XML - A Solution For LDBs, Eds and MRDs?

MARIE-HELENE CORREARD[¶] - MATHIEU MANGEOT-LEREBOURS[¶ß]

Xerox Research Centre Europe, 6 Chemin de Maupertuis, 38240 Meylan[¶]

GETA- CLIPS, IMAG, BP 53, 38041 Grenoble Cedex 9 France[ß]

Abstract

Lexical resources are a key element of NLP applications. They come from different sources and in different formats. Users of lexical resources have to produce their own or process the available ones in order to make them compatible with their own environments and applications. The use of DML (Dictionary Markup Language), could make working with lexical resources easier. The nature of the resources available will be briefly examined, then the solution that adopted to "unify" them will be presented with concrete examples to illustrate the approach.

Introduction

Dictionaries and other lexical resources are a key element of NLP (natural language processing) applications. Often they come from different sources and in different formats.

Currently, users of lexical resources must either write their own dictionaries, not a trivial task, or process them to make them compatible with their own environments and applications.

This paper describes how the use of DML could simplify working with lexical resources, make possible their reuse and improve their shareability. Firstly, the available resources will be briefly examined, secondly the solution that was adopted to "unify" these resources will be shown, including a look at constraints and requirements and thirdly concrete examples will be presented before concluding on the future of such an approach.

1. Starting point

Dictionaries used in NLP vary greatly, according to the final purpose of the application they are used by, but essentially they are of two kinds: either they were designed specifically for computer applications or they were written for human users. In this paper the word dictionary is used to refer to ordinary, general language dictionaries created for humans whose texts exist in electronic format.

First, a quick survey was performed to find out more about what kind of lexical resources were needed and how the available ones could be improved. Accessibility of the data was one of the items that came high on the list of possible improvements. One way to make the data more accessible without altering the contents was 'standardisation' of the format. Then, several dictionaries were used to test this approach.

A brief description of each one of them is given below.

- The Oxford-Hachette French Dictionary (OHD) is a bilingual dictionary. It consists of two sections roughly equal in size: French-English and an English-French. The dictionary is encoded in SGML. Its structure is fairly complex; a great number of elements are embedded.
- The New Oxford Dictionary of English is a new monolingual dictionary. It contains most of the elements of a monolingual dictionary, including etymology, sample material and encyclopedic information.
- The Password semi-bilingual English-French dictionary consists of one developed semi-bilingual section and one French index which cross-refers to English entries in which the French word is given as a translation.
- The FeM (French-English-Malay) dictionary supplies English and Malay translations of the French entry. English was used as a help for lexicographers during the dictionary development.

2. Solution

In order to get around the difficult it was decided to adopt, at a higher level, a common format for all the dictionaries. This common standard format had to be easily readable and to make it possible to keep all the information which was present in the original format. Then tools based on this common format could be built.

This section describes the requirements and gives an explanation of how the common format was defined.

2.1. Requirements

The design of the solution was driven by a list of specifications. These specifications came from previous experiments in computational lexicography and lexicology such as the indexation of the French-English-Malay dictionary [Lafourcade96], the building of the French-UNL database [Mangeot97], [Mangeot98] or the computerization of the I. Melčuk's Combinatory and Explanatory Dictionary of contemporary French [Sérasset98].

It was essential to find a way to preserve all the information present in the original format of the dictionary during the conversion. The dictionaries might be used for various applications, so it was not possible to predict in advance the kind of information that should be kept or left out.

In order to guarantee a maximum of compatibility for the new format and to reuse previous work in the domain, the obvious approach was to use existing norms and standards as much as possible. Furthermore since most of the resources available at the time were encoded in SGML [ISO86] it seemed reasonable to try and chose a format which did not need a lot of conversion work.

On the one hand, the power of object programming as well as that of relational database query facilities were attractive. On the other hand, the opacity of data repositories and portability problems were decisive factors for the choice of a textual format, either for storage or exchange when manipulating dictionaries.

2.2. Format adopted

All these considerations led to the choice of eXtended Markup Language (XML) [Connoly97] for encoding the dictionaries. XML is a W3C recommendation [W3C98a]. It is also UNICODE [ISO93] compliant. XML makes it possible to represent a large variety of information. All these features guarantee readability, perenniality and compatibility with an increasing number of tools.

Furthermore, because XML is a subset of SGML, the conversion of SGML dictionaries, well formed according to XML, into XML is unnecessary. Also, XML is a textual format, therefore it will always be possible to read the original files encoded in XML.

Now that the format is defined, a problem remains: how to encode the structure of the dictionaries? Two alternatives are possible:

Using a general DTD

The first option was to define a general DTD. This DTD would have to be generic enough to allow the encoding of all the dictionaries currently available. The conception of tools would then be easy because all of them would be based on the same DTD. This solution, despite its simplicity, was rejected because it was not possible to convert all the dictionaries following the same DTD without loss of information. It was also obvious that each dictionary has its own particular structure and, except for some rare cases, it was impossible to convert all the contents of one dictionary into another dictionary structure.

Keeping the original structure

An easier solution was to keep the original structure of each dictionary. A difficulty then rises at the stage of designing a tool for more than one dictionary. It appears quickly that each dictionary requires its specific tool. Therefore this solution does not solve all the problems.

A hybrid solution

A hybrid solution was then envisaged. XML is designed to be used with namespaces [W3C99]. It seemed appropriate to introduce a new one, specialized for dictionaries: DML for Dictionary Markup Language. This namespace is used for a hierarchised restricted set of tags. This set is composed of tags describing the same information in different dictionaries. For example, `<dml:entry>` always refers to an entry or `<dml:headword>` to the headword of an entry.

When some information in a dictionary cannot be represented with a tag from the DML set, it is still possible to copy it from the source file without transforming it. Specific tools manage it as they would the original file. If this type of information is present across several dictionaries, a new tag is then added to the DML set. The DML tags are used by the various tools as points of reference in an unknown converted dictionary.

The set of tags is composed of tags coming from standards like TEI/MARTIF [Ide95], [Johnson95], [Melby94], [ISO95]; GENELEX/EAGLES [GENELEX93] and GENETER [GENETER98]. The matching between a DML tag and an original tag is performed by a linguist to avoid possible conflicts between the tags.

Here is an alpha version of the DML tagset. The tags were chosen on the basis of their frequency. If an element occurred in more than 2 dictionaries (this figure may change at a later stage) it was added to the tagset. The tagset itself is evolving as new dictionaries are explored and converted.

`<dml tag>` (tei equivalent)

```
<dictionary
     name=""
     date=""
     source-language=""
     target-language="">
<letterset letter="">
<entry>     (entry)
     <headword homograph-number="">                    (hom)(orth)
     <headword-variant>                                (oVar)
     <pronunciation>                                   (pron)
        <phonetic encoding="">
     <etymology>                                       (etym)
     <syntactic-cat>                                   (sense level="1")
        <part-of-speech>                               (pos)(subc)
        <semantic-cat>                                 (sense level="2")
        <indicator>                                    (usg)
        <label>                                        (lbl)
        <definition>                                   (def)
        <example>                                      (eg)
        <translation language="">                      (trans)(tr)
        <collocate>                                    (colloc)
        <xref>                                         (xr)
           <x-headword homograph-number="">
           <x-syntactic-cat>
     <note> (note)
```

The next section shows how conversion was performed using DML tagset and how the results were exploited.

3. Examples

3.1. Conversion

According to the source format of the dictionary, there are three types of conversion. The simplest type occurs when the source format is well-formed SGML; the second type, when all the information is under the form attribute-value, and the third, the most complex one, relates to typographic formats which have to be parsed in order to extract as much information as possible.

3.1.1. Well-Formed SGML

If the dictionary is encoded in SGML and “well formed” in the XML sense (ie all opening tags are closed and the file is parsable by a context-free grammar), the conversion is very easy, since the structure is already, de facto, in XML. In this case, the only tasks are: conversion of characters into UNICODE characters set, changing the file encoding to UTF-8 and adding as much DML tags as possible.

If some information is redundant between the DML tag and the original tag, the latter is replaced and a note is kept of the changes. If the replacement DML tag is less precise than the original one, the original one remains in the text, embedded inside the DML tag. If some information is not in the same format (eg an element instead of attribute), it is altered to conform to DML and a note is kept of the changes.

The example is taken from the OHD [OUP-H94].

First, here is a sample of the entry `abréger` in original format:

```
<se><hw>abr&ea.ger</hw><pr><ph>abKeZe</ph></pr><hg><ps>vtr</ps></hg><s2
num=1>(<ic>rendre court</ic>) to shorten [<co>mot,expression</co>]; to summarize
[<co> texte, discours</co>]; <sl>&hw. &oq.t&ea.l&ea.vision&cq. en
&oq.t&ea.l&ea.&cq.</sl> to shorten &oq.television&cq. to &oq.TV&cq; (...) </se>
```

The headword `abréger` is followed by its pronunciation in Alvey notation, its part of speech `vtr` and its English translation `to shorten` then `to summarize`; the translations are differentiated by context (collocates). An example follows: `abréger ‘télévision’ en ‘télé’` then its translation: `to shorten ‘television’ to ‘TV’`. Translations were left untagged.

The sample below is the same entry with DML tags. Modified parts are in italics

```
<dml:entry><dml:headword>abr&#xE9;ger</dml:headword>
<dml:pronunciation><dml:phonetic encoding="ALVEY">
```

```
abKeZe</dml:phonetic></dml:pronunciation><hg><dml:part-of-
speech>vtr</dml:part-of-speech></hg><dml:semantic-sense>
<ic>rendre court</ic> to shorten <co>mot, expression</co>; to
summarize <co>texte, discours</co>; <sl>&hw;
&oq;t&#xE9;l&#xE9;vision&cq; en&oq;t&#x E9;l&#xE9;&cq;</sl> to
shorten &oq;television&cq; to &oq;TV&cq;;</dml:semantic-
sense></dml:entry>
```

3.1.2. Attribute-Value

When the original dictionary is represented by series of attribute-value pairs, the conversion remains simple. It consists in devising a DTD for the dictionary and converting the attribute-value pairs into <tag>value</tag>. Characters and file encoding are also converted.

The example for `abréger` below is taken from the FEM [Lafourcade96].

```
(:fem-entry
(:ENTRY "abréger")
(:FRENCH_PRON "abre-je-")
(:FRENCH_CAT "v.tr.")
(:FRENCH_GLOSS "un texte")
(:ENGLISH_EQU "to shorten")
(:ENGLISH_EQU "to abridge")
(:MALAY_EQU "memendekkan")
(:MALAY_EQU "meringkaskan")
)
```

The entry after conversion looks as follows:

```
<dml:entry><dml:headword>abr&#xE9;ger</dml:headword>
<dml:pronunciation><dml:phonetic encoding="GETA">abre-je-
</dml:phonetic></dml:pronunciation>
<dml:part-of-speech>v.tr.</dml:part-of-speech>
<FRENCH_GLOSS>un texte</FRENCH_GLOSS>
<dml:translation language="en">to shorten</dml:translation>
<dml:translation language="en">to abridge</dml:translation>
<dml:translation language="ml">memendekkan</dml:translation>
<dml:translation language="ml">meringkaskan</dml:translation>
</dml:entry>
```

3.1.3. Typographic Format

The most complex case occurs when a dictionary needs to be converted from a typographic format such as typesetters' tape, word processor, HyperText Markup Language (HTML). One particularly complex aspect these formats is that they represent knowledge designed to be readable by humans who can infer structure and disambiguate senses easily. In order to extract the information and, above all, build a deep structure for such a dictionary, a powerful tool built by [Haï98] called RECUPDIC was used. This tool combines two methods: a string transducer and a special tree parser. The structure of the result is described as a grammar and the tool extracts as much information as possible.

Here is the entry `babble` from Password semi-bilingual English-French dictionary:

```
>U43<babble >U1<[\B.270babl] >U2<verb >U23<1\N>U1<to talk indistinctly or
foolishly: >U2<What are you babbling about now? >U8< bafouiller, bavarder\L
>U23<2\N>U1<to make a continuous and indistinct noise: >U2<The stream babbled
over the pebbles.>f5h8<. >U8< gazouiller\L
```

and the same entry converted in XML:

```
<dml:entry><dml:headword>babble</dml:headword>
<dml:pronunciation><dml:phonetic
encoding="Password">'babl</dml:phonetic ></dml:pronunciation>
<dml:syntactic-cat><dml:part-of-speech>verb</dml:part-of-speech>
<dml:semantic-cat num="1"><dml:definition>to talk indistinctly or
foolishly </dml:definition><dml:example>What are you babbling
about now?</dml:example>
<dml:translation>bafouiller</dml:translation>
<dml:translation>bavarder</dml:translation><dml:semantic-cat
```

```
num="2"><dml:definition>to make a continuous and indistinct
noise</dml:definition><dml:example>The stream babbled over the peb
bles.</dml:example><dml:translation>gazouiller</dml:translation></
dml:semantic-cat>
</dml:syntactic-cat></dml:entry>
```

A summary of conversion operations is described in the table below:

Dictionary	Format	Size (in bytes)	Time spent
OHD - en/fr	SGML	17 Mb	1 day
OHD - fr/en	SGML	15 Mb	1/2 day
NODE - en	SGML	38 Mb	1 day
Password - en/fr (letterset)	typesetter's tape	300 Kb	5 days
Password - en/ja (letterset)	typesetter's tape	250 Kb	1 days
FeM - fr/en/ml	attribute-value	9 Mb	1/2 day

3.2. Usage

Some dictionaries do not contain information corresponding to some of these DML tags and some others contain information that is not covered by the DML tagset. However, as the tools are based on the DML tagset they will always find those elements which are represented by the tagset and present in a given dictionary, eg the `<dml:headword>` tag will always refer to the headword of an entry. Tools must be evolutive, to take into account the changes of DML.

As the resources are encoded in XML, all XML-compliant tools can be used. For example, a dictionary can be exported into a specific format with the help of XSL [W3C98b] or DSSSL [ISO96]. Tree transformations operations become possible. Dictionary readability can be improved with an associated stylesheet and an XML-compliant browser. Because of the relative youth of XML, few good tools are available yet but there should be more in the near future.

Two applications realised with XML/DML-encoded dictionaries are presented below.

3.2.1. Dicoweb

Dicoweb is a dictionary webserver. It was designed for human usage. It is used for experiments and research[1]. For legal reasons, not all these dictionaries are accessible to the public. The Dicoweb user first selects the source language of the headword she is looking up, then she selects the target language(s). The user can select as many target languages as are available. Before consulting the dictionaries, she can process the headword through a morphological analyser. Two buttons, labelled "previous" and ""next", give access to the preceding and following entries in dictionary order. For clarity reasons, each language is visualised in a specific colour and font.

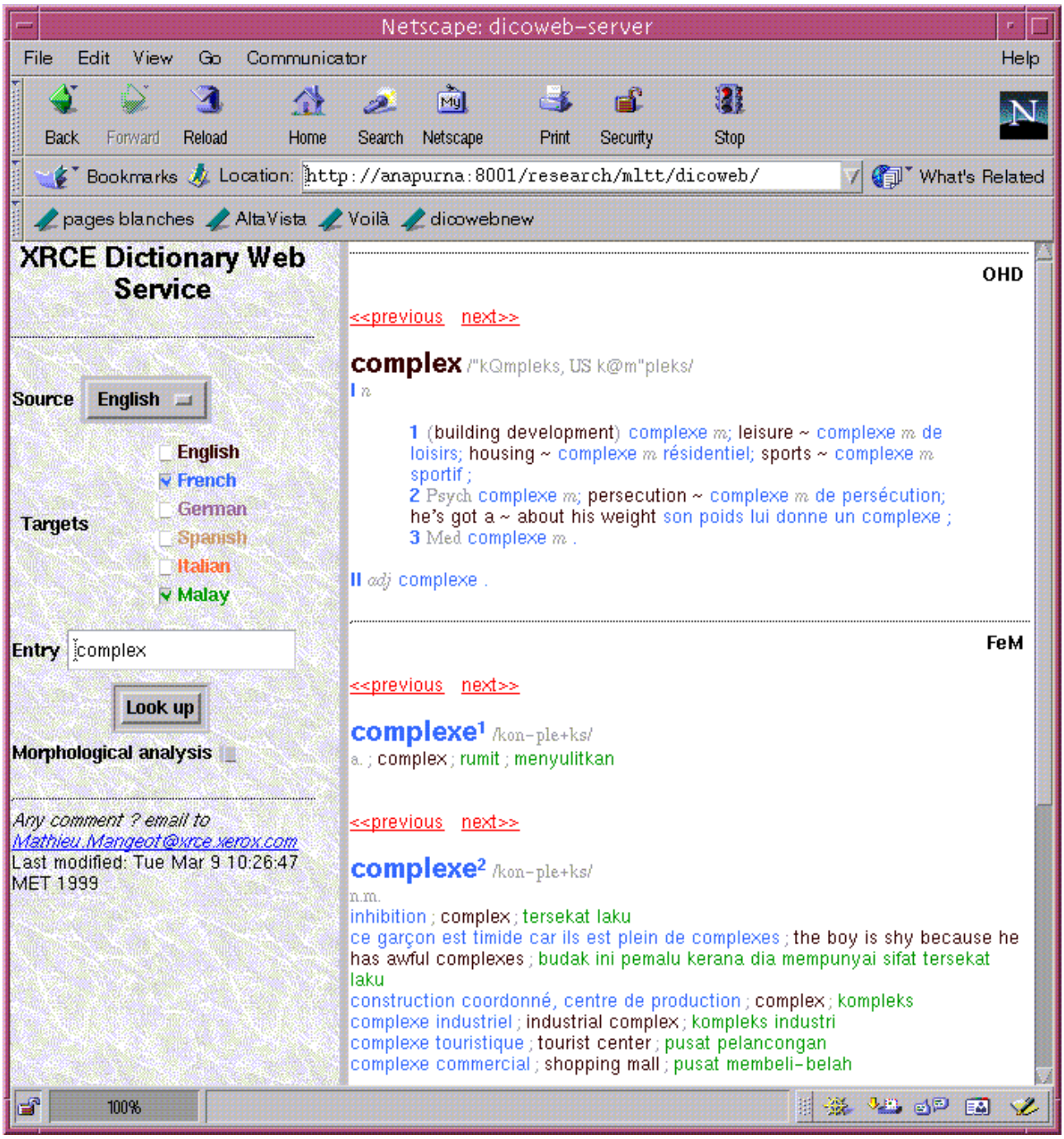
A Common Gateway Interface (CGI) script written in PERL [Wall91] works as a link between the user, the morphological analysers and the dictionaries. Dictionaries are selected according to the languages checked by the user. The files corresponding to the dictionaries are browsed by the script looking for a PERL regular expression such as: "`/<dml:headword[^>]*>$ENTRY<\/dml:headword>/`" where `$ENTRY` represents the headword entered by the user.

XML browsers are not widespread so it was decided to convert the result into HTML before sending it back to the user. The pages are built on the fly, with no breach of copyright and the possibility to modify directly the rendering of the final page.

Adding a new resource simply means adding the location of its file and the languages it covers to the script.

---

[1] URI: http://silfide.imag.fr



3.2.2. XeLDA

XeLDA (Xerox Linguistic Development Architecture) was built to provide developers and researchers with a common development architecture for the open and seamless integration of linguistic services. These services may include such applications as translation aids, syntax checking, terminology extraction, and authoring tools in general.

The above sample of Password English-French semi-bilingual dictionary was integrated into XeLDA. The dictionary was converted to comply with XeLDA DTD with the help of an XSL stylesheet. Here is the result of the transformation:

```
<xbdict>
<entry>
<headword><spl>babble</spl></headword>
<hwinfo><pronunciation><phonetic>['babl]</phonetic>
</pronunciation></hwinfo>
<syntactic><senseinfo><pos>verb</pos></senseinfo>
<semantic>
<subsense>to talk indistinctly or foolishly: What are you babbling about
now?</subsense>
<subsense><trans>bafouiller</trans></subsense>
<subsense><trans>bavarder</trans></subsense></semantic>
```

```
</syntactic>
</entry>
```

Conclusion

The work presented in this paper is still under development. The current results are satisfactory. However, further investigation is needed to establish the adaptability and coverage of DML. In the longer term it is planned to build new tools that will enable a user to set personal parameters according to the task at hand.

References

[Atkins94] **Atkins, B. T. S. and Zampolli, A. (1994)** *Computational Approaches to the Lexicon.* Oxford University Press, 480 p.

[Bauer94] **Bauer D., Segond F. & Zaenen A. (1994)** *Enriching a SGML-tagged Bilingual Dictionary for Machine-Aided Comprehension.* Technical Report, Xerox Research Center Europe, 21 p.

[Boguraev89] **Boguraev et al. (1989)** *Computational lexicography for natural language processing.* B. Boguraev & T. Briscoe, ed., Longman, London & New York, 310 p.

[Briscoe93] **Briscoe et al. (1993)** *Inheritance, Defaults and the Lexicon.* Cambridge University Press, Cambridge.

[Connoly97] **Connolly, D. (1997)** *XML Principles, Tools and Techniques* World Wide Web Journal, Volume 2, Issue 4, Fall 1997, O'REILLY & Associates, 250 p.

[Fedder91] **Fedder et al. (1991)** *Typed Feature Logic and its role in MULTILEX.* Centre for Computational Linguistics, UMIST, november 1991, 30 p.

[GENELEX93] **GENELEX (1993)** *Projet Eureka Genelex, modèle sémantique.* Rapport Technique, Projet Eureka, Genelex, 4 march 1994, 185 p.

[GENETER98] **GENETER (1998)** M*odèle générique de représentation des données terminologiques.* URI: http://www.uhb.fr/Langues/Craie/balneo/nen3.zip.

[Haï98] **Haï, D. (1998)** *Techniques génériques d'accumulation d'ensembles lexicaux structurés à‡ partir de ressources dictionnairiques informatisées multilingues hétérogènes.* Thèse de nouveau doctorat, Spécialité Informatique, Institut National Polytechnique de Grenoble, 168 p.

[Heid92] **Heid et al. (1992)** *Extracting linguistic information from machine-readable versions of traditional dictionaries, a metalexicographic method and some tools.* Proc. COMPLEX'92, Conference on Computational Lexicography an Text Research, Budapest, Hongrie, Linguistics Institute, Hungarian Academy of Sciences, Budapest, pp 161-174.

[Ide95] **Ide, N. and Veronis, J. (1995)** *Text Encoding Initiative, background and context.* Kluwer Academic Publishers, 242 p.

[ISO86] **ISO (1986)** *IISO 8879 (SGML) Information processing -- Text and office systems -- Standard Generalized Markup Language*, Geneva, 155 p.

[ISO93] **ISO (1993)** *ISO/IEC 10646 (UNICODE) Information technology -- Universal Multiple-Octet Coded Character Set (UCS)*, Geneva, 754 p.

[ISO95] **ISO (1995)** *ISO DIS 12620 (MARTIF) Terminology - Computer Applications– Data Categories.* ISO TC 37/SC 3/WG I, Geneva.

[ISO96] **ISO (1996)** *IISO/IEC 10179 (DSSSL) Information technology -- Processing languages -- Document Style Semantics and Specification Language*, Geneva, 292 p.

[Johnson95] **Johnson, E. (1995)** *The Text Encoding Initiative.* TEXT Technology vol. 5, n°3, Autumn 1995, pp 174-175.

[Lafourcade96] **Lafourcade M. (1996)** *Structured Lexical data: how to make them widely available, useful and reasonable protected? - a practical example with a trilingual dictionary.* Proc. COLING-96, Copenhagen, Denmark, Vol 2/2, pp. 1106-1110.

[Mangeot97] **Mangeot-Lerebours, M. (1997)** *Outils pour lexicographes naïfs (en informatique).* DEA Informatique Systèmes et Communications, GETA-CLIPS-IMAG, Université Joseph Fourier Grenoble 1, 19/06/97, 58 p.

[Mangeot98] **Mangeot-Lerebours M. (1998)** *Conception, implémentation et indexation de BaLeM, une base lexicale multilingue.* Proc. TALN'98, Paris, pp 215-217.

[Melby94] **Melby et al. (1996)** *The Machine Readable Terminology Interchange Format (MARTIF), Putting Complexity in Perspective*, Termnet News, vol 54/55.

[OUP-H94] **Oxford-Hachette (1994)** *Le dictionnaire Hachette-Oxford* Oxford University Press & Hachette, 1950 p.

[Sérasset98] **Sérasset, G. & Mangeot-Lerebours M. (1998)** *L'édition lexicographique dans un système générique de gestion de bases lexicales.* NLP-IA'98 Moncton, NB, Canada, vol 1/2, pp 110-116.

[Thurmair98] **Thurmair, Gr. et al. (1998)** *The Open Lexicon Interchange Format (OLIF).*
URI: http://www.otelo.lu/seite2.htm

[W3C98a] **W3C (1998)** *XML 1.0* URI: http://www.w3.org/TR/1998/REC-xml-19980210

[W3C98b] **W3C (1998)** *XSL 1.0* URI: http://www.w3.org/TR/1998/WD-xsl-19981216

[W3C99] **W3C (1999)** *XML namespaces*
URI: http://www.w3.org/TR/1999/REC-xml-names-19990114

[Wall91] **Wall L. & Schwartz R. L. (1991)** *Programming PERL*, O'Reilly and Associates.