ISO 1951: a revised standard for lexicography

André Le Meur and Marie-Jeanne Derouin

Why a revision of this standard?
Times are changing, also in matters of dictionary making. Lexicographical methods are well established, both on the publisher and the user’s side. For centuries, paper was the only media for which publishers had developed an impressive know-how. But as everything is now going digital, new methods for data management are being found. Although everyone is aware of the growing importance of electronic devices that are full of promises, few are ready – in spite of the numerous prophecies of recent years – to get rid of the well-established traditional methods of handling large printed data collections.

With the introduction of digital supports and networking, dictionary lifecycle has been considerably extended. The original manuscript has now become a unique source that can be accessed many times in order to be reused and even integrated into other language applications. For data manipulations such as merging dictionaries, inverting language directions, extracting and merging nomenclatures, integrating lexicographic data in terminological tools or lexical databases, etc, dictionary publishers and individual compilers are increasingly aware of the necessity to structure their contents according to standards recognized by other professionals, in order to avoid time-consuming and expensive data manipulations.

In the past decade, different proposals have either used existing printed dictionaries as a basis, including their “fuzzy” aspects and inconsistencies (TEI, for instance), or have deliberately chosen compatibility with strictly structured computer-based lexical databases that don’t allow for well-established habits of lexicographers. Therefore, there was a need for a method that takes into account both of these aspects: tradition and strictness, lexicography and computational linguistics. Thus, within ISO TC37 SC2, publishers, researchers, lexicographical and terminological experts have merged their experience to propose a revision of the ISO 1951 standard “Presentation/representation of entries in dictionaries”, which aims to bridge the above-mentioned traditional methods of dictionary-making with future oriented ones. This revision is due for publication in 2007.

XmLex: a generic model for dictionaries

XmLex (previously called LEXml)
is the formal model proposed in this new standard, and applies to any type of dictionary. It aims at finding a balance between strict formal structures (which allow automation) and user friendliness for the human editor, while preserving conformity to traditional lexicographic methods. It satisfies four requirements, which enable data that conform to this model to be independent from both the tools (free or commercial) and the media (paper, internet, cdrom):

- Complete separation between logical structure and display: all the punctuation and other structure markers can be automatically generated at the display stage, which means that data are independent from the media used for display.
- Non ambiguity: all the relations between elements can be computed so that XmLex data can be interfaced with any lexical database (e.g. the ISO Lexical Markup Framework project) or other linguistic applications relying on a clearly specified model.
- Flexibility: the XmLex model is generic. By applying XML rules of subsetting, as defined in ISO 16642 annex C, it is possible to specify subsets corresponding to specific needs. A subset accepts any order of elements, so that the editing structure can be strictly parallel to the display order (e.g. XSL stylesheets for transforming dictionary entries can be written in pure “push style” OR in pure “pull style”).
- Compatibility with currently available XML tools: it is now widely accepted that linguistic applications should not use proprietary formats and tools. XML and its associated specifications have become industrial standards. XmLex can be implemented as an XML schema and operated by commonly available XML editors and by XSL stylesheets.

XmLex uses data elements defined in ISO 12620, if they already exist. Moreover, it defines data elements specific to lexicography that have been observed in existing dictionaries. These new data elements will be proposed for inclusion in the forthcoming ISO TC 37 Data Category Registry.

First applications: an XML model, a subset for bilingual dictionaries
XmLex is an abstract model that can be applied to any type of dictionary (monolingual, bilingual, general, specialized, etc). For informative purposes only, an XML implementation has been specified, including a subset that represents currently available bilingual dictionaries. The XmLexIntro document describes the ‘XmLexWorkbench’, which contains:

- The generic DTD (XmLex_V00.dtd), corresponding to the generic XmLex model.
- The subset corresponding to bilingual dictionaries (XmLexForBilingualDictionaries_V00.dtd).
- Examples of bilingual entries.
- The following tools for transforming XML entries:
  - XmLexDisplayer.XSL transforms entries into HTML with a print-like preview. This XSL stylesheet and its CSS must be adapted for specific needs. Their major role is to show that, although the XmLex model deals only with content, presentational issues (such as numeration and punctuation) can be solved automatically.
  - XmLexInverter.XSL shows how to “invert” lexicographical entries (i.e. to find for any linguistic unit in the target language all related information in the source language). It illustrates the fact that since XmLex structures are non-ambiguous, methods like backtracking can be used for exploring any path in any direction when data have to be reused in a different context.
  - NomenclatureLib is a set of XSL stylesheets that extracts and lists the nomenclature (the list of the linguistic units in the source language of a dictionary) in bilingual dictionaries.
  - LexTermLib is a set of XSL stylesheets used to transform XmLex entries into terminological entries compatible with ISO TC37 terminological model and with concept-oriented tools like Translation Memory systems.

Note that this library is given “as is”. Its aim is only to illustrate the use of XmLex, and to initiate a public “open source” collection of useful and reusable algorithms for lexicographic data management that may help newcomers to evaluate the potential of XmLex.

Perspectives
The revised ISO 1951 document, with its specific model based on current professional practices, is intended to allow all possible lexicographic production, exchange and management procedures. Some publishers have already modified their editorial work-flow accordingly. The first integration of dictionaries using this model for providing Translation Memory tools, in parallel with traditional dictionary production, will be put on the market this year.
Particular acknowledgment is due to the other members of the editorial board of the revised ISO 1951: Elisabeth Blanchon, Oliver Schweiberer and Christine Tauchmann, as well as to Mariusz Idzikowski, Elena Mantzari, Claude Nimmo and Yuka Sasaki who assisted us with very useful comments. Last but not least, we thank Ilan Kernerman for his contribution in coining the term XmLex.

Notes
3 ISO CD 24613:2006 Lexical Markup Framework (Committee Draft)
5 ISO 12620:1999 Computer applications in terminology – Data categories
6 http://www.xmlex.net/lexicography/xmlexintro.pdf

An extract of the XmLex structure for bilingual dictionaries.

The full DTD and explanations about the data elements are available online as indicated above.

<ELEMENT Dictionary>
  (DictionaryEntry*)>
<ELEMENT DictionaryEntry (OriginationCtn?, CreationGrp?, ModificationGrp*, StatusCtn?, WorkingNoteCtn?, (Headword|HeadwordCtn|HeadwordBlock), (HomographGrp|SenseGrp|SenseBlock)*, (MultiWordUnit|MultiWordUnitCtn)*, (CompositionalPhrase|CompositionalPhraseCtn)*)>
<ELEMENT HomographGrp (Headword|HeadwordCtn?, (PartOfSpeech|PartOfSpeechCtn)*, Subcategorisation*, GrammaticalNote?, (SenseGrp|SenseBlock)*, MultiWordUnitCtn*, CompositionalPhraseCtn*)>
<ELEMENT SenseBlock (SubjectField*, Register?, SenseGrp*)>
<ELEMENT SenseGrp (Subcategorisation*, SubjectField?, SenseQualifier*, SenseIndicator*, GeographicalUsage*, RangeOfApplication*, TypicalComplement*, AbbreviatedForm*, FullForm*, GrammaticalPattern*, Register*, (Definition|DefinitionCtn?), Explanation?, (Translation|TranslationCtn|TranslationBlock)*, (Example|ExampleCtn)*, (Synonym|SynonymChn|SynonymBlock)*, (Antonym|AntonymCtn|AntonymBlock)*, SeeAlso*, CitationCtn*, Note?, SenseGrp?, MultiWordUnitCtn*, CompositionalPhraseCtn*, (SeeCtn|See)?)>