European Co-operation in standardisation of lexicographical resources and merging of existing specialised dictionaries for Internet purposes

Marie-Jeanne DEROUIN and André LE MEUR,
München, Germany and Rennes, France

Abstract

The traditional task of the specialised dictionary is to provide users with bilingual or multilingual tools for translation purposes. This has been undergoing important changes in recent years in the dictionary-making process.

In the meantime, the increase in communication and the legal obligations in the enlarged European community are creating new specialised dictionary needs in language combinations. At the same time, globalisation of these language-markets requires quick and easy access to multilingual information. Making such information available to a world-wide public through Internet or Intranet solutions will be a considerable asset for the European publishing industry in the next few years.

These were reasons enough for the motivation which led to taking part in the *EU-MLIS Publishnet* project the scope of which has been to bring together the complementary know-how of potential partners at European level and experiment with new techniques which will be current and necessary in the future.

1 The new specialist dictionary publishing needs

- Related to the electronic versions: need for numerisation of the data

  In the past five years most specialist dictionary publishers have launched a CD-ROM programme in addition to the range of print dictionaries. Publishers have first to make a choice in the matter of data management. In most cases SGML is given preference.

- Related to the lexicographical sources: need for a generic data-exchange format

  Less and less experts have now time for compiling lexicographical entries. For new publication and updating of existing dictionaries, publishers of specialist multilingual dictionaries will have to outsource existing terminology collections. There is the need for a generic data-exchange format.

- Need for merging existing dictionaries

  In all European countries English has become the first language for communication in business, science and technology . Nevertheless bilateral communication between non-English-speaking communities exists and the use of English as a basis for communication is very often unreliable and inconvenient. Publishers need to merge existing dictionaries and create new language pairs from them. This solution might well enable to reduce the efforts necessary to publish a new edition for a small market and provide users with the specialised dictionaries they need.

- Related to the globalisation of the market: need for online publishing solutions.
  With the development of information technology, individual consumers and professional users now require quick and easy access to multilingual information from where they are. Publishers have to offer them new purchase opportunities through internet or intranet solutions and thereby develop their online-publishing and marketing strategies.

2 Presentation of the Publishnet MLIS project

The objective of the *MLIS Publishnet project* focussed on setting up a network of publishers in the Internet with a view to harmonising and standardising terminological data belonging to technical domains covered by conventional dictionaries. The purpose of this project has been to create an experimental web-site and later include all publishers interested in this network.

The Publishnet project included three European specialist dictionary publishers: *La Maison du Dictionnaire* (France), *Diaz de Santos* (Spain) and *Langenscheidt Fachverlag* (Germany); a software editor: *LCI* (France); experts in terminology and language technology: *Université de Rennes 2* (France).

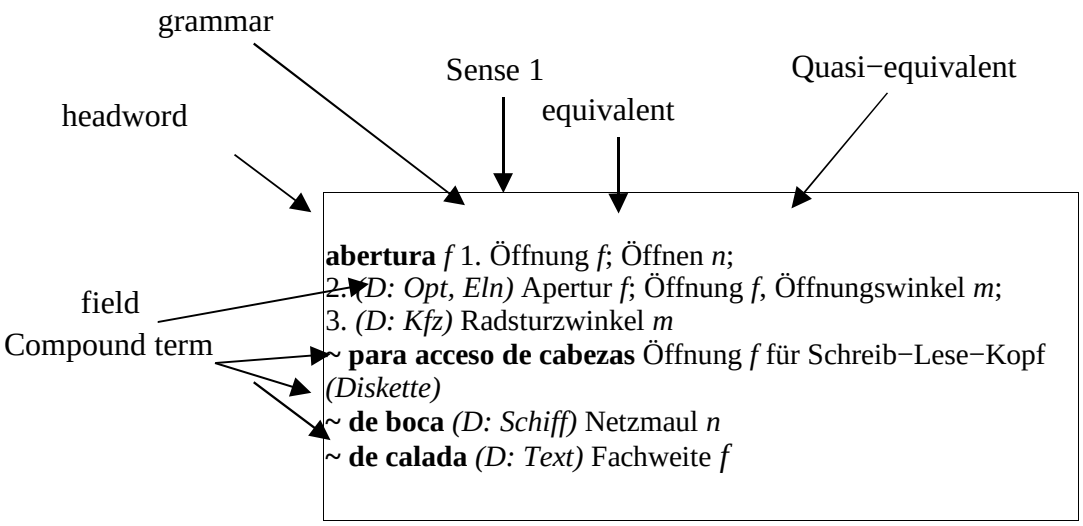
Seven technical dictionaries of different size and format have been chosen.

The project was divided into seven stages: Identification of terminological resources; Study of the file formats; Conversion of data to a standard format; Data access modules; Merging existing dictionaries; Copyright; Dissemination of data. We have chosen to focus our paper on points 3 and 5, data conversion an merging of dictionaries.

3 Conversion of data to a standard format

It is obvious that each editor has its own lexicographical format but it is also obvious that all of them share a very general model for their lexicographical entries, namely the underlying structure of *ISO 1951* standard. This has meant that it was possible in the Publishnet project, after a study of the source formats and a review of existing formal models such as TEI (Text Encoding Initiative), to draw a generic model - a hierachical tree – so that each existing element in the dictionaries could be assigned to one sole node in this tree.

The following example (*Langenscheidt Fachverlag: LFG*) shows how typographical signs (bold, italics, comma, semicolon) are used, in accordance with *ISO 1951*, in order to mark the structure of the entry :



In line with this observation, an extension to the SGML *GENETER* DTD (Document Type Definition) was elaborated to represent this structure. It is described in Annexe F of the ongoing Geneter Work Item (*ISO TC 37/SC3/WG4*).

An entry (<lexical-entry>) deals with one and only one lexeme belonging to a language source.

If the lexeme belongs to several grammatical categories, each "homograph" is described separately.

For each target language, the lexeme is described in the element <language>. In a language, a description may be given in one or more senses (<sense>).

In a <sense> block, application fields (<field>) may be used to specify the domain or subject area the lexeme expresses or belongs to. At this level it is possible to give a definition (<definition>), the contexts and examples.

The <target> element indicates the equivalent of the <headword> in the target language.

It is accompanied by collocations in both source and target languages (<equivalent-collocation>).

An entry may include a cross-reference (<see>) to other entries. Derivatives (<lexical-derivation>) and compounds of or involving the lexeme (<compound-term>) can also be described.

It is worth noting that:

- <headword> has two representations i.e., <pr> (Printable Representation) and <rr> (Real Representation). A distinction can therefore be made between the displayable form of a headword (including optional and/or alternative elements) and its indexable form(s). A headword may have one <pr> and many <rr>.

- administrative elements (address pointers, creation dates etc.) and descriptive elements (phonetics, grammar, geographical restrictions or usage, etc.) are identical to the data categories of a Terminological entry. As far as possible they come from IS0 12620.

- For each data category the DTD indicates the content model, which is composed of *GENETER* content elements (15 elements specific to Geneter for expressing dates, responsibility, pointer etc., and 64 elements from XHTML).

4 Research and Experimentation on merging the data of existing dictionaries

It was decided to limit this experiment to one domain: *Computer Sciences*.

A software tool was developed to align each dictionary according to its corpus. This procedure first created two databases, each with a pivot language, either English or Spanish: Database1: En-Sp-De; Database 2: Fr-En-Sp.

- First step: Automatic fusion:
  About 4% of the multilingual databases were capable of automatic fusion. The main reasons for this low score are as follows:

  1. The structure of the content of specialist dictionaries according to domains and subdomains can be vary greatly from publisher to publisher;
  2. The problem of extraction of the data according to the domain. Most bilingual specialised dictionaries have not been compiled in a database.

- Second step: Completing the database manually:
  This task was carried out by the authors of the different specialised dictionaries. The adding of missing equivalencies ranged from 30% to 80% of the missing entries.

- Third step: Validation and quality assurance of the missing entries:
  The authors chose the mostly used in up-to- date equivalency. Nevertheless the risk of error is still high as the existing entries are out of context and a very careful and expensive process of validation is necessary.

- Fourth step: Validation of the multilingual entry:
  After having merged and partly completed the two computer scienaces data-bases, every new language pair had to be validated by authors as for a new edition. In this respect the merging of specialist dictionary data as we experienced it in this project can only be considered as a step towards a "semi-raw" manuscript which has to be thoroughly edited afterwards.

5 Conclusion: Limits and perspectives of this experience

Without doubt the most promising results were provided by the conversion of the data of the existing dictionaries. This operation emphasised the necessity for a consistent data structure within the dictionaries. Secondly the successful conversion of the data proved that the *SGML*

format compliant *GENETER meta-model* originally dedicated to terminological resources could be also be used for lexicographical resources.

Concerning the merging of the specialised dictionaries, the limited experiment with sample points out important needs which will have to be taken into account by publishers in the future such as:

- a profound requirement for standardisation in the field of lexicography for harmonising lexicographical resources. The existing *ISO Standard 1951*: *Lexicographical symbols and typographical conventions for use in terminology* and its national versions will have to be thoroughly updated to meet the needs of information technology. In Germany a working group has started updating the equivalent *DIN 2336* and a new *ISO 1951* working group is due to start in 2000.
- The necessity to develop more intelligent tools for merging specialised dictionaries on a conceptual level in order to reduce and optimise the validation procedure and assure the quality of the new generated dictionaries.