schemes
- multilingualism issue
- possibility to increase precision and recall trough Z39.50 protocol and its profiles and to apply authority control in subject retrieval process
- need for cooperation

Standardization

It has been agreed that the standardization applied in subject analysis area should

- minimize duplication of work in sharing information in a networked environment
- ensure consistency and compatibility of subject access points used by different national agencies
- support shared cataloguing process at national level
- make it easier for end-users to overcome retrieval difficulties in cross-domain searching process at national and international level

Subject retrieval methods: browsing and searching

Browsing enables a user to view all the records listed under a particular subject term. Browsing method can be useful for inexperienced users who do not have any specific search term in mind or are not familiar with the subject field.

Searching enables a user to enter the search terms and retrieve records containing those terms. Advanced searching techniques such as field-searching, phrase searching and the use of Boolean operators could be allowed. It has been suggested that in MSAC initiative both retrieval methods should be supported.

Interoperability

The need for interoperability among different indexing and classification systems and its usefulness for the end-users have been identified as follows:

The information resources accessible on Internet – the library catalogues and digital collections of cooperating institutions included - are heterogeneous and have been indexed with different vocabularies and organized according to different tools and schemes. The best solution is to prepare for the users a “one-stop” seamless searching by creating a single search apparatus instead demanding from the user to search individual databases or collections separately (Chan and Zeng 2002).

Regarding the methods used for achieving and improving interoperability, there were mentioned

- intellectual mapping, which consists of establishing equivalents between terms in different controlled vocabularies or between verbal terms and classification numbers
- using of a switching language, which could serve as an intermediary for moving among equivalent terms in different vocabularies, above all multilingual

Regarding the methods used in the link storage and management it was declared that special fields in authority records/formats may be used to store and maintain links between indexing terms and their complex relationships.

Authority control

It has been declared that the authority control of subject search terms is essential, because it improves subject access dramatically by providing consistency in the form of headings used to identify the subjects.

Z 39.50 and its profiles

The use of network communication tools like Z39.50 between systems has changed the environment we operate in from a local to an international one. When using Z 39.50 communication both precision and recall will improve, provided that the authority control is applied whenever possible – in all databases searched through, introducing the same subject search criteria both in remote and local databases.

Multilingualism issue in online environment

Multilingualism is a complex issue. Users may want to search a multilingual collection by using queries in one language or to retrieve documents in a number of specific languages, preferably also via an interface in the language of their choice. In some cases they may require some translation or summary in another language than that of the document. The best solution seems to be that the users are provided with the language support they need. However, multilingual access to information will be limited as it requires sophisticated technologies, the language skills of the staff involved in cataloguing process, and immense financial means. Therefore, there have been only few attempts to create a multilingual subject access tool or to integrate already existing library systems in the area of multilingual subject access. One of the most important is the Multilingual access to subjects (MACS) Project. The goal of this project is to integrate the most developed and used subject indexing systems Library of Congress Subject Headings (LCSH), Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU) and Schlagwortnormdatei (SWD). The feasibility of linking mentioned Subject Heading Languages to provide multilingual access to their collections was investigated; the approach by creating links between LCSH, RAMEAU and the SWD/RWSK was tested in the fields of sport and theatre. A prototype was created, and the ways how to extend the use of MACS project has been discussed. It is felt that MACS may show users and other partners how the language barrier may be crossed and encourage further efforts, either in adding new subject indexing systems or investigating the use of other tools such as classifications if the same are available across several institutions. It is recognized however that the creation and management of links may demand significant resources (Clavel 2003).

The possibility to join the MACS project had been discussed; however, it appeared that there were several obstacles, such as

- lack of authorized national indexing systems in some countries
- substantial differences between MACS and national indexing systems, if existed
- lack of a common “LCSH” in Central and East European countries
- critical lack of financial resources

Recently a new discussion is going on, mainly at the level of CENL, about the possibility and ways of mutual cooperation of the MACS and MSAC project. It has been agreed that the feasibility of merging MACS and MSAC would be investigated.

Need for cooperation

A multilingual organizing and retrieving tool could be created only in collaboration with other national libraries. The responsibilities for adding equivalents in participating languages should be divided among them.

Subject analysis process in online environment

In order to make the subject analysis process more efficient, easier and faster in current online environment, it seems to be useful

- to prefer post-coordinated indexing system,[1] which can provide the same precision as the pre-coordinated one, when including equivalent notation of a classification scheme into the retrieval process to avoid “false drops”. In addition, a notation of classification scheme gives a context to the verbal search term.
- to simplify application syntax - to minimize the structure of subject headings strings

- to facilitate automatic validation of subject headings and automatic maintenance of subject authority files
- to reduce the size of authority files
- to support conceptual compatibility of indexing formulas/preferred terms used in various indexing languages
- to support harmonisation between various indexing languages
- to support mapping between verbal terms (indexing formulas) and equivalent notations of classification scheme
- to improve hierarchical structure of subject authority file
- to make the assignment of controlled terms more efficient
- to enhance and encourage co-operative cataloguing efforts
- to improve subject access for OPACs and for Web resources

Classification systems in on-line environment

The role of classification systems in on-line environment has been discussed as well. The role of classification in flexible and interactive retrieval systems is to serve as an underlying knowledge structure that provides systematic subject organisation and thus complements the search using natural language terms.

Universal classification scheme applied in networked environment can enhance subject access, because it

- covers all subjects
- is able to create collections of related resources in a hierarchical structure quickly and easily
- improves subject access to large databases using sophisticated methods
- provides context to search terms
- supports interoperability between information systems
- enables multilingual access to collections
- enables language independent notations to be linked to search terms of various verbal languages
- enables to search in more than one language at the same time
- enables other languages to be joined later without the need to classify the resources again
- could serve as reference or switching language, which ensures convertibility between information languages
- is able to provide the same level of specificity in all participating languages
- offers language independent coding

Universal Decimal Classification (UDC)

On the Internet where information flows freely in all formats across national and linguistic boundaries, it is important to have a standard knowledge organization system that can represent concepts in a language-independent fashion. UDC, with its worldwide user community, meaningful notation, well-defined categories, well-developed hierarchies, and rich network of relationships, is such a system. UDC classification represents a universal synthetic, faceted classification which can be adopted and used as mapping mediator between indexing controlled terms of participating languages.

It is based mostly on numerical notations and uses language independent coding. The scheme UDC MRF is available among others in English and Czech languages and in machine readable form. It is flexible more than other universal classification schemes; it supports very detailed expressions of complex subjects using a variety of common and special auxiliaries, specific symbols and punctuation. Another useful feature of the UDC is its ability to indicate entities which occur in more than one domain (class). Concepts such as Incest may occur in sociology, criminal law, psychiatry; Water in inorganic chemistry, hydrology, water management.[2]

MSAC and UDC

As a result of initial discussions about the feasibility of a multilingual project (MSAC), only UDC system appeared to be the most suitable for creation of a multilingual common indexing tool. In addition, the positive aspect was that all the participating libraries used it, even if in different versions.

In MSAC, like in other subject gateways and portals on the Internet using UDC, this classification system is applied as an enumerative classification. The UDC numbers – single and complex (pre-combined) are treated as single UDC numbers (set of characters without meaning). When applied as an enumerative, non-synthetic classification, UDC has functionality very similar to that of DDC, the only difference being that UDC has a more comprehensive and more specific vocabulary and does not contain as many enumerated, in advance prepared compound terms as is the case with DDC. With the trend of present revisions moving UDC towards more faceted structure, the need to use some kind of combination of numbers may appear more frequently. This is the case, for instance, at 821 Literature and 94 History. To obtain the number for “national” literature, the number 821 for literature has to be combined with the common auxiliary for language, e.g. 821.162.3 Czech literature. Similarly, for history of individual countries the number for history 94 and common auxiliary for place to denote the country, e.g. 94(437.3) History of Czechia, have to be used. The use of UDC as an enumerative classification (either with simple numbers or with pre-composed numbers which are treated as simple) may well serve its main purpose if class number captions (descriptions) are added to the retrieval system so that beyond numbers terms are available for search and are added to the systematic display at the end-user interface (Slavic 2003).
In MSAC system UDC class numbers are used alongside their descriptions.

Example of UDC index (English and Czech equivalent of UDC class numbers)

602.44 -- biotransformation / biotransformace
602.6 -- gene engineering / genové inženýrství
602.6 -- genetic engineering / genetické inženýrství
602.6 -- transgenosis / transgenoze
602.641 -- viral vectors / virové vektory
602.7 -- cloning / klonování
604.4 -- secondary metabolites / sekundární metabolity
604.6 -- genetically modified organisms / geneticky modifikované organismy
606:616-056 -- gene therapy / genová terapie
608.1 -- bioethics / bioetika
608.3 -- biological safety / biologická bezpečnost

608.34:663/664 -- genetically modified foods / geneticky modifikované potraviny

Citation order of compound/complex UDC notations

One of the advantages of the UDC is the facility to adapt the citation order to fit in with local requirements. However, international exchange of information demands consistency in building UDC class numbers, therefore the same citation order should be adopted.

UDC MRF in electronic form

The precondition for creation of a subject multilingual retrieval tool based on UDC classification to be applied in online environment is that all participating libraries have their national language version of UDC MRF in electronic form. This prerequisite, unfortunately, have not been fulfilled in all respective countries, yet. Therefore, a special solution has been accepted: the language equivalents of controlled terms created by participating libraries are being added to the Czech Subject Authority file.

Czech National Subject Authority File - CZENAS

Subject authority file of NL CR is an integrated indexing and retrieval tool in which verbal controlled terms are being linked to the UDC equivalent notations. When creating the subject authority file we respect IFLA recommendation - to consider possible relationships between subject authority records and classification.

The Czech National Subject Authority File consists of three specific authority files: geographic, genre/form and that of topical terms.

Czech authority file of topical terms serves as a base for multilingual controlled vocabulary.

Topical authority file is a thesaurus in which following kinds of relationships between terms are defined: equivalence (expressed: USE), hierarchy (expressed: BT-Broader term; NT-Narrower term) and association (expressed: RT-Related term).

Equivalent relationship is identified between preferred terms and its variants (synonymous, variant spelling forms, alternative forms, in some cases opposite terms and specific narrower terms).

Hierarchical relationships can be established between valid (preferred) controlled terms, which form part of the same semantic domain and are connected by reciprocal NT, BT references.

Associative relationship can be identified between terms which are related other than hierarchically; the relationship is reciprocal.

The authority record includes among others:

089	UDC class number and UDC explanatory term
150	Main heading – topical term heading

450	Cross-References (or Used For References – UF)
550	Broader term (BT) Narrower term (NT) Related term (RT)
750	Linking entries - Equivalent of topical term headings in different languages and source of heading or language code

Mapping of indexing terms, UDC numbers and English equivalents

Controlled vocabulary structure is tied to a classification scheme so that relationships between indexing terms can be expressed more definitely.

Mapping process between Czech verbal expressions and UDC numbers is being done intellectually. Candidates of controlled terms are being chosen with document in hand (from bottom up) in order to suggest terms as specific as needed (not as specific as possible). Single or complex UDC numbers (pre-combined) are being linked, English equivalents of preferred terms, mostly LCSH terms are being chosen. Sometimes, we are not successful in finding LCSH equivalents since the LC terms are too broad: in this case, the reference sources like LC titles and subtitles file, encyclopaedias, manuals, language vocabularies, www pages, full text databases are consulted. The proposals of preferred terms linked to the UDC class numbers and English equivalents are being sent to the editorial staff for approval, then the approved authority records are entered via special programme procedure into the authority database.

Mapping process

As already mentioned, mapping process of Czech verbal expressions, UDC numbers and languages equivalents are being done intellectually. Czech National Library team is responsible for UDC notations and English equivalents.

On the other hand, the participating libraries are responsible for subject terms expressed in their native language. Mapping process of English verbal expressions, UDC numbers and local language equivalents is being done intellectually as well. Cooperating libraries receive special tables of preferred terms in English accompanied by UDC numbers; they add data in their language, then verbal terms are incorporated in authority records as equivalent headings entered in fields 7XX MARC 21.

Principles of mapping

The proposal consists in establishing equivalents between the subject controlled terms used in indexing systems of participating libraries through a switching language.

The switching language represents UDC notations based on UDC MRF and English equivalents. The mapping links are defined between preferred terms represented by isolated lexical units only. The subject headings strings as a whole are excluded, are not mapped. The authority records as a whole are excluded, are not mapped; the links are established only between topical main headings (main entries) and language equivalents.

Special language based indexes have been defined

Subjects-Czech
Subjects-English
Subjects-Croatian
Subjects-Latvian
Subjects-Lithuanian
Subjects-Macedonian
Subjects-Slovak
Subjects-Slovenian

The subject authority file can be browsed by controlled terms, English and other participating languages equivalents and UDC notations.

The authority file is searchable by subject terms and UDC class numbers using CCL Search module. Boolean operators (AND, OR), combination of the first element of UDC notation and right truncation is allowed.

MSAC is being developed in two phases. The first phase includes the development of Czech topical authority file and the integration of language variants of participating libraries in Czech subject authority file.

In the second phase of the project the combinations of UDC-natural language verbal expressions or UDC-English expressions (or both of them) are to be inserted into the special fields of respective bibliographic records in the databases of cooperating libraries. The process could be done (semi)automatically, intellectual checking of data is supposed.

Cooperating libraries should provide access via Z39.50 protocol. All the libraries that have got a working Z39.50 server will be integrated directly, while for those not supporting this protocol a small testing database can be created at the NL CR. NL CR can offer access to databases of cooperating libraries via one single interface in the Uniform Information Gateway (UIG). After joining the system, UIG users can (after accomplishing the procedure of authorization and authentication) work in their own environment. They can work in both Czech and English environment and can use both Czech and English languages for searching.

Future development

The Internet European multilingual community uses more than 30 languages, represented by many character sets with different repertoires and encodings.
The idea to create a multilingual subject retrieval tool or to introduce a mapping scheme in existing systems is considered as an essential element of The European Library service.

We are still at the very beginning of the MSAC project. The biggest problem is that the creation of MSAC is dependent on voluntary work of teams of participating libraries and that communication takes place almost only via e-mails. Until now there has been no external financial support for the project.

A new perspective for future development of the MSAC project is appearing in joining the TEL-ME-MOR project (The European Library: Modular Extensions for Mediating Online Resources) funded by the European Commission under the Sixth Framework Programme of the Information Society Technologies (IST) Programme, where the ten new member states of European Union have been invited.[3]

[1] Postcoordination is a combinnation of elements by a searcher at the time he/she looks for materials. Individual terms are assigned to specific works, and the searcher bears the burden of combining terms for the topics deemed pertinent

[2] Examples:

Heading	|a water
UDC	|a 546.212 |c anorganic chemistry
UDC	|a 556-032.2 |c hydrology
UDC	|a 628.1.03 |c water management

Heading	incest
Seen from	krvesmilstvo [o]
Broader term	psychosexuální poruchy
	sexuální trestné činy
Related term	psychopatologie
	sexuální morálka
UDC	316.835.2 (sociology)
	343.542.5 (criminal law)
	616.89-008.442.38 (psychiatry)
English	incest
Lithuanian	kraujomaiša (baudžiamoji teisė)
	kraujomaiša (psichiatrija)
	kraujomaiša (sociologija)
Macedonian	Incest (sociologija)
	Incest (krivično pravo)
	Incest (psihijatrija)
Slovak	incest (sociológia)
	incest (trestné právo)
	incest (psychiatria)
Slovenian	Incest (kazensko pravo)
	Incest (psichiatrija)
	Incest (sociologija**)**

MARC 21 format

089	|a 316.835.2 |c sociologie |d sociology
089	|a 343.542.5 |c trestní právo |d criminal law
089	|a 616.89-008.442.38 |c psychiatrie |d psychiatry
150	|a incest
450	|a krvesmilstvo |0 o
550	|a psychopatologie
550	|a sexuální morálka
5509	|w g |a psychosexuální poruchy
5509	|w g |a sexuální trestné činy
75007	|a incest |2 eczenas
75047	|a Incest |b 316.835.2 |c sociologija |2 mac
75047	|a Incest |b 343.542.5 |c krivično pravo |2 mac
75047	|a Incest |b 616.89-008.442.38 |c psihijatrija |2 mac
75057	|a incest |b 316.835.2 |c sociológia |2 slo
75057	|a incest |b 343.542.5 |c trestné právo |2 slo
75057	|a incest |b 616.89-008.442.38 |c psychiatria |2 slo
75037	|a kraujomaiša |b 343.542.5 |c baudžiamoji teisė |2 lit
75037	|a kraujomaiša |b 616.89-008.442.38 |c psichiatrija |2 lit
75037	|a kraujomaiša |b 316.835.2 |c sociologija |2 lit
75067	|a Incest |b 343.542.5 |c kazensko pravo |2 slv
75067	|a Incest |b 616.89-008.442.38 |c psychiatry |2 slv
75067	|a Incest |b 316.835.2 |c sociologija |2 slv

[3] References

Chan, Lois Mai and O'Neill, Edward T. (2003) *Fast (Faceted Application Of Subject Terminology) : A Simplified LCSH-Based Vocabulary.* [Paper Presented at the 69th IFLA Council and General Conference, Berlin, 2003]

Slavic, Aida (2003) *UDC implementation: from library shelves to a structured indexing language*. [Paper Presented at the 69th IFLA Council and General Conference, Berlin, 2003]

Howarth, Lynne C. (2003) *Metadata Schemas for Subject Gateways*. [Paper Presented at the 69th IFLA Council and General Conference, Berlin, 2003]

Clavel-Merrin, Genevieve. (2003) *National libraries as access points: the role of TEL and MACS*. [Paper Presented at the 69th IFLA Council and General Conference, Berlin, 2003]

Chan, Lois Mai and Zeng, Marcia Lei (2002) *Ensuring Interoperability among Subject Vocabularies and Knowledge Organization Schemes: a Methodological Analysis*. [Paper Presented at the 68th IFLA Council and General Conference, Glasgow, 2002]

Freyre, Elisabeth and Max Naudi. (2001) MACS: *Subject access across languages and networks*. *In Subject Retrieval in a Networked Environment* [Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on InformationTechnology, OCLC, Dublin, Ohio, USA, 14-16 August 2001]. Dublin, OH: OCLC.

Hudon, Michele. (1997) *Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge concepts*. In *Knowledge Organization* 24(2): 84-91.
IFLA Section on Classification and Indexing. (2001) *Newsletter* Nr.24, December 2001.

Koch, Traugott, Heike Neuroth, and Michael Day. (2001) *Renardus: cross-browsing european subject gateways via a common classification system (DDC).* Available at http://www.lub.lu.se/tk/renardus/preifla-final.html (last accessed February 25, 2005)

GERHARD - German Harvest Automated Retrieval and Directory. Available at http://www.gerhard.de/gerold/owa/gerhard.create_index_html?form_language=99

DESIRE Information Gateways : *Handbook* Available at http://www.desire.org/handbook/

Kuhr, Patricia S. (2001) *Putting the world back together: mapping multiple vocabularies into a single thesaurus.* In *Subject Retrieval in a Networked Environment* [Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001]. Dublin, OH: OCLC.

Olson, Tony. (2001) *Integrating LCSH and MeSH in information systems*. In *Subject Retrieval in a Networked Environment* [Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001]. Dublin, OH: OCLC.

Riesthuis, Gerhard J.A. (2001) *Information languages and multilingual subject access*. In *Subject Retrieval in a Networked Environment* [Papers Presented at an IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, Dublin, Ohio, USA, 14-16 August 2001]. Dublin, OH: OCLC.

Robert P Holley, Dorothy McGarry, Donna Duncan and Elaine Svenonius ed. *Subject Indexing: Principles and Practices in the 90's;* [Proceedings of the IFLA Satellite Meeting Held in Lisbon, Portugal, 17-18 August 1993, and Sponsored by the IFLA Section on Classification and Indexing and the Instituto da Biblioteca Nacional e do Livro, Lisbon, Portugal] München: K.G. Saur, 1995.