Report on the Revision of the Lexicographical Standard ISO 1951
Presentation/Representation of Entries in Dictionaries

Marie-Jeanne Derouin, Dr. André Le Meur

Marie-Jeanne Derouin
Managing Director
Langenscheidt Fachverlag GmbH
Postfach 40 11 20
D-80711 München
Germany
marie-jeanne.derouin@langenscheidt.de

Dr. André Le Meur
Maître de conférence en Informatique
Laboratoire RESO – CNRS – UMR 6590
Université Rennes 2
6, Avenue Gaston Berger
F-35043 Rennes, France
andre.lemeur@uhb.fr

Abstract

The two authors of this paper belong to the expert commission of the standardization bodies in France (AFNOR) and in Germany (DIN) and are, within the ISO/TC37/SC2, project leader and expert for the revision of the ISO-standard.
In this paper we will report on the revision of the standard ISO 1951 *Presentation/Representation of entries in dictionaries* which will give recommendation regarding the organization of lexicographical entries and take in account the computer-based dictionary manuscript and its various uses and reuses on different print and electronic devices.

1. Introduction

The first steps towards the revision of this standard started in 2000 when the German ISO-delegation reported in London on the ongoing updating of the equivalent national DIN-standard 2336 *Lexikographische Zeichen für manuell erstellte Fachwörterbücher* according to the needs we have presented in the first part of this paper. Consequently it has been decided to check whether the needs expressed in Germany can apply to most countries in the world or not and a feasibility study has been carried out in every ISO-member country.

2. Feasibilty Study

The following questionnaire has been sent to lexicographers in universities or special schools, specialized dictionary authors, specialized dictionary publishers, terminology department of industrial companies and national or international bodies:

1. To what extent does the above mentioned paper meet or not meet the needs in your country ?
2. To what extent does the existing description of the lexicographical symbols and conventions meet the new needs of data-management and electronic dictionaries for the different language combinations inclusive the ideogram languages such as Chinese, Japanese etc..?
3. Is there a demand in your country for a standard regarding the representation of entries in specialized dictionaries and database?
4. If some of you never used the ISO 1951 standard, to which extent will you be able to work in the future without taking a standard into account and the reasons for it ?
5. Which experts in your country would be ready to take an active part in a workshop concerning this ISO-work item?

The results of the feasibility study show that most countries insist on the fact that ISO 1951 does not anymore meet the current needs in lexicography. In Sweden, for example, the ISO-standard has not been adopted as a national standard and it has not had any impact on current lexicographical and terminographical practice. The Nordic Association for Lexicography applies its own model for the presentation of entries in specialized dictionaries, terminological vocabularies and databases. In most countries there is an urgent need for a standard for representing and exchanging data of special languages which should take in account the needs of the computer-based lexicography in order to get a consistent representation of the entries and therefore homogeneous dictionaries.

The French AFNOR wishes that the redefined standard should define a solid XML-based format (see below an example) for representing and exchanging data so that each collaborating partner would need one single and export routine. According to AFNOR, the scope of the standard should be larger that the only “specialized dictionaries” and it would be worth enlarging it to general- monolingual and multilingual- dictionaries.

In Germany, the revision of the DIN 2336 with its new title and scope, *Darstellung von Einträgen in Fachwörterbücher und Terminologie Datenbanken* is nearly finished. The German DIN is ready to propose the German revised standard as basis for the development of the revised ISO 1951.

3. First Steps towards the Revision of the Lexicographical ISO-Standard 1951

During the Toronto ISO-meeting 2001, according to the positive feasibility study, a resolution to revise the ISO-standard 1951 has been approved. The revised standard will apply to general and specialized dictionaries and give a specific model for lexicography. Its objective is to facilitate the management, use, reuse and exchange of data for dictionaries. Its new title is: *Presentation/Representation of entries in dictionaries*.

Experts from nine countries Austria, Belgium, Canada, Finland, France, Germany, Greece, United Kingdom and Ukraine have started to develop this revision on the basis of the new revised German standard in November 2001. The following proposals will be taken in account in a working draft due to be circulated before the next ISO-Meeting which will take place in Vienna in August 2002.

a) Although the forthcoming new German DIN 2336 provides a variety of possible layouts for presenting data in different electronic environments it appears to be too much focused on the print specialised of dictionaries since only one subclause is devoted to the presentation of databank entries. It is restricted to the presentation issues concerning typographical characters and conventions and types of entry arrangement, without working on the systematical structuring of the presented lexical information, such as sequence of information. Consequently it can serve as basis for starting the revision but will have to be significantly extended and restructured.

b) The future new ISO-standard 1951 will have to cover the different options of organisation and management of data for print and electronic environments. In other terms, a formal model independent from presentation of data is needed. This model should be build in order to obtain any layout (and particularly the DIN 1336) through stylesheets, and to fulfil the requirements for electronic editing, storing, querying and dissemination.

c) It will have to cover a wide range of lexicographical resources such as general and specialised dictionaries, monolingual and multilingual, Machine Readable Dictionaries (MRDs) etc.

d) Uniformity at the exchange of data should be ensured. Except for the specifications for typographical conventions, already described in the present ISO-standard 1951, we need a more generic data exchange format.

e) Moreover, a DTD should be initiated so as the creators and the users of the lexical collections to be confident that can (re)produce and use unambiguously parts or the whole of the included information. That DTD should also cater for optionality issues of the data, combination of data categories, which may influence the presentation options providing a structured generic exchange format.

For that, the experts will have to take into account the published specifications for dictionaries and lexicons like TEI1 (Text Encoding Initiative), EAGLES2, ISLE3 and other works related to this matter such as Pierre Corbin's EURALEX 2002 paper on "Composants lexicographiques et contenus informationels des dictionnaires".

Moreover lexicographical description models have to be compatible with other models for linguistic resources description like lexicons4 and terminologies5.

1 TEI P4 - http://www.tei-c.org/P4X/DTD/teidict2.dtd

2 Preliminary Study of the Structure of Lexicon Entries http://www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html

3 Survey of Major Approaches Towards Bilingual/Multilingual Lexicons : http://lingue.ilc.pi.cnr.it/EAGLES96/isle/clwg doc.html

4 Open Lexicon Interchange Format : http://www.olif.net

4. First steps : towards a new formal representation of entries in dictionaries

In a previous paper [DEROUIN, LE MEUR 2000] a first inventory of data categories for printed or machine-readable dictionaries has been presented, based on the observation of seven technical dictionaries. This inventory[6]. considers now thirty technical, general, bilingual or monolingual dictionaries. It shows that :

- more than sixty elements are required in order to represent all the informations we can find in dictionaries,
- many elements (administrative information for instance) are commun to all linguistic resources,
- applying the principle of **subsidiarity**, an accurate description of many elements can be borrowed to existing more specialized formats : for instance **ontological relations** can be borrowed to concept oriented terminological formats and **morphological**, **syntactical** and **semantical** descriptions can be borrowed to machine oriented lexicons.

A first draft of such a formal dictionary model with a XML Document Type Definition is under developpment. It takes into account most of the structural features that are described in the previouly mentioned analysis (TEI, ISLE, etc.).

The example below shows how a classical entry of a technical german-english technical dictionary maps on this structure.

Läufer *m* **1**. (*El*) rotor *m*, induit *m* (*bei Gleichstrommaschinen*); **2** (Strm) rotor *m*, roue *f* mobile (*s.a.* Laufrad 1.); **3**. curseur *m* (*z. B. einer Spinnmaschine*); **4**. couette *f* (coiffe *f*) vive, coulisse *f* vive (*Stapellauf*); **5** garant *m* (*Tau*); **6**. panneresse *f* (*Mauerwerk*); **7**. coulure *f* (*Ansttrichfehler*); **8**. *s.* Cursor

FIGURE 1 : Sample

The following illustrations (figure 2 and XML encoding) show how to represent such an entry.

Note the fact that for this example, the lexicographical description has been enriched with two main features:

- a morphological description (for german inflections) of the headword,
- a concept relation with an hyperonym.

As far as possible Generic Identifiers (tag names) are self-explanatory.

Two XML namespaces are used :

- LEX (for lexicography),
- OLIF (for the Olif format).

The implicit namespace is GEN for Geneter, which plays here the role of a framework in which these three points of view on linguistic resources.can collaborate.

5 Terminology Markup Framework http://www.loria.fr/projets/TMF

6 LEX : Elements for a formal representation of lexicographical data categories -AFNOR - X03 A - G1 N7: http://www.genetrix.org/lexicography/texts/Lex-en.doc

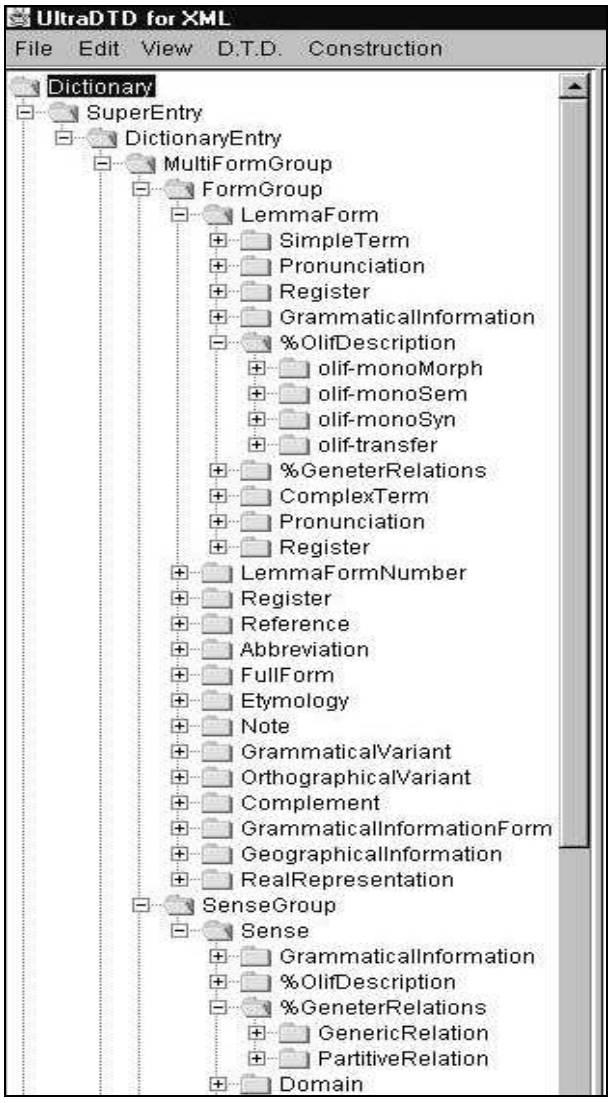

Figure 2 :Tree structure of an entry

This figure illustrates the general outline of a lexicographical entry for Machine Readable Dictionnaries which keeps the traditional features of printed dictionnaries such as printed layout (see [DEROUIN, LE MEUR 2000]) but is enriched with morphological, syntactical and semantical features (%OlifDescription in figure 2, prefix olif: in the encoding) coming from Translation Oriented Lexicons (Olif) as well as with ontological relations (%GeneterRelations in figure 2, <GenericRelation> in the encoding) coming from Geneter, a concept oriented markup language specified in ISO 16642[7].

More information about this technique of **hybriding semasiological** and **onomasiological** descriptions[8] of linguistic resources based on shared XML namespaces is available at http://www.genetrix.org/texts/subsidiarity.doc

The full encoding of this example and tools for validation and presentation (XSL stylesheet) are available at http://www.genetrix.org/lexicography/

[7] http://www.loria.fr/projets/TMF/ - Annex C - Geneter

[8] http://coral.lili.uni-bielefeld.de/EAGLES/WP5/termdeliv97/node13.html

XML encoding

```
<?xml version ="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE Geneter SYSTEM
' http://www.genetrix.org/dtd/GeneterV06.dtd' >
<Geneter >
<LEX:Dictionary>
<LEX:DictionaryEntry id=' boch2'
sourceLanguage=' de' >
<LEX:FormGroup>
<LEX:LemmaForm>
<LEX:SimpleTerm>Läufer</LEX:SimpleTerm>
<olif:monoMorph>
<olif:inflection>
<olif:paradigm>
<olif:inflectedForm>
<olif:form>Läufers</olif:form>
<olif:monoMorph>
<olif:case>g</olif:case>
<olif:number>sg</olif:number>
</olif:monoMorph>
</olif:inflectedForm>
<olif:inflectedForm>
<olif:form>Läufern</olif:form>
<olif:monoMorph>
<olif:case>g</olif:case>
<olif:case>d</olif:case>
<olif:number>pl</olif:number>
</olif:monoMorph>
</olif:inflectedForm>
</olif:paradigm>
</olif:inflection>
</olif:monoMorph>
</LEX:LemmaForm>
</LEX:FormGroup>
<LEX:SenseGroup>
<LEX:Senseid=' boch3' >
<GenericRelationvalue=' superordinateConcept' >motor
</GenericRelation>
<LEX:TranslationGroup>
<LEX:TranslationEntity>
<LEX:Translation>
<LEX:SimpleTerm>rotor</LEX:SimpleTerm>
</LEX:Translation>
</LEX:TranslationEntity>
</LEX:TranslationGroup>
</LEX:Sense>
</LEX:SenseGroup>
</LEX:DictionaryEntry></LEX:Dictionary></Geneter>
```

Xml encoding of the sample

References

[DEROUIN; LE MEUR 2000] Derouin, M.-J., Le Meur A. 2000. European Co-operation in standardisation of lexicographical resources and merging of existing specialised dictionaries for Internet Purposes, *Proceedings of the Ninth Euralex International Congress, Euralex 2000*