

## **Models for multilingual subject access in online library catalogues : the ILO experience**

**Ron Davies**

International Labour Organization  
Geneva, Switzerland

### **Abstract**

*Different models exist for providing users of library catalogues with subject access in the language of their choice. One model relies on online processing to translate subject terms at the time the catalogue record is accessed for searching, for display or for export. Another, less frequently discussed model uses pre-processing, at a point in time before the catalogue record is available, to support multilingual subject access. This article discusses the experience of the International Labour Organization, where systems in use now or in the very recent past illustrate both models. The advantages and disadvantages of each model are summarized.*

### **Introduction**

Many libraries serve clients who speak a variety of different languages. Libraries located in officially bi- or multi-lingual countries, libraries in areas with large immigrant populations and libraries of international organizations must provide multilingual subject access to their online public-access catalogues (OPACs) in order to serve their varied clientele. There are, however, at least two different models for supporting multilingual subject access in a modern OPAC. In the first model, online processing at the time the catalogue is accessed provides access in different languages to the records in the library catalogue. In the second model, processing to support multilingual subject access is done offline, in batch, at some point in time before the user approaches the computer to do an OPAC search. This article describes both models by means of a case study of library systems formerly or currently in use at the International Labour Organization in Geneva, Switzerland, and discusses the advantages and disadvantages of each.

First, what is multilingual subject access? In a simple unilingual library catalogue, subject headings or descriptors are entered in the language of the user population— only subjects in that language that can be retrieved and displayed. A French book catalogued by a library in London, for example, can normally only be found using subject headings in English, just as a Spanish book catalogued by a library in Paris must be searched for using subject headings in French and an English book in a library in Madrid would usually be retrieved using subjects in Spanish. In a multilingual English/French/Spanish environment, however, we need to allow users to search using subjects expressed in *any one* of those three languages and to retrieve *all* books dealing with that subject. For example, three users searching in the library OPAC for information on employment should be able to perform a subject search for "employment", "emploi", and "empleo" respectively and to retrieve exactly the same documents. When those users display their search results, they should see the subject terms in the library catalogue record displayed in the language of their choice, e.g. a user searching

---

for the French term "emploi" should see French headings displayed in the record, not "employment" and other subject terms in English or in Spanish. Finally, when the library shares its catalogue records with other institutions by exporting them to other library catalogues, union catalogues or commercially published databases, subject terms in all languages should be available to these other systems. These different functions-- search, display and export-- are the basic functional requirements for multilingual subject access.

### **Case study**

The Library of the International Labour Organization (ILO) has a long history of multilingual support. The ILO, the oldest and one of the largest of the United Nations' specialized agencies, has seven official languages<sup>i</sup>, of which English, French and Spanish are considered working languages. The Library serves a worldwide clientele, including the constituents of the organization (the governments of 176 member states as well as trade unions and employers' organizations); researchers in labour and social issues; and more than 1900 staff in headquarters in Geneva Switzerland and in 40 offices around the world. *Labordoc*, the Library's catalogue and bibliographic database, contains over 350,000 bibliographic records. Records in *Labordoc* have been indexed since 1965 using the *ILO Thesaurus*<sup>ii</sup>, a multilingual thesaurus in English, French, Spanish and German.

From 1980 until late 2002, the ILO Library's catalogue was automated with the MINISIS software developed by the International Development Research Centre, a Canadian development agency. MINISIS was very suitable for the ILO because of its support for searching using a multilingual thesaurus and its ability to call user-developed software routines that could extend multilingual support to other functions. The multilingual subject access implemented in MINISIS at the ILO was based implicitly on the "online processing" model because all processing took place at the time the catalogue was accessed. While there are many descriptions in the literature of systems that have implemented at least the search functionality of this model<sup>iii</sup>, the following outline of the way in which this model was implemented in the MINISIS software at the ILO is typical.

### **Online processing model (MINISIS)**

In the MINISIS application at the ILO, subject descriptors from the *ILO Thesaurus* were entered in bibliographic records in one language only, English<sup>iv</sup>. The MINISIS software validated the descriptor entered against the values for that language in the thesaurus authority file. When it came time to search for a subject in the OPAC, the MINISIS software recognized that a subject search was called for and invoked special query expansion processing. The search term in the query was not searched directly in the index; instead it was looked up in the subject authority file (thesaurus file) in order to retrieve the equivalent descriptors in all the languages of the thesaurus. Each of the different language forms of the descriptor (including the original descriptor as entered by the user) were then searched separately and the result sets from each search were combined with a Boolean OR before the final, combined result set was returned to the user (Fig. 1)<sup>v</sup>. Therefore even entering a search strategy in French or Spanish, all documents relevant to the concept in *Labordoc* could be retrieved.

---

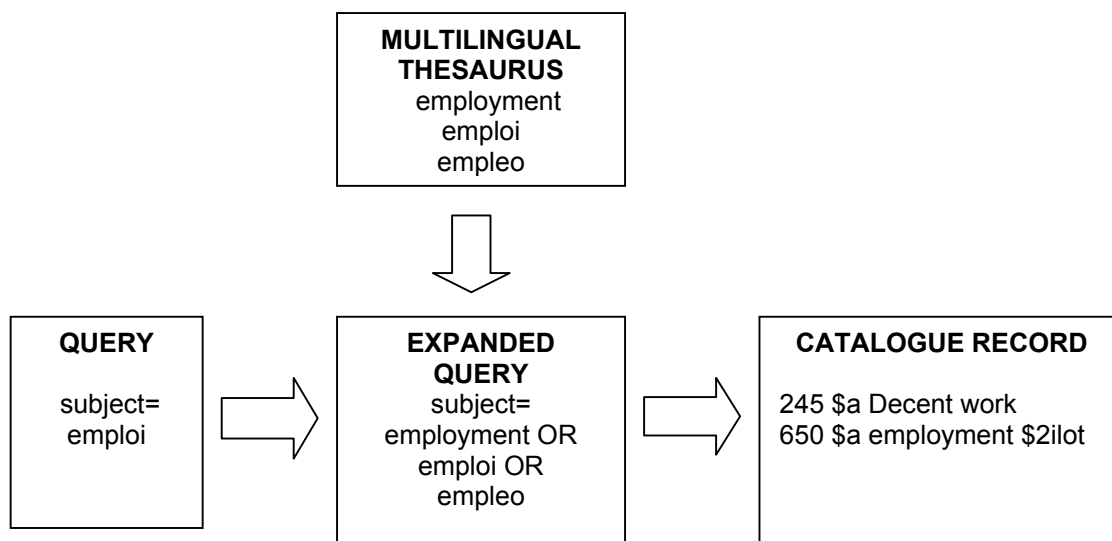


Fig. 1 - Query expansion in the online processing model

While the basic MINISIS software supported searching in other languages, MINISIS did not have the ability to display descriptors in languages other than in the language in which the descriptors were entered. However MINISIS did offer the ability to define a variety of different display formats and to call user-developed software routines at various points during its processing cycle. This allowed the ILO to develop software that would also display subject descriptors in the language the user preferred, independent of the language in which it had been entered. Whenever a user wanted a record displayed in a language other than English, a particular format for the language was invoked. When that format indicated that the subject field was to be displayed, the subject descriptor was passed off to the user-written routine, which looked up the descriptor in a special translation file, found the corresponding descriptor in the appropriate language, and passed that replacement value back to MINISIS for display. Similar external processing allowed descriptors in other languages to be added as locally defined fields in a record when it was exported from the database and sent to partner institutions, union catalogues and re-publishers of *Labordoc*<sup>vi</sup>. These other institutions could then index or display the subject descriptors in whatever language or languages they chose, independently of whether their software supported the online processing model or not.

### **Transition to a new system**

While the MINISIS software and the online processing model satisfied the need of the users of the ILO Library for multilingual subject access, by the latter half of the 1990's, it had become evident that the ILO Library needed a new library system. The version of MINISIS in use in the ILO (Version G) had last been updated in 1986 and was no longer supported. It ran on expensive and outdated hardware; it was complex to manage, expensive to operate and labour-intensive to maintain. MINISIS had not originally been designed for the MARC format (introduced at the ILO Library in 1996), and maintaining the system to use this format was complex and unwieldy. Furthermore the high cost of the hardware

---

and the number of layers of software on top of the basic database system that were required to support MARC records and provide a Web interface made it impractical to offer Labordoc to a wide audience over the Internet.

In 1997, the Library began the process of defining user requirements, evaluating options, and selecting an integrated library system (ILS) to replace MINISIS. Multilingual functionality was just one factor among many governing the choice of a system including cost, computer platform, other functional features, openness and ease of integration with other ILO systems and longtime viability of the vendor. It was clearly not desirable to restrict the choice of a system only to those relatively few ILS systems that supported the online processing model for multilingual subject access. ILO staff felt, however, that it would be possible to provide multilingual subject access using other processing models, and in late 2001, the ILO selected as its integrated library system Voyager from Endeavor Information Systems. Voyager was a relatively new software, although already in use in more than 900 libraries around the world. It used Oracle, the ILO's database standard as its database engine; it ran on common UNIX platforms; and it was developed and supported by a company that was part of the large Reed Elsevier publishing group. An especially important factor was that Voyager had a very open design, offering prospects for closer integration of the ILS with other Oracle information applications being envisaged for the ILO. However while Voyager allowed the OPAC user interface to be translated into other languages, the system, like a number of other integrated library systems, did not provide multilingual subject access using the online processing model described above. Therefore one of the challenges of implementing Voyager at the ILO was the implementation of a different model for providing multilingual subject access, what we call the "pre-processing" model.

### **Pre-processing model (Voyager)**

In the pre-processing model, the online processing required with the previous model is traded for offline processing and extra data storage. Instead of having software translate descriptors at the moment that subjects are being searched, displayed or exported, descriptors are translated in a batch process immediately after they have been entered into the bibliographic record, and the equivalent descriptors in other languages actually entered into the bibliographic record. Once they exist as data in the bibliographic record, normal library system processing can be used during the search, display and export of bibliographic records. We describe how multilingual subject access with the new Voyager integrated system was implemented at the ILO using this pre-processing model.

First, in order to move to the Voyager system, some authority file work was required. Previously in the MINISIS system, subject authority records had been maintained in a special thesaurus authority file where one record, representing a single concept, contained the descriptors in all the different languages. Now subject authority records were loaded in Voyager into the standard MARC format authority file in the different languages of the thesaurus. A single concept was now represented by different subject authority records, each given a different code assigned by the Network Standards Office at the Library of Congress to the different language versions of the *ILO Thesaurus*. (For the English language *ILO Thesaurus*, the code is *ilot*; for the French version entitled *Thesaurus du BIT*, the code is *tbit*; and for the Spanish version *Thesauro de la OIT*, the code is *toit*.)<sup>vii</sup>. While this structure works very well for Voyager purposes, the file structure is not

---

easily processed by external programs, so for reasons of convenience, an additional table was loaded on the server containing one record for each concept, with terms in each of the thesaurus languages.

Once the subject authority file has been prepared, descriptors in English can be entered into the appropriate 6xx subject fields in the MARC record (i.e. the \$a subfield of the 650, 655 or 610 fields) with the appropriate indicators, including the indicator for primary or secondary descriptor. This continues the practice the ILO had established with the MINISIS system of using English as the language for entry and revision of subject descriptions. The *ilot* code for the English language of the *ILO Thesaurus* is entered into the \$2 subfield in each subject field. When the record is saved, the English descriptors in the record are validated against the English language subject authority records, and invalid descriptors flagged for operator correction, as Voyager would do with any unilingual subject system.

Once the subject description is complete, and has been reviewed if necessary by a supervisor, pre-processing for multilingual subject access can take place. An automatic batch process is run against all newly catalogued records. It reads from each record each English language descriptor in turn, looking up the descriptor against the table containing all the thesaurus descriptors in English with their equivalent form in French and Spanish. The French and Spanish descriptors are then added to new occurrences of the 650, 651 and 655 fields with a value in the \$2 subfield indicating the language of the thesaurus (or more properly speaking the version of the *ILO Thesaurus* in that language). Finally the updated records are written back to the Voyager database, replacing the older version of the record.

Figure 2 illustrates this process. If the descriptor "employment" is found in the 650 field of a newly catalogued record, "employment" is used to search the translation table, and retrieve a record where "emploi" and "empleo" are found in the French and Spanish fields respectively. "Emploi" is then entered in a new occurrence of the 650 field in the MARC bibliographic record with a \$2 subfield value of *tbit* and "empleo" is entered into another occurrence of the same field with a \$2 value of *toit*. Similar processing takes place for the other 65X subject fields.

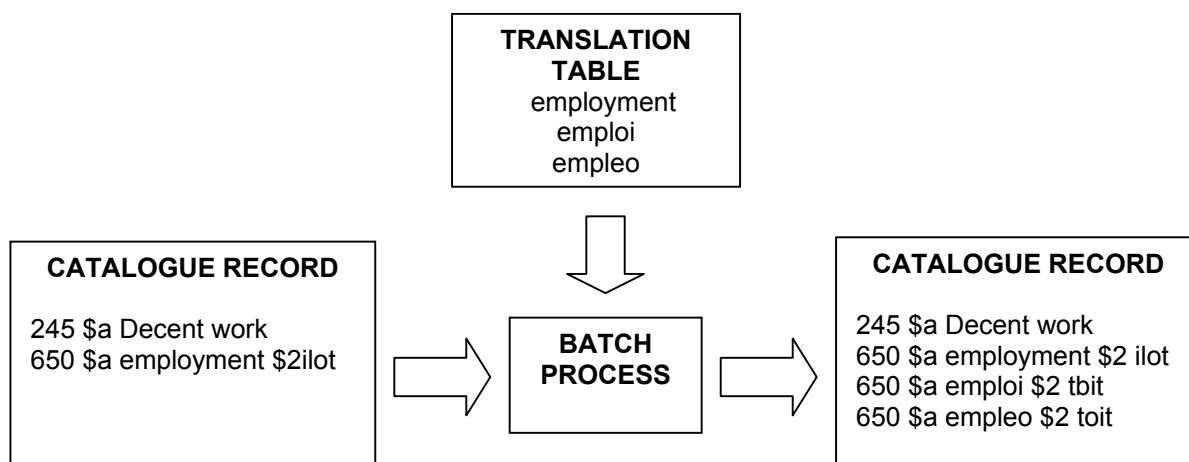


Fig. 2 Batch translation process

---

Once French and Spanish descriptors are added to the 65X subject fields in the MARC record, the French and Spanish descriptors are indexed in the same subject index as the English descriptors. The standard Voyager OPAC then allows users to do a "subject browse" search, whereby they can display and browse up and down lists of headings in alphabetical order, irrespective of the language of the heading. The subject browse search also allows OPAC users to look at the full subject authority record in a given language, complete with hyperlinked Broader Term, Narrower Term and Related Term relationships leading to new lists of terms starting with that particular term: this facility provides users with at least a limited way to navigate the thesaurus structure in order to discover other terms that might improve or expand the original search strategy. Since the thesaurus terms in other languages are actually entered in the bibliographic record, they are also indexed by Voyager's keyword search engine, and are available for keyword searching for a specific field or for general "keyword anywhere" searching with relevance ranking.

This addition of equivalent language descriptors to the topic or genre/form subject fields of the MARC record is not sufficient in itself to support all the requirements for a multilingual OPAC. Assuming an average of 12 descriptors is assigned to a bibliographic record, after pre-processing we now have a MARC bibliographic record with 36 different occurrences of the subject fields (12 for each of the three working languages of the catalogue). Voyager permits selective display of fields on the basis of indicator values, but it does not support filtering of fields based on a value in a particular subfield in the field (in this case, the thesaurus language value in the \$2 subfield). The large number of occurrences of the subject fields in all three languages would have rendered displays for the user of the library OPAC long, illegible, and full of irrelevant or redundant information.

Therefore another solution had to be found for record display. The solution adopted was to copy data into locally defined fields in the record. As descriptors are being translated and added to new occurrences of the 655 or 650 fields, they are also added to fields in the 900 block, reserved for local use. English descriptors for genre/form, primary topical descriptors and secondary topical descriptors are added to 900, 905 and 910 fields respectively; French descriptors to 901, 906 and 911 fields, and Spanish descriptors to 902, 907 and 912 fields. The actual 65X subject fields are never displayed. Instead in the configuration files that control record display in the Voyager OPAC, only these local 9xx fields are indicated for display so as to restrict the display to one language only. Since the display configuration files can be tailored to each individual interface language, the display files for the English language version of the OPAC display only English language descriptors, the display files for the French language version of the OPAC only French descriptors, and similarly for Spanish.

The fact that thesaurus descriptors are now found in locally defined fields in the MARC bibliographic record in all languages of the thesaurus as well as in the 6xx subject fields also simplifies the export of bibliographic information. No further processing is required to share records over the Internet, in union catalogues or to supply them to re-publishers of *Labordoc* on CD-ROM. This feature is particularly valuable for the approximately forty ILO external offices located in countries around the world, most of which have small local library or publications databases. Previously these offices had to wait for records to be specially

---

processed on export in order for them to be able to copy ILO library cataloguing with French or Spanish descriptors into their local catalogue. Now, wherever the offices have an Internet connection, these records are available as soon as they are processed in Geneva.

The automatic processing of subject terms in newly catalogued records required by this pre-processing model has been implemented with a Perl script. The script, which has been run daily since October 2002, begins by exporting records from Voyager based on the date on which the subject description was completed, writing the records to a file on the Sun server on which Voyager runs. The script uses the MARC::Record Perl module<sup>viii</sup> to read each record from the file in turn and to extract from each record the descriptors entered in English. These descriptors are then used to search (via an ODBC connection) an Oracle table consisting of English descriptors and their equivalent descriptors in French and Spanish. The French and Spanish descriptors are added to the MARC subject fields, the descriptors in all languages are added to the MARC local fields, and the updated record is written to a second MARC record file on the server<sup>ix</sup>. This second file is used as input into the Voyager bulk import program, where an import rule matches the bibliographic record identifier in the record in the import file to the bibliographic record identifier in the Voyager database, and automatically replaces the latter with the former. The increased size of the modified records, and the increased disc space and time required to build index files used for retrieval has not proven to be a problem with the robust Voyager/Oracle solution.

## **Conclusion**

Each of the two models described here has advantages and disadvantages. The pre-processing model requires development of a batch script, and involves a significant amount of extra data storage and indexing, though in the current information technology environment where hardware costs are low, neither of these is a particularly significant constraint. More importantly, it does require additional processing when making changes to subject terms, since the batch translation process has to be re-invoked when subject terms change. It also requires a tighter control over workflows to ensure that records that undergo such changes are in fact re-processed. As library collections increase in size to over a million records all these factors may begin to take on more importance.

However the pre-processing model also has advantages over the online processing model. The fact that records contain distinct locally defined fields for different languages make re-use of these records by exchange and database publishing partners easier, particularly if the systems in use at those institutions do not themselves support the online model of multilingual processing. If library systems use add-on keyword retrieval functions which depend on data being found in the record itself, the pre-processing model is clearly more effective. The most important advantage of the pre-processing model is that it frees small to medium-sized multilingual libraries from restricting a choice of an integrated library system to just those ILS vendors that support the online processing model. Instead of being an essential feature, perhaps imposing less than optimal choices in terms of other areas of functionality, multilingual subject support can be evaluated in an even-handed way along with all the other considerations that go into such a complex and difficult choice.

---

## **Bibliography**

Adler, Elhanan. "Multilingual and Multiscript Subject Access: the Case of Israel" *Proceedings of the 66<sup>th</sup> IFLA Council and General Conference, Jerusalem, August 13-18, 2000*. Available at <http://www.ifla.org/IV/ifla66/papers?035-130e.htm>. Verified Feb. 16, 2003

Chachra, Vinod. "Subject Access in an Automated Multithesaurus and Multilingual Environment" in *Automated Systems for Access to Multilingual and Multiscript Library Materials. Proceedings of the Second IFLA Satellite Meeting, Madrid, August 18-19, 1993*. Ed by Sally McCallum and Monica Ertel. Munich: K.G. Saur, 1994 p 63-76.

Clavel-Martin, Genevieve. "The Need for Cooperation in Creating and Maintaining Multilingual Subject Authority Files". *Proceedings of the 65<sup>th</sup> IFLA Council and General Conference, Bangkok, August 20-28, 1999*. Available at <http://www.ifla.org/IV/ifla65/papers/080-155e.htm>. Verified Feb 16, 2003.

Cousins, S. A. and R.J. Hartley. "Towards Multilingual Online Public Access Catalogues" *Libri* 44 (1): 47-62. 1994.

Goosens, Paula. "Across Language Barriers in Multinational OPACs". In *Les bibliothèques, traditions et mutation. Mélanges offerts à Jean-Pierre Clavel à l'occasion de son 65e anniversaire*. Lausanne: Bibliothèque cantonale et universitaire, 1987, p. 401-416.

Landry, Patrice. "The MACS Project: Multilingual Access to Subjects (LCSH, RAMEAU, SWD)" *Proceedings of the 66<sup>th</sup> IFLA Council and General Conference, Jerusalem, August 13-18, 2000*. Available at <http://www.ifla.org/IV/ifla66/papers/165-181e.pdf>. Verified Feb. 16, 2003.

Lebowitz, Abraham I, Robert Portegies Zwart and Helga Schmid. "Multilingual Indexing and Retrieval in Bibliographic Systems: the AGRIS experience". *IAALD Quarterly Bulletin* 36, 3 (1991) p. 187-191.

MARBI. *Recording Language of Heading in USMARC Authority Records* [Discussion Paper No. 108]. Washington: Library of Congress, 1998. Available at <http://lcweb.loc.gov/marc/marbi/dp/dp108.html>. Verified Feb. 16, 2003.

MARBI Multilingual Record Task Force. *Multilingual Authority Records in the MARC 21 Authority Format* [Discussion Paper 2001-DP05]. June 8. 2001. Available at <http://www.loc.gov/marc/marbi/2001/2001-dp05.html>. Verified Feb. 16, 2003

Michos, S., E. Stamataos and N. Fakotakis. "Supporting Multilinguality in Library Automation Systems Using AI Tools." *Applied Artificial Intelligence* 13(7) : 679-704. Available from <http://slt.wcl.ee.upatras.gr/papers/michos3.pdf>. Verified Feb. 16, 2003.

Peters, Carol and Eugenio Picchi. "Across Languages, Across Cultures: Issues in Multilinguality and Digital Libraries" *D-Lib*, (May 1997) Available at <http://www.dlib.org/dlib/may07/peters/05peters.htm>

---

Pollitt, Stephen et. al "A common query interface for multilingual document retrieval from databases of the European Community Institutions" *Online Information 93: Proceedings of the 17<sup>th</sup> International Online Information Meeting, London, 7-9 December 1993*, p. 47-61.

Powell, James and Edward A. Fox "Multilingual Federated Searching Across Heterogeneous Collections" *D-Lib Magazine*, Sept. 1998. Available at <http://www.dlib.org/dlib/september98/powell/09powell.html>.

Soergel, Dagobert. "Multilingual Thesauri in Cross-Language Text and Speech Retrieval. In *Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, Stanford, CA. 1996. p. 164-170. Available at: [www.ee.umd.edu/medlab/filter/sss/papers?soergel.ps](http://www.ee.umd.edu/medlab/filter/sss/papers?soergel.ps). Verified Feb. 16, 2003.

Slater, Ron "Authority Control in a Bilingual OPAC: MultiLIS at Laurentian" *Library Resources and Technical Services* 35 (4): 422-458.

---

<sup>i</sup> English, French, Spanish, Russian, Chinese, Arabic and German.

<sup>ii</sup> International Labour Organization. *ILO Thesaurus: Labour, Employment and Training Terminology* 5<sup>th</sup> edition. Geneva: ILO, 1998.

<sup>iii</sup> To mention a few examples, the works cited in the bibliography by Chachra, Landry, Lebowitz, Michos, Powell, Pollitt and Slater all describe systems based on this model, used either in library catalogues or bibliographic databases.

<sup>iv</sup> There are in practice two variants of the online processing model: either all documents in the database are described in a single language (e.g. all documents have descriptors in English), or documents are indexed in the language most appropriate to the language of the document and/or to the indexer (e.g. French documents are indexed with French descriptors, Spanish with Spanish and all others in English). Because the concepts and the terms in the various languages are defined in the multilingual thesaurus, using a single language or any combination of languages for descriptors in the bibliographic record is transparent to the end user and indicates in itself no language bias. Cf. Soergel's assertion (1996) that "a cross-language retrieval system with controlled vocabulary must support the indexing of documents or other objects-- the assignment of controlled vocabulary descriptors-- in the various languages."; this is true only where there is no multilingual display capability.

<sup>v</sup> With a trilingual thesaurus, and only a single language used for entering descriptors, the search for two of the subject terms will of course always result in 0 hits, since these terms do not in fact appear in any bibliographic record. However this is irrelevant, since the Boolean OR in the expanded query will always result in the complete set of documents representing the concept, regardless of the language of the search term entered.

<sup>vi</sup> such as the Helecon CD-ROM

<sup>vii</sup> This approach is consistent with the recommendations of the MARC Multilingual Record Task Force, where cataloguing context is indicated with special coding in a proposed \$7 subfield. The coding in the ILO Voyager system uses the \$2 subfield, but the coding of the language versions of the thesaurus is logically equivalent to the an encoding of the *ILO Thesaurus* as the subject source followed by an encoding for language, e.g. \$2 tbit is logically equivalent to the proposed \$7 ilot//fre as per the MARBI recommendations.

<sup>viii</sup> Available at <http://www.cpan.org>.

<sup>ix</sup> In fact, some other additional processing is performed at the same time, e.g. a status code and date modified are added to the record to help track workflow, and descriptors are validated a second time, ensuring that, even in the case of operator error, no invalid descriptors get through the database.

---