

Presentation of the new ISO-Standard for the representation of entries in dictionaries: ISO 1951

Marie-Jeanne Derouin, Dr. André Le Meur

Managing Director
Langenscheidt KG/ Langenscheidt Fachverlag
Postfach 40 11 20
D-80807 München
Germany
marie-Jeanne.derouin@langenscheidt.de

Maître de Conférences en Informatique
Université du Québec en Outaouais
283 boulevard André Taché
Case Postale 1250, succursale Hull
Gâtineau (Québec) Canada J8X 3X7
andre.lemeur@uqo.ca

Abstract

Times have changed over the last ten years in terms of dictionary production. With the introduction of digital support and networking, the lifespan of dictionaries has been considerably extended. The dictionary manuscript has become a unique data-source that can be reused and manipulated many times by numerous in-house and external experts. The traditional relationship between author, publisher and user has now been extended to include other partners: data-providers – either other publishers or institutions or industry-partners -, software developers, language-tool providers etc. All these dictionary experts need a basic common language to optimize their work flow and to be able to co-operate in developing new products while avoiding time-consuming and expensive data manipulations. In this paper we will first of all present the ISO standardization for Lexicography which takes these new market needs into account, and then go on to describe the new standard ISO 1951: “Presentation /Representation of entries in dictionaries” which was published in March 2007. In conclusion, we will outline the benefits of standardization for the dictionary publishing industry.

1. General points on ISO-Standardization

Standardization is achieved on national, regional and international levels. Most countries in the world have their own national standardization body which is organized into technical committees according to the application fields and develop their own national standards.

Most countries also contribute to the development of international standards which are accepted worldwide. Stakeholders from industry, academia and societies from all over the world are involved in creating international standards within the ISO (International Standardization Organisation).

International standards aim to prevent national variations and give priority to describing performance requirements and interfaces. Products are not standardized.

Standardization is achieved according to the following principles:

a) Market adequacy: A project will only be launched if a need has been identified through a market survey in all member countries.

b) Voluntary implementation: Experts work on standards on a voluntary and anonymous basis

c) Consensus based: At each of the main stages of development, experts from all member countries have to give their comments and to vote for or against the circulated draft.

d) One vote per country.

e) Worldwide acceptance

2. The current situation in dictionary standardization

International standardization in Terminography and Lexicography is the responsibility of the TC 37/SC2 (Technical Committee 37/Subcommittee 2). Only one ISO-standard for dictionaries has been published: the ISO 1951: “Lexicographical symbols and typographical conventions for use in terminography”. This appeared in 1973 and was revised in 1997. It mainly concerned the layout of print-dictionaries and did not address the actual needs of dictionary making. Now standards have to take into account the broadest variety of dictionary-products: monolingual, bilingual, multilingual, general language and specialist dictionaries, in print form or electronic, as isolated products or integrated into other language tools.

With the introduction of digital support and networking, the lifespan of dictionaries has been considerably extended and the manuscript has become a unique data-source that can be reused and manipulated many times by editors and content management experts. Dictionary publishers are importing and accommodating more and more data from other sources, such as industry and other institutions or from other publishers when acquiring perhaps a new dictionary series or co-publishing a new dictionary. In many cases they have to deal with international partners. Dictionary publishers are also content-providers for other language tools such as translation software or translation memory systems or other internet products.

In order to optimize their work flow and to avoid time-consuming and expensive data manipulations, all these different players need to be implementing the

same current best practices in terms of data representation. Taking these new challenges into account, the ISO TC 37/SC2 has just published the completely revised ISO-Standard 1951.

3. The revised ISO Standard 1951: Presentation/Representation of entries in dictionaries

3.1 This standard aims

- a) to support the creation and management of various types of dictionary
- b) to allow dictionary content to be reused in different printed and electronic formats
- c) to facilitate necessary production, exchange and management procedures
- d) to propose a specific model based on current best professional practices

It has been approved by experts from over 20 countries and it applies to every kind of dictionary.

3.2 The normative section indicates requirements to be met and includes:

- a) Definitions concerning basic and unambiguous terms relating to dictionary structure and presentation, common in most types of dictionary, which are within the scope of this International Standard.
- b) A formal description of dictionary entries with an overview of data elements and compositional elements. It describes data elements using numerous examples from over 40 existing dictionaries and their grouping into compositional elements needed for representing most common dictionary entries. For each data element a conventional name and description is provided, based as far as possible on [ISO 12620:1999]. Free data elements, definable by the user, allow extensions of the model for “negotiated interchange”.
- c) A formal generic dictionary model called XmlLex which provides a balance between strict formal structures, which permits automation, and user friendliness whilst still conforming to traditional lexicographic methods. This model also satisfies four requirements which enable data to be independent from tools and media (paper, CD-ROM, internet). They are:
 - 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 media used for display
 - Non ambiguity: all the relations between data-elements can be computed so that XmlLex data can be interfaced with any lexical database (e.g. the ISO Lexical Markup Framework project) or any linguistic applications based on a clearly specified model
 - Flexibility: by applying XML rules of subsetting, as defined in ISO 16642 annex C, it allows subsets to be specified according to specific needs.
 - 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. XmlLex can be implemented as an XML schema and operated by commonly available XML editors and by XSL stylesheets

XmlLex uses existing data elements defined in ISO 12620. It also 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.

Information about the XmlLex-model is to be found at <http://www.xmllex.net/lexicography/xmllexintro.pdf>

3.3 The informative section provides useful recommendations and presents

- a) Means of presentation of entries in print and electronic dictionaries with layout aids and compacting mechanisms. Depending on the medium, one or more layout aids are used. Text formatting and typographical conventions can be used for all media to differentiate entry elements. The tabular form is especially recommended for dictionaries with more than two languages. For electronic media, data bank templates can be used as a specific form of presentation.
 - b) Numbering systems
 - c) Tables of the functions of lexicographical symbols
- a) Examples of XML-Encoding

dam [dʰ40;m], (past tense & past participle **dammed**, continuous form **damming**)

dam noun

* [barrier] barrage *m* (de retenue)

* [reservoir] réservoir *m*

* [animal] mère *f*

dam transitive verb

construire un barrage sur

dam up

separable transitive verb

* *literal* construire un barrage sur

* *figurative* [feelings] refouler, ravalier;

[words] endiguer

Figure 1: Example of a general bilingual dictionary entry

```
1. <Dictionary sourceLanguage="en" targetLanguage="fr">
2. <DictionaryEntry identifier="LEX_ex.1">
3. <HeadwordCtn>
4. <Headword>dam</Headword>
5. <GrammaticalNote> past tense & past participle <Fragment>dammed</Fragment>, continuous form
<Fragment>damming</Fragment></GrammaticalNote>
6. <PronunciationCtn>
7. <Pronunciation> dph40;m</Pronunciation>
8. </PronunciationCtn>
9. </HeadwordCtn>
10. <HomographRecord>
11. <PartOfSpeech value="noun"/>
12. <SenseRecord>
13. <TranslationCtn>
14. <RangeOfApplication>barrier</RangeOfApplication>
15. <Translation>barrage<GrammaticalGender value="masculine"/><Optional>de retenue</Optional></Translation>
16. <SearchForm>barrage</SearchForm>
17. <SearchForm>barrage de retenue</SearchForm>
18. </TranslationCtn>
19. </SenseRecord>
20. <SenseRecord>
21. <TranslationCtn>
22. <RangeOfApplication>reservoir</RangeOfApplication>
23. <Translation>réservoir<GrammaticalGender value="masculine"/></Translation>
24. </TranslationCtn>
25. </SenseRecord>
26. <SenseRecord>
27. <TranslationCtn>
28. <RangeOfApplication>animal</RangeOfApplication>
29. <Translation>mère<GrammaticalGender value="feminine"/></Translation>
30. </TranslationCtn>
31. </SenseRecord>
32. </HomographRecord>
33. <HomographRecord>
34. <PartOfSpeech freeValue="transitive verb"/>
35. <SenseRecord targetLanguage="fr">
36. <TranslationCtn>
37. <Translation>construire un barrage sur</Translation>
38. </TranslationCtn>
39. </SenseRecord>
40. </HomographRecord>
41. <MultiWordUnitCtn>
42. <MultiWordUnit>dam up</MultiWordUnit>
```

```

43. <PartOfSpeech freeValue="separable verb"/>
44. <SenseRecord>
45.   <SenseQualifier>figurative</SenseQualifier>
46.   <TranslationBlock>
47.     <RangeOfApplication>feeling</RangeOfApplication>
48.     <Translation>refouler</Translation>
49.     <Translation>ravaler</Translation>
50.   </TranslationBlock>
51.   <TranslationCtn>
52.     <RangeOfApplication>words</RangeOfApplication>
53.     <Translation>endiguer</Translation>
54.   </TranslationCtn>
55. </SenseRecord>
56. </MultiWordUnitCtn>
57. </DictionaryEntry>
58. </Dictionary>

```

Figure 2 : Encoding

3.4 ISO 1951 is a useful guide

- a) for the development of new dictionaries in all publishing media
- b) for the production of new dictionary products based on existing data collections (extraction and merging of dictionaries)
- c) for the integration of dictionaries from other series in an existing dictionary programme for providing dictionary data in other language tools (Translation Software, Translation Memory tools)
- d) for optimizing the editorial work flow

It is available from every national standardisation body such as DIN in Germany, AFNOR in France etc.

4. Conclusion

Standardization of dictionary products is beneficial for dictionary publishers and all their business partners from the point of view of data-manipulation and exchange. New market demands require new dictionary products and dictionary users need global language solutions. Interoperability of data saves time and cuts costs. The importance of standardization is growing as dictionary prices have to be kept low because of the increasing development of free internet products and consequently of price consciousness among dictionary users. The dictionary market is becoming global and it is becoming increasingly necessary for there to be a basic common language which is used throughout dictionary publishing. While standardization may not be as indispensable as in other branches of industry, conformity to an ISO-standard will improve quality assurance and contribute to a better appreciation of dictionary products.

Standardization is not intended to regulate dictionary production, however. Dictionary publishers should not be worried about the regulation of choice, content or layout of dictionary products. All the players in the dictionary industry are invited to participate in the development of lexicographical standards by providing expert advice to their national standardization bodies. This will ensure that these standards continue to reflect market requirements

and allow everyone to stay informed about new developments in this field.

References

[DEROUIN; LE MEUR, 2002] Report on the revision of the lexicographical Standard ISO 1951: Presentation/Representation of entries in dictionaries, *Proceedings of the third International Conference on Language Resources and Evaluation, LREC 2002*