# An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language.

**Mathieu MANGEOT-LEREBOURS**
Software Research Division, NII
Hitotsubashi, 2-1-2 Chiyoda-ku
101-8430 Tokyo, Japan
mangeot@nii.ac.jp

## Introduction

Lexical data resources are growing rapidely thanks to the Internet. Unfortunately, despite numerous existing standards like TEI, MARTIF, GENELEX, EAGLES/PAROLE, etc. each resource has its own format and own structure. Furthermore, the existing lexical data is generally developed for a specific purpose and can't be reused easily in other applications.
In this paper, we intend to define a complete framework for developing multilingual lexical database for multipurpose. The framework is generic enough in order to accept a wide range of dictionary structures and proposes for manipulating heterogeneous dictionaries a set of common pointers into these structures.
We will first present the organisation of Dictionary Markup Language (DML) framework.
Then we will describe more precisely the DML language based on XML schemata.
Next, we explain how to describe dictionary macro and microstructures with the DML.
Lastly, we will explain our concept of common pointers defined in a Common Dictionary Markup (CDM) set.

## 1. Presentation of the DML Framework

The DML Framework described first by Mangeot-Lerebours (2001) is a complete framework for the consultation of heterogeneous dictionaries, cooperative construction of new dictionaries and communication with other lexical databases or lexical data client and supplier applications. The framework is completely generic in order to manage heterogeneous dictionaries with their own proper structures.
The consultation of heterogeneous dictionaries is possible as soon as they are encoded in XML, consultation of other resources via remote servers through API, possibility of adding pre-consultation help modules such as spell checking and morphological analysis before consultation or post-consultation modules like syntethisers, conjugation of verbs, learning drills, etc. Possibility of automatic consultation of the database via client API.
The construction of new dictionaries can be done by a community of contributors and validated by a group of head lexicographers specialists.
The management of user profiles, preferences and weights for consultation, annotation and edition of lexical data with inheritance and sharing possibilities among groups of users is also handled by the framework.
The <database> element describes a lexical database and lists the dictionaries that are stored in it.

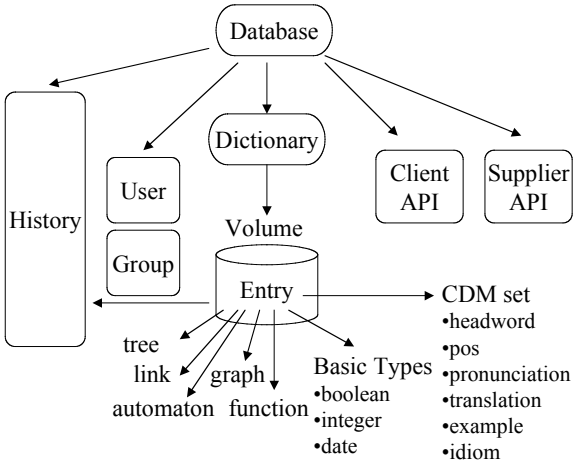


**Figure 1.** Logical Organisation of a Lexical Database

The <dictionary> element describes the metadata linked to ther dictionary. It links all the volumes of the dictionary.
The <volume> element describes a dictionary part. The content is principally a list of dictionary entries. For example, a bilingual bidirectional French-English dictionary will be described by only one <dictionary> element. The French->English entries will be in one <volume> element and the English->French entries in another <volume> element

## 2. The DML Language

### 2.1. The DML Namespace

To describe the structure of all the documents, elements, attributes and XML types, we use an XML namespace [XML Namespaces]. Our namespace is called DML for Dictionary Markup Language. The namespace URI points to an XML schema [XML

Schemas] describing the contents of the namespace. It is available online[1] to allow users to edit and validate their files online with an XML schema validator.

```
<MyElement
xmlns:dml="http://www-
clips.imag.fr/geta/services/dml">
  ...
  <dml:MyDescendant/>
  ...
</myElement>
```

Figure 2: Usage Example of the DML Namespace

## 2.2. DML Common Types and Attributes

For some information, we define type and attributes common to all DML elements. It allows to standardize the data. The XML schemata have originally simple predefined types. We selected and reused some in our definitions.

### *2.2.1. Dates and Time*

Dates are represented by the `date` DML attribute of the XML schema type dateType taken from the extended format of the ISO 8601 standard.

### *2.2.2. Response Delay*

The `delay` DML attribute of an element indicate the response delay when a request has been launched on this element.
This delay is a duration of the XML schema `durationType` type. For example, 5 seconds and 10 cents will be indicated "5.10S".

### *2.2.3. Unique ID*

The `id` DML attribute of an element is a unique ID in all the lexical database. It allows to create links between elements. It redefines the XML schema ID simple type.

### *2.2.4. Modifications History*

The modifications history of an element has a unique ID. The element links to its history thanks to the DML attribute `history` that gives the value of the history ID. The type redefines the XML schema ID simple type.

### *2.2.5. Languages Notation*

To note the various languages, we use the ISO-639-2/T (T for Terminology) [ISO98] standard that defines a 3 letter code for each language (French->fra; English->eng, Malay->msa, etc.). It is far more complete that the two letters code standard ISO-639-1. We also add our proper codes like "unl" for the UNL language. This codes list represents the lang DML type. The `lang` DML attribute is from this type.

[1] http://www-clips.imag.fr/geta/services/dml/

### *2.2.6. Documents Encoding*

To note the encodings of the various documents in the database, we define the `encodingType`. DML type. The values are those described by the IANA (Internet Assigned Number Authority) for the encodings. These are also the values used for MIME types (Multipurpose Internet Mail Extension). Among the most used, we find ASCII on 7 bits, ISO-8859-1 on 8 bits for latin languages, Shift-Jis on 8 or 16 bits for the Japanese, UTF-8 on 8 bits for UNICODE characters, etc.

### *2.2.7. Status of an Element*

The `status` DML attribute is used to indicate its status. The values can be among others `auto` if the element has been obtained automatically, `rough` if the element has not been revised and `revised` if so, etc.

# 3 DML Architecture

## 3.1. Macrostructure Definitions

To describe the macrostructure of our dictionaries as well as our lexical database, we use XML elements. We principally based our definitions on the LEXARD language defined by Serasset (1994) and added some information

### *3.1.1. Description of a Lexical Database*

To describe a lexical database, we use the `<database>` element formally described inthe DML schema.
The modifications of the `<database>` element and its descendants are stored in a document linked with the `history-ref` atttribute.
We add to LEXARD the possibility to define various users and groups in the database. At the beginning three groups are predefined `universe` contains all the users of the database, `administrators` contains the administrators of the database and `lexicologists` contains the users in charge of the control of the data.
The information relative to each user are stored in another element referenced by the `<user-ref>` element.
All the dictionaries of the database are referenced by pointers on XML documents that describe them. The pointers are the `href` attributes of the `<dict-ref>` elements grouped in the `<dictionaries>` element.

### *3.1.2. Description of a Dictionary*

To describe a dictionary, we use the `<dictionary>` element. The modifications information is stored in a document pointed by the `history-ref` attribute.
We indicate meta-information on the resources.

The elements <category> , <type> and <links> describe the dictionary macrostructure. The <category> element indicates the dictionary type (monolingual, bilingual, multilingual, interlingual). The <type> element indicates if the dictionaries are unidirectional, bidirectional or pivot based.
The <links> element indicates the links between the volumes of the dictionary. For example, if a dictionary is pivot based with 3 languages English, French and Malay, it contains 4 volumes Interlingual, English, French and Malay linked as follows:

```
<links>
  <link from="English"
to="Interlingual"/>
  <link from="French"
to="Interlingual"/>
  <link from="Malay"
to="Interlingual"/>
</links>
```

The dictionary volumes are referenced by their unque name. The <volumes> element gathers all the reference to the volumes files noted with the <volume-ref> element.
The source and target languages are indicated with the 3 letter code DML lang type.
The <content> element describes the content of the dictionary. The <domain> element indicates the domain covered by the dictionary (general, medecine, computer, etc.)
We indicate also the size of the dictionary in bytes by <bytes>, and the headword number by <hw-number>.
For the version management, we indicate the version number (<version>), the creation-date of the dictionary (<creation-date>) and the date of the integration of the dictionary into the database (<installation-date>).
For the non-DML resources, we need to indicate the file format (<format>) and the encoding (<encoding>). The encoding values are determined by the DML type encodingType.
We also indicate meta-information on the dictionary like the resource supplier (<source>), the owner (<owner>), the responsible at the database level (<responsible>), the rights attached to the dictionary (<legal>) and miscellaneous comments (<comments>).
The CDM (see chapter 4) elements list (<cdm-elements>) is stored with for each element, its real name in the resource and the maximal response delay. The (<corpus>) element is special, it allows to indicate that we search a string anywhere in the dictionary.

#### *3.1.3. Description of a Volume*

The <volume> elements gathers dictionary entries with the same source language. The modificaitons history is referenced with the history-ref attribute.

### 3.2. Microstructure Definitions

To represent dictionary microstructures, we propose to redefine in XML the structures defined with LINGARD (see serasset (1994).

#### *3.2.1. Trees*

To represent a dependance tree associated to the sentence "Le chat mange une souris.", for example, we can use a “decorated node” <dn> with attributes corresponding to the grammatical variables.

```
<dn ul="manger" time="present"
aspect="imperfectif">
  <dn ul="chat" determ="defini"
gnr="masc" pos="-1"/>
  <dn ul="souris" determ="indefini"
gnr="fem" pos="+1"/>
</dn>
```

#### *3.2.2. Links*

The definition of a link is done with the xlink standard [XLink 1.0]. We alslo add our attributes:

- The attribute type="bidirectionnal" or type="oriented" indicates if the link is bilingual or not;
- The attribute id is of the DML id type. It allows to attribute a unique id for each link;
- The content text of the element allows to tag the links.

Here is a link example:

```
<link type="oriented" id="l001"
href="example.xml#xpointer(//node[x
l:label='n002'])"/>
```

The reference to the external element is done with the href attribute. The reference is noted as a URI. If the object does not have a unique id (id), the link is described with the [XPointer] standard. Otherwise, it is pointed as follows:

```
<link type="oriented" id="l001"
href="example.xml#n002"/>
```

#### *3.2.3. Graphs and Automatons*

The xlink standard [XLink 1.0] is used to describe arcs. The arcs type is oriented type="oriented" or bijective type="bijective". The source and the target of the arc are noted with the node identifiers from="n001" and to="n002".
The definition of an automaton follows the definition of a graph. The starting node is noted with the xl:title="starting-node" attribute. The ending nodes are noted with the xl:title="ending-node" attribute.

#### *3.2.4. Functions*

The following example represents the lexical function [lambda]x1 (CausOper$_1$x0x1). The

results of its application to the French lexie DÉSESPOIR are the following: pousser, réduire quelqu'un au désespoir, jeter quelqu'un dans le désespoir, frapper quelqu'un de désespoir. The function is noted in XML as follows:

```
<function name="CausOper1">
  <arguments>
    <first value="desespoir"/>
  </arguments>
  <valgroup>
    <value>pousser</value>
    <value>réduire [qqun au
désespoir]</value>
    <value>jeter [qqun dans le
désespoir]</value>
    <value>frapper [qqun de
désespoir]</value>
  </valgroup>
</function>
```

#### *3.2.5. Feature Structures*

If the features are typed, the type is noted with an attribute. If the feature has several values, the element is duplicated.

```
<feature1
type="type1">valeur1</feature1>
<feature1
type="type2">valeur2</feature1>
```

#### *3.2.6. Sets and Disjonction*

Sets and disjunctions are defined directly at the XML schema level with the two elements <xsd:choice> and <xsd:sequence>

#### *3.2.7. Basic Types*

The basic type of an XML document is the character string. Thanks to XML schemata, we can use many other basic types like boolean, entity, decimal, float,etc.

## 4. The Common Dictionary Markup Subset

We defined a subset of DML element and attributes that are used to identify which part of the different structures represent the same lexical information. This subset is called Common Dictionary Markup (CDM).

### 4.1. Definition of the Subset

The DML framework may be used to encode many different dictionary structures. Indeed, two dictionary structures can be radically different. So, in order to handle such heterogeneous structures with the same tools, we need a common formalism. Standards like TEI [Ide95], MARTIF [Melby94], [ISO99]; GENELEX/EAGLES [GENELEX93] and [GENETER] aim to be universal but very few resources implement them.

We made a more pragmatic work with identifying the information in the existing resources as well as their meaning and naming them ina unique way in the DML namespace

This hierarchized subset is called Common Dictionary Markup and comes principally from the detailed examination of the FeM, DEC, OHD, OUPES, NODE, EDict, ELRA-MÉMODATA dictionaries and the 12th chapter of the TEI about dictionaries. It contains the most frequent elements found in these resources like the headword, the pronunciation, the part-of-speech, the examples, the idioms, etc. These elements have always the same semantics. For example, <dml:entry> always refer to a dictionary entry and <dml:headword> to the headword.

For some elements with closed lists of values, we define a list representing the intersection of the values and conversion rules for each resource. An example is the list of parts-of-speech for each language.

This set is in constant evolution. If the same kind of information is found in several dictionaries then a new element representing this piece of information is added to the CDM set. It allows tools to have access to common information in heterogeneous dictionaries by way of pointers into the structures of the dictionaries. The table 1 lists a first version of the CDM subset.

| **<CDM tag>** | **(TEI equivalent)** |
|---|---|
| <entry> | (entry) |
| <headword hn=""> | (hom)(orth) |
| <headword-var> | (oVar) |
| <pronunciation> | (pron) |
| <etymology> | (etym) |
| <syntactic-cat> | (sense level="1") |
| <pos> | (pos)(subc) |
| <lexie> | (sense level="2") |
| <indicator> | (usg) |
| <label> | (lbl) |
| <definition> | (def) |
| <example> | (eg) |
| <translation> | (trans)(tr) |
| <collocate> | (colloc) |
| <link href=""> | (xr) |
| <note> | (note) |

**Table 1: CDM Elements Subset**

### 4.2. CDM Correspondance Examples

When a resource is recuperated, a correspondance table is established between the original element names and CDM elements. The table 2 has been used for the FeM, OHD and NODE dictionaries.

| CDM | FeM | OHD | NODE |
|---|---|---|---|
| <entry> | <fem-entry> | <se> | <se> |
| <headword> | <entry> | <hw> | <hw> |
| <pronunciation> | <french_pron> | <pr><ph> | <pr><ph> |
| <etymology> | | | <etym> |
| <syntactic-sense> | | <sense n=1> | <s1> |
| <pos> | <french_cat> | <pos> | <ps> |
| <lexie> | | <sense n=2> | <s2> |
| <indicator> | <gloss> | <id> | |
| <label> | <label> | <li> | <la> |
| <example> | <french_ sentence> | <ex> | <ex> |
| <definition> | | | <df> |
| <translation> | <english_equ> <malay_equ> | | <tr> |
| <collocate> | | <co> | |
| <link> | <cross_ref _entry> | <xr> | <xg> <vg> |
| <note> | | <ann> | |

**Table 2:** Equivalents of the CDM elements in the FeM, OHD and NODE

## Conclusion

This framework has been extensively used for the Papillon project (see Serasset & Mangeot-Lerebours (2001)) of mutualized construction and consultation of a pivot multilingual lexical database. This experiments allowed us to correct and adapt some parts of the DML.
Nevertheless, the framework need to be opened to the public in order to receive feedback and comments. We plan to open a web site dedicated to the DML soon.

## References


GENELEX (1993) Projet Eureka Genelex, modèle sémantique. Rapport Technique, Projet Eureka, Genelex, mars 1994, 185p.

Nancy Ide & Jean Veronis (1995) Text Encoding Initiative, background and context. Kluwer Academic Publishers, 242p.

ISO (1998) ISO 639-1 & 2 Code for the representation of names of languages Part 1 & 2 Alpha-3 code. Geneva, Part 1: 17 p., Part 2: 90p.

ISO (1999) ISO DIS 12200 (MARTIF) Computer applications in terminology - Machine-readable terminology interchange format - Negotiated interchange.ISO TC 37/SC 3/WG I, Geneva, 118 p.

Mathieu Mangeot-Lerebours (2001) *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. Thèse de nouveau doctorat, Spécialité Informatique, Université Joseph Fourier Grenoble I, 27 September 2001, 280 p.

Allan Melby et al. (1996) The Machine Readable Terminology Interchange Format (MARTIF), Putting Complexity in Perspective. Termnet News, vol.54/55, pp.11-21.

Gilles Sérasset (1994) *Interlingual Lexical Organisation for Multilingual Lexical Databases in NADIA*. In Proc. COLING-94, Kyoto, 5-9 August 1994, M. Nagao ed. vol. 1/2 : pp. 278-282.

Gilles Serasset & Mathieu Mangeot-Lerebours (2001) *Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links*. Proc. NLPRS'2001 The 6th Natural Language Processing Pacific Rim Symposium, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan, 27-30 November 2001, vol 1/1, pp. 119-125.


## Bookmarks


GENETER modèle GENErique pour la TERminologie.

http://www.uhb.fr/Langues/Craie/balneo/demo_geneter.pl?langue=1

XLink 1.0 W3C Recommendation.

http://www.w3.org/TR/NOTE-xlink-req/

XML 1.0 eXtended Markup Language 1.0. W3C Recommendation.

http://www.w3.org/TR/REC-xml

XML Namespaces XML Namespaces. W3C Recommendation.

http://www.w3.org/TR/REC-xml-names

XML Schemas XML Schemas. W3C Recommendation.

http://www.w3.org/TR/xmlschema-0

XPath XPath Language. W3C Recommendation.

http://www.w3.org/TR/xpath

XPointer XML Pointer Language W3C Recommendation.

http://www.w3.org/TR/xpt

# Annexs

## Annex 1: XML Document Describing a Database

```
<database xsi:schemaLocation="http://www-clips.imag.fr/geta/services/dml
http://clips.imag.fr/geta/services/dml/dml.xsd"
name="GETA Lexical Database"
creation-date="22/10/99"
 owner="GETA">
  <partner-servers>
    <user-ref name="XRCE Analyser" href="xrce.xml"/>
  </partner-servers>
  <users>
    <user-ref name="Mathieu.Mangeot" href="mangeot.xml"/>
    <user-ref name="Mutsuko.Tomokiyo" href="tomokiyo.xml"/>
  </users>
  <groups>
    <group name="universe">
      <user-ref name="Mathieu.Mangeot"/>
      <user-ref name="Mutsuko.Tomokiyo"/>
    </group>
   <group name="lexicologists"><user-ref name="Mutsuko.Tomokiyo"/></group>
    <group name="administrators"><user-ref name="Mathieu.Mangeot"/></group>
  </groups>
  <dictionaries>
    <dict-ref name="FeM" href="FeM.xml"/>
    <dict-ref name="Papillon" href="papillon.xml"/>
  </dictionaries>
</database>
```

## Annex 2: XML Document Describing a Dictionary

```
<dictionary
xsi:schemaLocation="http://clips.imag.fr/geta/services/dml
http://clips.imag.fr/geta/services/dml/dml.xsd"
category="multilingual"
creation-date="21/1/97 00:00:00"
encoding="ISO-8859-1"
format="rtf"
hw-number="192460"
installation-date="23/06/99 15:04:00"
fullname="dictionnaire français-anglais-malais"
name="FeM"
owner="GETA"
type="unidirectional"
version="1">
  <languages>
    <source-language lang="fra"/>
    <target-language lang="eng"/>
    <target-language lang="msa"/>
  </languages>
  <contents>general vocabulary in 3 languages</contents>
  <domain>general</domain>
  <bytes>9106261</bytes>
  <source>ML, YG, PL, Puteri, Kiki, CB, MA, Kim</source>
  <legal>all rights belong to ass. Champollion</legal>
  <cdm-elements>
    <headword delay="1s"/>
    <pronunciation delay="5s"/>
```

```
    <part-of-speech delay="5s"/>
    <translation lang="eng" delay="5s"/>
    <translation lang="msa" delay="5s"/>
    <corpus delay="10s"/>
  </cdm-elements>
  <administrators><user-ref name="Kim, ML"/></administrators>
  <volumes><volume-ref name="FeM" href="fem_fr_en_ms.xml"/></volumes>
</dictionary>
```

### Annex 3: XML Document Describing a Volume

```
<volume
xsi:schemaLocation="http://clips.imag.fr/geta/services/dml
http://clips.imag.fr/geta/services/dml/dml.xsd"
 name="FeM_fr_en_ms"
source-language="fra">
  <entry>…</entry>
    ...
</volume>
```

### Annex 4: XML Document Describing a User

```
<user
xsi:schemaLocation="http://www-clips.imag.fr/geta/services/dml
http://www-clips.imag.fr/geta/services/dml/dml.xsd"
name="Mathieu MANGEOT"
creation-date="22/10/2001">
  <login>Mathieu.Mangeot</login>
  <password>toto</password>
  <email>Mathieu.Mangeot@imag.fr</email>
  <profiles>
   <competences>
      <eng level="good">translation</eng>
      <fra level="mother tongue">phonetic, collocations, examples, grammar</fra>
      <jpn level="beginner"/>
      <spa level="good">translation</spa>
    </competences>
    <interests><interest lang="hun,jpn"/></interests>
    <activities>
      <activity dictionary="FeM">interface</activity>
      <activity dictionary="Papillon">administration</activity>
    </activities>
  </profiles>
  <credits>10</credits>
  <annotations href="mangeot-ann.xml"/>
  <contributions>
    <contribution source="French.xml" href="mangeot-cnt1.xsl"/>
  </contributions>
  <requests href="mangeot-req.xml"/>
  <xml-stylesheet type="text/css" href ="mangeot-sty.css"/>
  <groups>
    <group-ref name="universe"/>
    <group-ref name="administrators"/>
  </groups>
</user>
```

### Annex 5: XML Document Describing a supplier API

```
<api type="supplier" category="consultation" name="JMDict_en-ja">
  <info>Dictionnaire japonais-anglais de Jim Breen</info>
  <url href="http://www.csse.monash.edu.au/cgi-bin/cgiwrap/jwb/wwwjdic"/>
```

```
  <protocol type="get"/>
  <delay min="1s" average="1s" max="2s" timeout="10s"/>
  <encoding input="UTF-8" output="EUC-JP"/>
  <format input="txt" output="html"/>
  <arguments>
    <element name="source-language">
      <complexType>
        <restriction base="string">
          <enumeration value="jpn"/>
          <enumeration value="eng"/>
        </restriction>
      </complexType>
    </element>
    <element name="headword" type="string"/>
    <element name="regex" type="boolean"/>
  </arguments>
  <result><element name="output" type="string"/></result>
</api>
```

## Annex 6: XML Document Describing a client API

```
<api type="client" category="consultation" name="getabase">
  <info>API de consultation de la base lexicale du GETA</info>
  <url href="http://www-clips.imag.fr/cgi-bin/geta/dicoweb
mailto:dicoweb@imag.fr
telnet://www-clips.imag.fr:2628"/>
  <protocol type="post get mailto DICT" login="anonymous"/>
  <encoding input="ASCII ISO-8859-1 UTF-8" output="UTF-8"/>
  <format input="txt xml" output="xml html txt"/>
  <arguments>
    <element name="name" type="string"/>
    <element name="source-language" type="lang"/>
    <element name="word-order" type="string"/>
    <element name="cdm-elements" type="string"/>
    <element name="context" type="positiveInteger"/>
    <element name="input" type="string"/>
  </arguments>
  <result>
    <element name="output">
      <complexType>
        <sequence><element name="article" type="articleType"/></sequence>
      </complexType>
    </element>
  </result>
</api>
```