HimalCo project: LMF and dictionaries

Céline Buret

October 23, 2015

1 What is LMF?

LMF is an ISO (*International Standard Organisation*) standard of Technical Committee 37 and Sub-Committee 4: ISO-TC37/SC4 24613.
This standard is suitable for general and specialised dictionaries, monolingual and multilingual. It describes a formal generic structure indepent of publication supports: from a well-formatted unique lexicographical source, we can obtain a printable form and an electronic form of data.
LMF follows a lexicographical approach centered on lemma. It is a two layers model: morphological and semantic.

LMF model is divided into two main parts: what is called the *core package*, a simple, rigid and mandatory skeleton, which is the heart of the model ; and extensions.

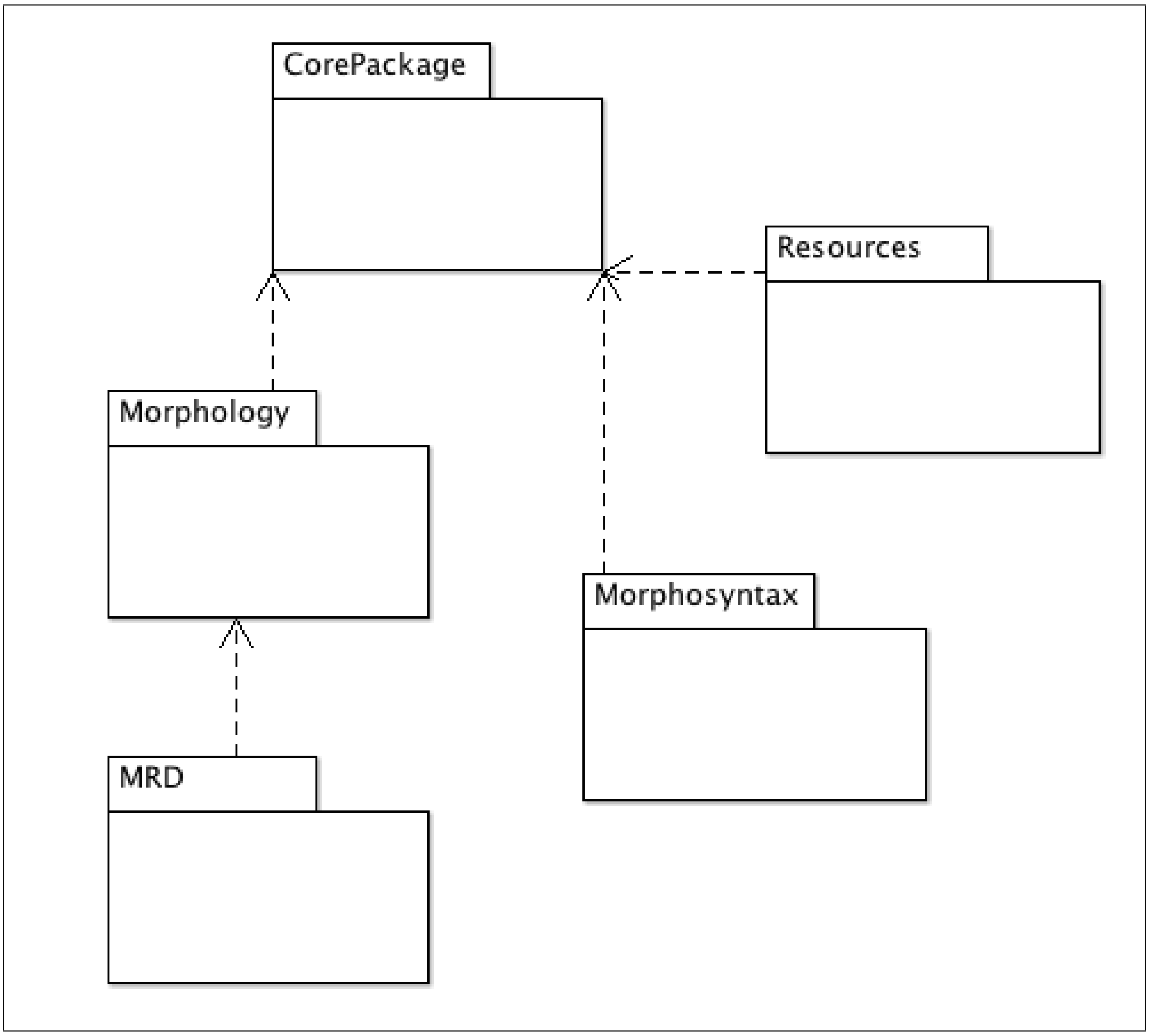

Figure 1: LMF packages

The *core package* is divided into two sub-systems:

- the lexical entry, *Lexical Entry*, and its different forms, *Form* (signifier) ;
- the sens or senses, *Sense* (signified).

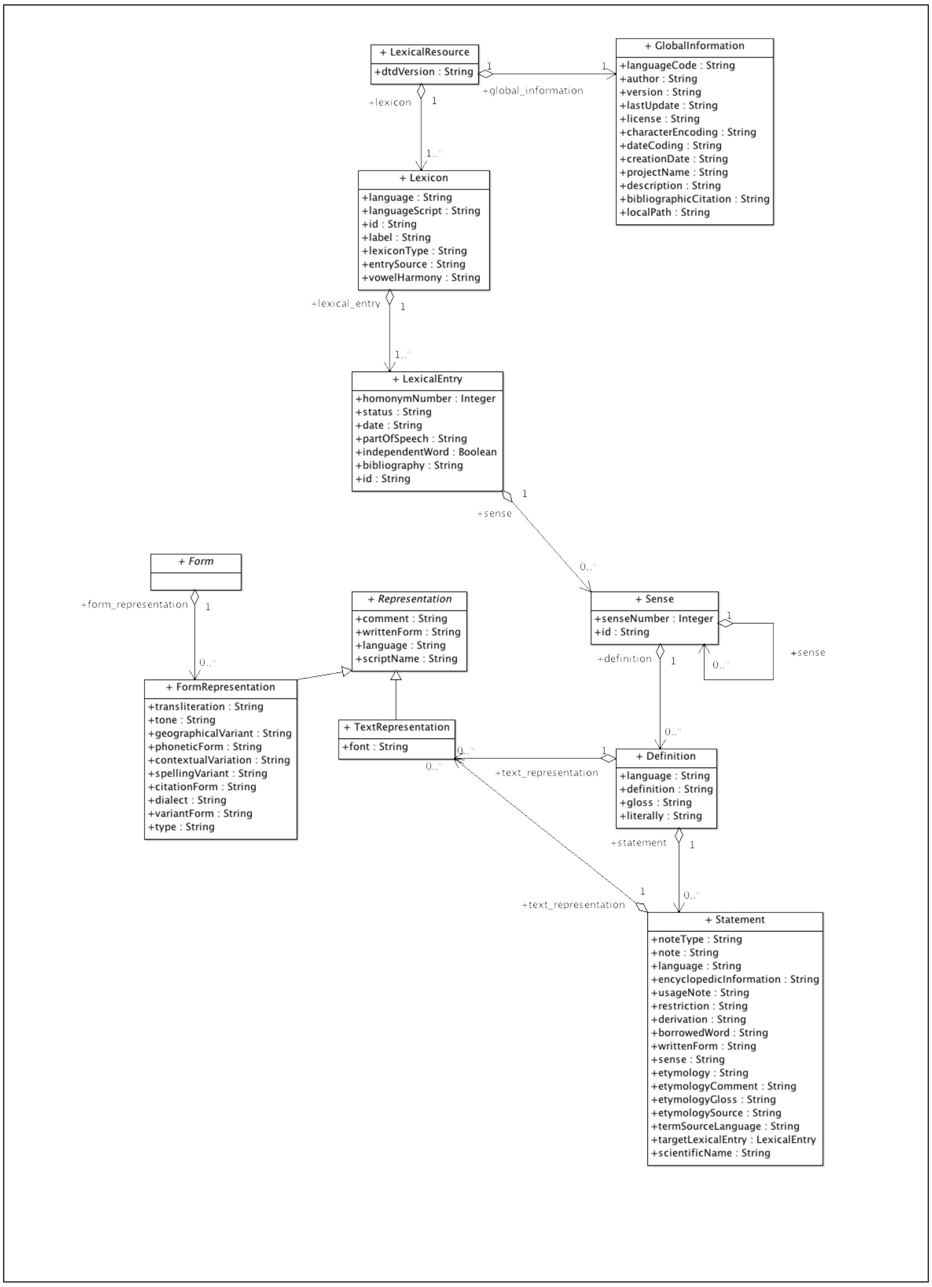

Figure 2: Core Package

Peripheral systems (extensions) are flexible, optional but powerful. Among the 8 proposed extensions, I have selected some that I think are relevant for our needs.

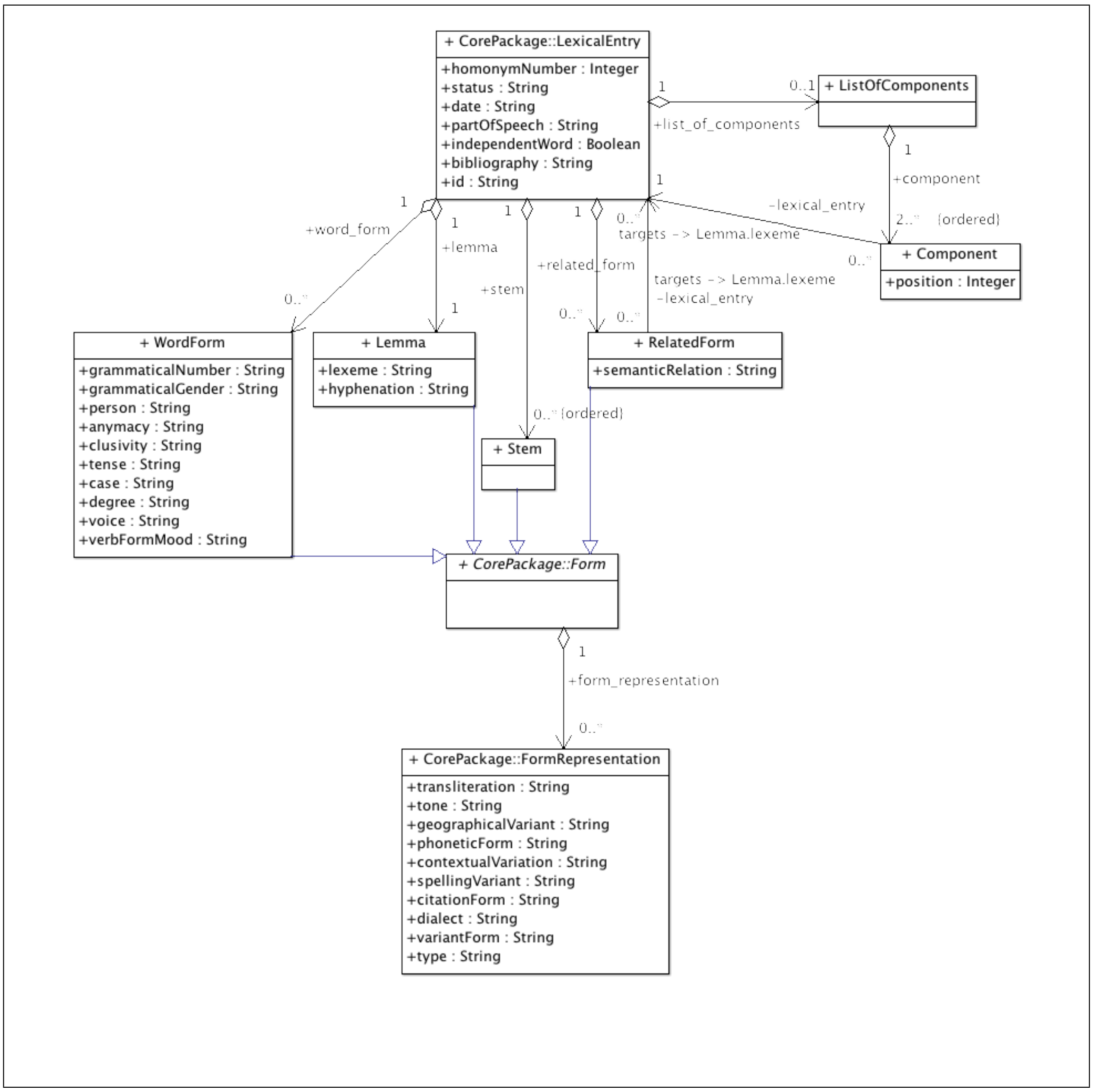

Figure 3: Morphology

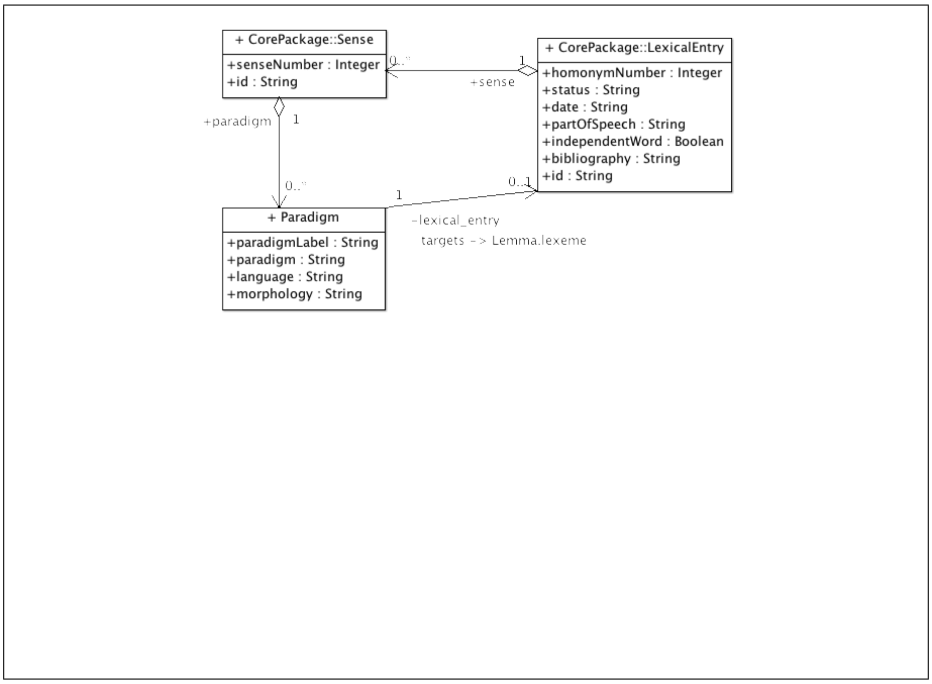

Figure 4: Morphosyntax

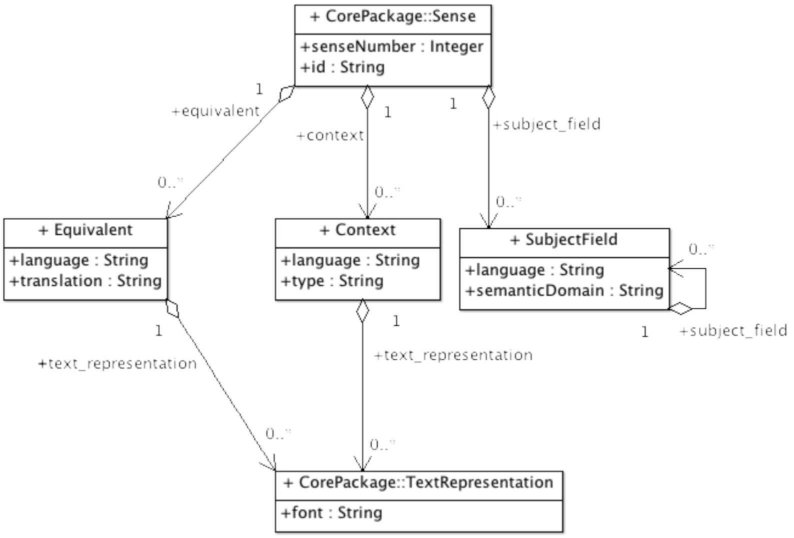

Figure 5: MRD (Machine Readable Dictionary)

In addition to existing extensions, we can create new ones. That is what I propose to do for audio ressources and speakers management.

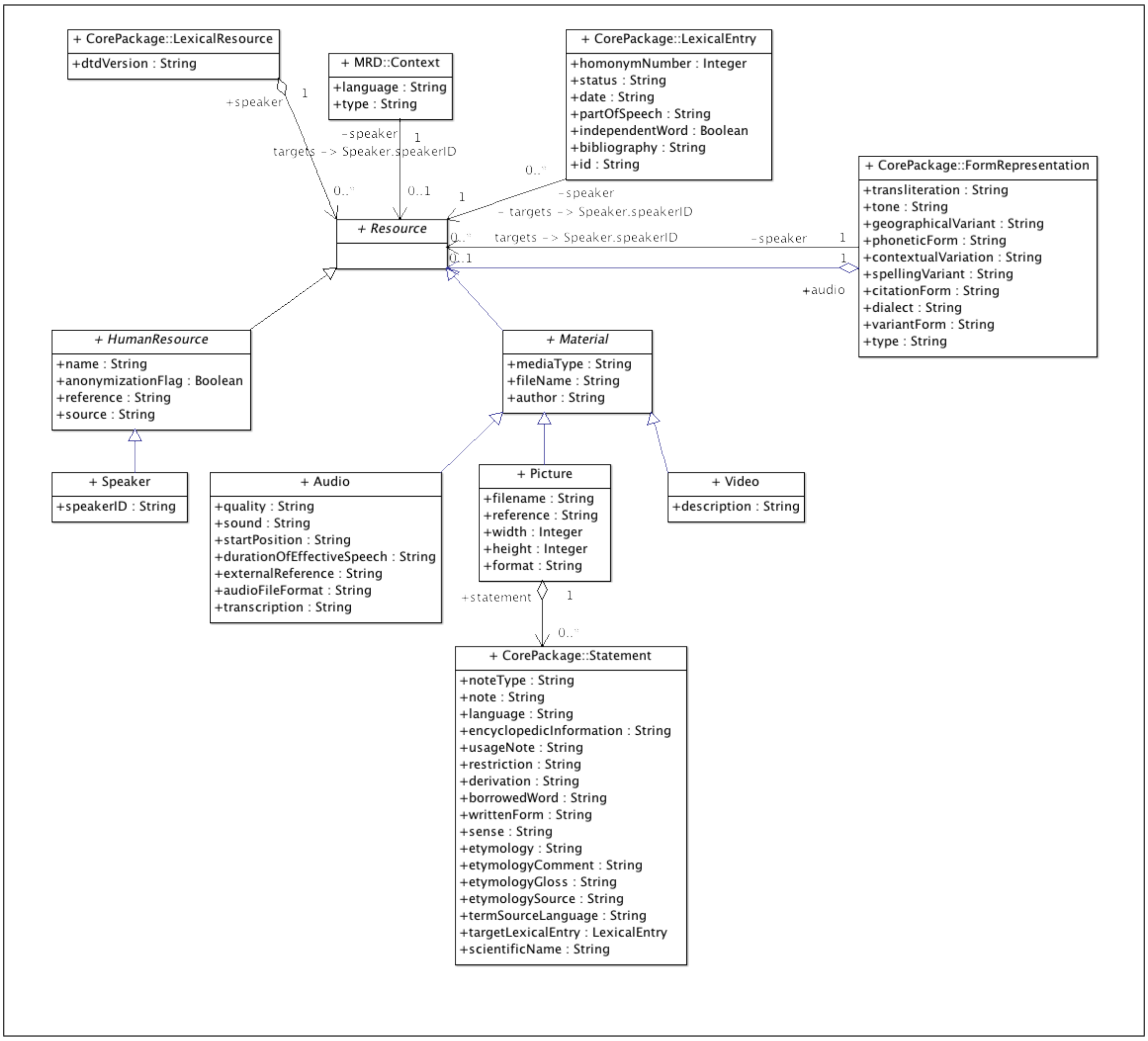

Figure 6: Resources

2 Classes and attributes

In this section, I will focus on what are a class and its attributes - in a simplified way, do not worry. Why? Because there is in fact a direct match between the used software architecture and the chosen XML LMF format.

2.1 Matching between UML and XML

A small example in order to have an overview: let us take the *Statement* class of the *Core Package* (at the bottom right of the figure). This class is composed of many attributes, including the 2 following ones:

- *borrowed word*
- *written form*

By following LMF recommendations, if we wish to represent for instance a borrowing from English of the word *cool* in French, we obtain following XML lines:

```
<Statement>
        <feat att=''borrowed word'' val=''eng''/>
        <feat att=''written form'' val=''cool''/>
</Statement>
```

Several comments about this example:

- In LMF, class attributes are structured as pair of attributes of specific tag *feat*.
  - The name of the attribute is indeed the value of the attribute *att* of the tag *feat* ;
  - The value given to this attribute is the value of the attribute *val* of the tag *feat*.
- In this example, it should be noted that according to LMF (and by the way also MDF), the borrowing language must be filled in the attribute *borrowed word*, while the borrowing word itself is filled in the attribute *written form*.

3 For novices: what is a class?

A class is an abstract entity that represents an object, for example a car, and that consists of some attributes, for example the brand or the color of the car. A class also has methods, that are functions that it implements: for the car, it would be for example *start*, *accelerate*, etc. Whereas attributes are generally materialized by common names, methods are named by action verbs.

On the other hand, a class can inherit from another class, that is, by simplifying, that it inherits from attributes and methods from its mother class. This heritage is represented on preceeding UML schematics by a full arrow. For instance, we could imagine a vehicle class, from which would inherit car, motorcycle, and so on, classes. They would all have common attributes (number of wheels, of doors, brand, color of the vehicle, etc.) that would then be attributes of the vehicle class, and specific attributes as for example the crutch for a motorcycle or a bike.

A class can have an aggregation or a compisition relation with another class, i.e. it is part of it. If we take again the basic example of the car and if we create a wheel class, we could say that the car is composed of, among other things, 4 wheels. This relation is represented by a lozenge in UML.

Another realtion used in UML schematics of the preceeding section is a simple arrow, which means that a class references another class. For instance, a car and its owner are two distinct entities that exist independantly from each other. However, a link exists between these two entities, represented by an association.

At last, in UML, abstract classes are written in italics.

3.1 Classes and attributes defined in LMF

For each package described in the previous section, classes and relations between classes are defined and not alterable (note that some existing projects deviate from the standard by proposing enhancements). However, we are (more or less) free to define attributes

that we want for each class. But each attribute must be referenced in the DCR (*Data Category Registry*). We can use existing elements, or propose new ones if appropriate. It is an open database, available on the website http://www.isocat.org.
A difficulty that I encountered with this database is that there are a lot of redundancies and duplicates: lots of quite identical terms are defined 2 or 3 times. In this case, which one to choose? According to which criteria? I have tried to focus on the definition that is closest to the need, and at almost similar definition, I have focused on terms issued from MDF, or created by Gil Francopoulo (author of the LMF book). However, rather than follow the MDF principles about markers associated specifically to vernacular, regional and national languages, I have chosen to let more freedom by defining a general attribute associated with a language attribute (example : definition in the 'xxx' language rather than 'dn' that forces a definition in a national predefined language). Moreover, this solution avoids to define for instance 'df' for the French language.

In the table below, I have listed attributes of each class, but not methods, because it would weigh down specifications without bringing relevant informations. I have also noted MDF markers which the attributes refer if any. As for concerned LMF extension, it is in the column *LMF package*.

Table 1: LMF classes and their attributes

LMF package	**Class name**	**Attribute**	**Attribute type or example value**	**DCR PID and type**	**MDF marker**	**Comment**
	Lexical Resource (singleton)	dtd version	“16”	-	-	LMF DTD is an XML attribute
		global information	Global Information	N/A	N/A	
		lexicon	Lexicon	N/A	N/A	
		resource	Speaker	N/A	N/A	
	Global Information (no subclass)	language code	“ISO-639-3”	2008 open	-	
		date coding	“ISO-8601”	2090 open	-	
		creation date	“2001-03-24”	2510 open	-	
		last update	“2014-07-21”	2526 open	-	
		author	“Alexis Michaud, MICA & Guillaume Jacques, CRLAO”	6130 open	-	
		version	“0.1”	2547 open	-	
		license	“GPL”	2457 open	-	
		project name	“ANR HimalCo”	2536 open	-	
		description	“everything you want to tell about this resource”	2520 open	-	
		bibliographic citation	“Online dictionaries, CNRS, 2014”	6137 open	-	

Core Package

Table 1: (continued)

LMF package	Class name	Attribute	Attribute type or example value	DCR PID and type	MDF marker	Comment
	Lexicon (no subclass)	character encoding	“UTF-8”	2564 open	-	
		id	“na?”	1845 open	-	identifier is an XML attribute (not necessarily unique)
		label	“Na online dictionary”	1857 open	-	
		language	“fra”, “eng”	2482 constrained	-	ISO 639 ; vernacular language
		language script	“latn”	2485 open	-	ISO 15924
		lexicon type	“bilingual dictionary na - eng”	2487 open	-	
		entry source	“na_dictionary.txt”	207 open	-	
		vowel harmony		no existing DC	-	
		lexical entry	Lexical Entry	N/A	N/A	-
	Lexical Entry (no subclass)	id	“toto_1”	6196 open	lx <id>, se <id>	unique identifier or key form is an XML attribute
		part of speech (English)	“verb”	3748 closed (1)	ps	grammatical category

Table 1: (continued)

LMF package	Class name	Attribute	Attribute type or example value	DCR PID and type	MDF marker	Comment
		lemmatized form	Lemma	N/A	N/A	
		date	“2014-06-15”	3694 open	dt	
		status	“no print”, “done”, “check”	3760 open	st	
		homonym number	“1”	3714 open	hm	“0” if no homonym
		bibliography	“212”	3687 open	bb	
		independent word	yes, no	5285 closed		
		resource	Resource	N/A	N/A	Speaker, Audio, Picture, Video
		form	Form Representation	N/A	N/A	
		sense	Sense	N/A	N/A	
		word form	Word Form	N/A	N/A	
		related form	Related Form	N/A	N/A	
		stem	Stem	N/A	N/A	
		list of components	List Of Components	N/A	N/A	
		borrowed word	Borrowed Word	N/A	N/A	
	Form (abstract class)	variant form(s)	“woman”, “women”	3768 open	va, pdl <stem>	written or spoken

Table 1: (continued)

LMF package	Class name	Attribute	Attribute type or example value	DCR PID and type	MDF marker	Comment
		type	(2)	1971 open		variant type : spelling, pronunciation, archaic, etc.
		form representation	Form Representation	N/A	N/A	
	Form Representation	tone		517 open	np <tone>	
		geographical variant		1851 open	va	
		phonetic form (vernacular)		3745 open	ph	
		contextual variation		1977 open	lc	
		spelling variant		5612 open	a	
		citation form (vernacular)		3716 open	lc	
		dialect	"North German"	2466 open	ve	
		language	"fra", "eng"	2482 constrained	-	ISO 639 ; language used for variant comment
		transliteration	"readable characters"	1848 open	ph	
		script name	"Latin"	3809 open	-	script used for romanization

Table 1: (continued)

LMF package	**Class name**	**Attribute**	**Attribute type or example value**	**DCR PID and type**	**MDF marker**	**Comment**
		resource	Resource	N/A	N/A	Speaker, Video, Picture
		sound	Resource	N/A	N/A	Audio
	Representation (abstract class)	written form	"..."	1836 open	xv, xe, xn, xr, xf	example
		language	"fra", "eng"	2482 constrained	-	ISO 639 ; language used for variant comment
		comment	"..."	1846 open	ve, vn, vr, vf, xc	explanation
	Text Representation	font	font family / font weight / font size	1650 closed		'font-style', 'font-variant', 'line-height'
	Sense	id	"toto_1_1"	1845 open	-	identifier or key form is an XML attribute (not necessarily unique)
		sense number	"1"	3758 open	sn	
		sense	Sense	N/A	N/A	
		definition	Definition	N/A	N/A	
		etymology	Etymology	N/A	N/A	
		paradigm	Paradigm	N/A	N/A	
		equivalent	Equivalent	N/A	N/A	
		context	Context	N/A	N/A	

Table 1: (continued)

LMF package	Class name	Attribute	Attribute type or example value	DCR PID and type	MDF marker	Comment
		subject field	Subject Field	N/A	N/A	
	Definition	definition	“This is the lexeme definition”	1972 open	dv, de, dn, dr, df	
		gloss	“GLOSS”	244 open	gv, ge, gn, gr, gf	
		language	“fra”, “eng”	2482 constrained	-	ISO 639 ; language used for definition and gloss
		literally	’au pied de la lettre’	3721 open	lt	
		text representation	Text Representation	N/A	N/A	
		statement	Statement	N/A	N/A	
	Statement	note type	(3)	6178 open	nt <type>, np <type>, ng <type>	
		note		382 open	na, nd, ng, np, nq, ns, nt	
		language	“fra”, “eng”	2482 constrained	nt <lang>	ISO 639
		encyclopedic information	“...”	3828 open	ee, en, er, ev	
		usage note	“...”	526 open	uv, ue, un, ur	text
		restriction	“...”	1956 open	oe, on, or, ov	

Table 1: (continued)

LMF package	**Class name**	**Attribute**	**Attribute type or example value**	**DCR PID and type**	**MDF marker**	**Comment**
		derivation	"..."	188 open	-	
		borrowed word (English)	"Chinese"	3688 open	bw	source language
		written form	"..."	1836 open	bw	loan word
		sense	"..."	464 open	-	sense in borrowed language
		etymology	"aspirin: from acetyl + spiraeic acid (old name for salicylic acid)"	221 open	et	
		etymology comment (English)		3696 open	ec	
		target lexical entry	Lexical Entry		cf <type="et">	
		term source language	"fra", "eng"	3639 open	-	language
		etymology gloss		3698 open	eg	
		etymology source		3701	es	
		scientific name	"Canis lupus familiaris"	3754 open	sc	
		text representation	Text Representation	N/A	N/A	

Table 1: (continued)

LMF package	**Class name**	**Attribute**	**Attribute type or example value**	**DCR PID and type**	**MDF marker**	**Comment**
	List Of Components	component	Component	N/A	N/A	
	Component	position	"2"	2183 open	-	
		target lexical entry	Lexical Entry	N/A	N/A	
Morphology	Word Form	grammatical number	collective, dual, paucal, plural, quadrial, singular, trial	1298 closed		
		grammatical gender	common gender, feminine, masculine, neuter	1297 closed		
		person	first person, second person, third person	1328 closed		
		anymacy	animate, inanimate, other anymacy	1902 closed		
		clusivity	inclusive, exclusive	3031 closed		
		tense	future, imperfect, past, present	1286 closed		

Table 1: (continued)

LMF package	**Class name**	**Attribute**	**Attribute type or example value**	**DCR PID and type**	**MDF marker**	**Comment**
		voice	active voice, middle voice, passive voice	1413 closed		
		verb form mood	(4)	1427 closed		
		case	“accusative case”	1840 closed		
		degree	comparative degree, positive degree, superlative degree	2779 closed		
	Lemma	lexeme	“toto”	3723 open	lx	
		hyphenation	“pho-ne-ti-cian”	264 open	-	syllables separated by ’-’
	Stem			N/A	N/A	
	Related Form	semantic relation	(5)	6331 open	sy, an, cf <et>, cf <hm>, se, mn, lf, ev, ee, en, er	
		cross reference	Lexical Entry	164 open	cf, mn	also used for main entry cross-reference
Morpho-syntax	Paradigm	paradigm label (English)	(6)	3741 open	pdl	
		language	“fra”, “eng”	2482 constrained	-	ISO 639
		paradigm		3736 open	pd	

Table 1: (continued)

LMF package	Class name	Attribute	Attribute type or example value	DCR PID and type	MDF marker	Comment
		morphology (vernacular)		3738 open	mr	
		target lexical entry	Lexical Entry	N/A	N/A	in case of classifier
MRD	Context	language	"fra", "eng"	2482 constrained	-	ISO 639
		type	"proverb", "locution", "example", "combination"	1971 open	PHONO	
		resource	Audio	N/A	N/A	
		text representation	Text Representation	N/A	N/A	
	Subject Field	language	"fra", "eng"	2482 constrained	sd <lang>	ISO 639
		semantic domain	"arbre"	3755 open	sd, is, th	see appendix C of the MDF guide
		subject field	Subject Field	N/A	N/A	hyponym / hypernym
	Equivalent	language	"fra", "eng"	2482 constrained	-	ISO 639
		translation		6037 open	re, rn, rr, rf	reversal
		text representation	Text Representation	N/A	N/A	
	Resource (abstract class)					

Resources

Table 1: (continued)

LMF package	Class name	Attribute	Attribute type or example value	DCR PID and type	MDF marker	Comment
	Material (abstract class)	media type	unspecified, unknown, audio, video, document, text, image, drawing	2570 closed		
		file name		5435 open	sf, sfx	
		author	"Guillaume Jacques, CRLAO"	6130 open	-	
	Audio	quality	very low, low, normal, good, very good (high)	2574	sf, sfx <quality>	
		sound		2250 open	-	
		transcription		1849 open	-	
		start position	"00:05:00"	3896 open	-	
		duration of effective speech	"00:05:00", "3"	2691 open	-	
		external reference		1975 open	sf, sfx <numbering>	
		audio file format	"MP3", "Vorbis", "WAV", "AU", "uLaw"	2689 open	sf, sfx	

Table 1: (continued)

LMF package	**Class name**	**Attribute**	**Attribute type or example value**	**DCR PID and type**	**MDF marker**	**Comment**
	Video	description	“everything you want to tell about this video”	2520 open	-	
	Picture	size		2580open	pc	
		size unit		2583 open	pc	
		statement	Statement		N/A	
	Human Resource (abstract class)	name		6122 open	-	
		source		3759 open	so	
		reference		3751 open	rf	
		anonymization flag	false, true, unknown, unspecified	2548 closed	so <print>	
	Speaker	speaker id	“SpID-1”	3597 open		

(1) part of speech:

- adjective 1230
- adposition 1231
- adverb 1232
- affirmative particle 1918
- affix 1234
- article 1892
- auxiliary 1244
- bitransitive verb 1275
- classifier 2345
- comparative particle 1922
- conditional particle 2230

- conjunction 1260
- coordinating conjunction 1262
- declarative punctuation 2086
- demonstrative determiner 1269
- determiner 1272
- existential pronoun 3012
- ideophone 4192
- impersonal verb 1306
- indefinite determiner 1307
- interjection 1318
- interrogative determiner 1320
- interrogative particle 1921
- intransitive verb 1322
- modal 1329
- negation 2313
- negative particle 1894
- noun 1333
- numeral 1334
- particle 3372
- participle adjective 1598
- possessive pronoun 1359
- possessive relative pronoun 3005
- postposition 1360
- preposition 1366
- presentative pronoun 3015
- pronoun 1370
- proper noun 1371
- reciprocal pronoun 1924

- reflexive determiner 1377
- reflexive verb 5592
- relative determiner 1379
- time noun 3855
- transitive verb 1405
- verb 1424

Values not found in the DCS (*Data Category Selection*):

- onomatope
- function word
- stative intransitive verb
- linker

(2) type:

- unspecified 1908 (simple)
- orthography 2971 (simple)
- phonetics 2641 (simple)
- archaic form 504 (simple)

(3) note type:

- “comparison”
- “history”
- “semantics”
- “tone”
- “derivation”
- “case”
- “subord”
- “usage”
- “comment”
- “legend”
- “restriction”

- “encyclopedic”
- “anthropology”
- “discourse”
- “grammar”
- “phonology”
- “question”
- “sociolinguistics”
- “general”

(4) verb form mood:

- gerundive
- imperative
- indicative
- infinitive
- participle
- subjunctive
- conditional
- relative mood
- prohibitive mood
- debitive mood

(5) semantic relation:

- synonym
- antonym
- homonym
- etymology
- subentry
- main entry
- simple link
- derived form

- root
- stem
- collocation 340 (simple) (classifier)

(6) paradigm label:

- lexicalized affix (la)
- conjugation class (cc)
- thème du passé (past)
- comitatif (comit)
- construction (constr)
- directional (dir)
- irregularity (ir)

3.2 Remarks and limitations

1. Toolbox subentries are coded as *Lexical Entry* whose main entry has links with others.

2. With the proposed model, we can not establish a reference ('cf') from a sense to another. It is at the entry level that we can reference another lexical entry as a synonym for instance. Is there a need to do it at the 'sn' (*sense number*) level? It would add complexity to the model, but it is a possible enhancement. We can also simplify the model if you think that some attributes or even some classes are not necessary.

3. Case of complex predicates VV or NV: let us take the example of complex predicate NV. According to the LMF model, we would have 3 lexical entries:

   - V with the attribute *independent word* = *no* ;
   - N with the attribute *independent word* = *no* ;
   - NV with the attribut *independent word* = *yes*, having as list of components (*List Of Components*) a link to the 2 lexical entries defined above.

4 Examples

4.1 Na

Table 2: Na dictionary: matching between MDF and LMF

MDF	LMF
lx, se	Lemma lexeme
lx, se <id>	Lexical Entry id
sf	Material file name
sf <nb>	Audio external reference
hm	Lexical Entry homonym number
lc	Form Representation contextual variation
ph	Form Representation romanization
bw	Borrowed Word borrowed word / written form
et	Etymology etymology
ec	Etymology etymology comment
ec <lang>	Etymology language
ps	Lexical Entry part of speech
sn	Sense sense number
cf	Related Form cross reference
cf <type>	Related Form semantic relation
sd	Subject Field semantic domain
sd <lang>	Subject Field language
nt	Statement note
nt <lang>	Statement language
nt <type>	Statement note type
np	Statement note
np <type>	Statement note type
nd	Statement note
nd <arch>, ue archaic	Form type = archaic form
so	Human Resource source
so <print>	Human Resource anonymization flag
va	Form Representation variant form
va <speaker>	Form Representation resource
vf	Representation comment with Representation language = “fra”
vf <type>	Representation comment
pdl	Paradigm paradigm label
pdv	Paradigm paradigm with Paradigm language = “na”
pdf	Paradigm paradigm with Paradigm language = “fra”
de	Definition definition with Definition language = “eng”
ge	Definition gloss with Definition language = “eng”
dn	Definition definition with Definition language = “chn”

Table 2: (continued)

gn	Definition gloss with Definition language = “chn”
gr	Definition gloss with Definition language = “...”
df	Definition definition with Definition language = “fra”
gf	Definition gloss with Definition language = “fra”
xv	Representation written form with Representation language = “na?”
xe	Representation written form with Representation language = “eng”
xn	Representation written form with Representation language = “chn”
xf	Representation written form with Representation language = “fra”
rf	Context resource
xc	Representation comment
dt	Lexical Entry date

```
\lx æ˩˥
\sf <nb=”B”> 1789
\sf <nb=”2011”> 2642
\hm
\ph
\bw
\et
\ec <lang=”fr”>
\ps n
\sn
\cf
\cf <type=”hm”>
\sd <lang=”fr”> animal
\sd <lang=”eng”> animal
\nt <lang=”pumi” type=”comp” print=”n”>
\nt <type=”hist” print=”n”>
\nt <type=”hist” print=”n”>
\nt <type=”sem”>
\np LM confirmé type ”porc”
\np <type=”tone”> LM
\nd
\so <print=”n”> F4
\va <speaker=”F4”>
\vf <type=”tone”>
\va <speaker=”F5”> ID.
\vf <type=”tone”>
\va <speaker=”M18”>
\va <speaker=”M21”> ID.
\va <speaker=”M23”>
\pdl classifier
\pdv mi˩
\pdf
\de chicken
\ge chicken
\dn 鸡
\gn 鸡
\gr
\df poulet, poule
\gf poulet
\xv æ˩ dzɯ˧-ze˩
\xe ...has eaten (a/some) chicken
\xn 吃了鸡
\beginlstlisting
\xf ...a mangé (un/du) poulet
\xc PHONO
\xv æ˩ hwæ˧-ze˧
```

\xe ...has bought (a) chicken
\xn 买了鸡
\xf ...a acheté (un/du) poulet
\xc PHONO
\xv *æ˩˥, / kʰv˧, / bo˩˥, / hwɤ˥, / ɟi˧, / la˧, / tʰo˧li˧, / mv˧gv˧, / bv˧ʐv˧, / ʐwæ˧, / jo˧, / ʑi˩˥*
\xe the twelve years of the duodenary cycle
\xn 十二个生肖
\xf les douze signes astrologiques
\rf
\xv
\xf
\rf
\xv
\xf
\xc
\dt 15/Jun/2014

Listing 1: Na example

```
<?xml version="1.0" encoding="UTF-8"?>

<!DOCTYPE LexicalResource SYSTEM "DTD_LMF_REV_16.dtd">
<LexicalResource dtdVersion="16">
    <GlobalInformation>
        <feat att="languageCode" dcr:datcat="http://www.isocat.org/datcat/DC-2008" val="ISO-639-3"/>
    </GlobalInformation>
    <Speaker speakerId="F4" dcr:datcat="http://www.isocat.org/datcat/DC-3597"/>
    <Speaker speakerId="F5"/>
    <Speaker speakerId="M21"/>
    <Lexicon>
        <LexicalEntry id="æ_1" dcr:datcat="http://www.isocat.org/datcat/DC-6196">
            <feat att="partOfSpeech" dcr:datcat="http://www.isocat.org/datcat/DC-3748" val="noun" dcr:datcat="http://www.isocat.org/datcat/DC-1333"/>
            <feat att="date" dcr:datcat="http://www.isocat.org/datcat/DC-3694" val="2014-06-15"/>
            <Lemma targets="F4">
                <feat att="lexeme" dcr:datcat="http://www.isocat.org/datcat/DC-3723" val=" æ"/>
            </Lemma>
            <Audio>
                <feat att="externalReference" dcr:datcat="http://www.isocat.org/datcat/DC-1975" val="B:1789"/>
            </Audio>
            <Audio>
                <feat att="externalReference" val="2011:2642"/>
            </Audio>
            <FormRepresentation targets="F5">
                <feat att="variantForm" dcr:datcat="http://www.isocat.org/datcat/DC-3768" val=" æ"/>
            </FormRepresentation>
            <FormRepresentation targets="M21">
                <feat att="variantForm" val=" æ"/>
            </FormRepresentation>
            <Sense id="æ_1-0" dcr:datcat="http://www.isocat.org/datcat/DC-1845">
                <SubjectField>
                    <feat att="language" dcr:datcat="http://www.isocat.org/datcat/DC-2482" val="fra"/>
                    <feat att="semanticDomain" dcr:datcat="http://www.isocat.org/datcat/DC-3755" val="animal"/>
                </SubjectField>
                <SubjectField>
                    <feat att="language" val="eng"/>
                    <feat att="semanticDomain" val="animal"/>
                </SubjectField>
                <Definition>
                    <Statement>
                        <feat att="noteType" dcr:datcat="http://www.isocat.org/datcat/DC-6178" val="phonology"/>
                <feat att="language" dcr:datcat="http://www.
                    isocat.org/datcat/DC-2482" val="fra"/>
                <feat att="note" dcr:datcat="http://www.isocat.
                    org/datcat/DC-382" val="LM confirmé type "
                    porc""/>
            </Statement>
            <Statement>
                <feat att="noteType" val="tone"/>
                <feat att="note" val="LM"/>
            </Statement>
        </Definition>
        <Definition>
            <feat att="language" dcr:datcat="http://www.isocat.
                org/datcat/DC-2482" val="eng"/>
            <feat att="definition" dcr:datcat="http://www.isocat
                .org/datcat/DC-1972" val="chicken"/>
            <feat att="gloss" dcr:datcat="http://www.isocat.org/
                datcat/DC-244" val="chicken"/>
        </Definition>
        <Definition>
            <feat att="language" val="chn"/>
            <feat att="definition" val=" "/>
            <feat att="gloss" val=" "/>
        </Definition>
        <Definition>
            <feat att="language" val="fra"/>
            <feat att="definition" val="poulet, poule"/>
            <feat att="gloss" val="poulet"/>
        </Definition>
        <Paradigm targets="mi1">
            <feat att="paradigmLabel" dcr:datcat="http://www.
                isocat.org/datcat/DC-3741" val="classifier"/>
            <feat att="paradigm" dcr:datcat="http://www.isocat.
                org/datcat/DC-3736" val=" mi"/>
        </Paradigm>
        <Context>
            <TextRepresentation>
                <feat att="language" dcr:datcat="http://www.
                    isocat.org/datcat/DC-2482" val="na?"/>
                <feat att="writtenForm" dcr:datcat="http://www.
                    isocat.org/datcat/DC-1836" val=" æ  dz- ze"/>
            </TextRepresentation>
            <TextRepresentation>
                <feat att="language" val="eng"/>
                <feat att="writtenForm" val="...has eaten (a/
                    some) chicken"/>
            </TextRepresentation>
            <TextRepresentation>
                <feat att="language" val="chn"/>
                <feat att="writtenForm" val="  "/>
            </TextRepresentation>
            <TextRepresentation>
                <feat att="language" val="fra"/>
                <feat att="writtenForm" val="...a mangé (un/du)
                    poulet"/>
```

```
                        <feat att="comment" dcr:datcat="http://www.
                            isocat.org/datcat/DC-1846" val="PHONO"/>
                    </TextRepresentation>
                </Context>
                <Context>
                    <TextRepresentation>
                        <feat att="language" val="na?"/>
                        <feat att="writtenForm" val=" æ hwæ- ze"/>
                    </TextRepresentation>
                    <TextRepresentation>
                        <feat att="language" val="eng"/>
                        <feat att="writtenForm" val="...has bought (a)
                            chicken"/>
                    </TextRepresentation>
                    <TextRepresentation>
                        <feat att="language" val="chn"/>
                        <feat att="writtenForm" val=" "/>
                    </TextRepresentation>
                    <TextRepresentation>
                        <feat att="language" val="fra"/>
                        <feat att="writtenForm" val="...a acheté (un/du)
                            poulet"/>
                        <feat att="comment" val="PHONO"/>
                    </TextRepresentation>
                </Context>
                <Context>
                    <TextRepresentation>
                        <feat att="language" val="na?"/>
                        <feat att="writtenForm" val=" æ, | h kv, |  bo,
                            |  hw, |  i, |  l, | h toli, |  mvgv, |  bvv
                            , |  wæ, | jo, |  i"/>
                    </TextRepresentation>
                    <TextRepresentation>
                        <feat att="language" val="eng"/>
                        <feat att="writtenForm" val="the twelve years of
                            the duodenary cycle"/>
                    </TextRepresentation>
                    <TextRepresentation>
                        <feat att="language" val="chn"/>
                        <feat att="writtenForm" val="  "/>
                    </TextRepresentation>
                    <TextRepresentation>
                        <feat att="language" val="fra"/>
                        <feat att="writtenForm" val="les douze signes
                            astrologiques"/>
                    </TextRepresentation>
                </Context>
            </Sense>
        </LexicalEntry>
        <LexicalEntry id="mi_1">
            <Lemma>
                <feat att="lexeme" val=" mi"/>
            </Lemma>
        </LexicalEntry>
    </Lexicon>
```

```
</LexicalResource>
```

Note that attributes *dcr:datcat* can be defined in the DTD in order to lighten the XML document.

4.2 Japhug

Table 3: Japhug dictionary: matching between MDF and LMF

MDF	**LMF**
lx, se	Lemma lexeme
lx, se <id>	Lexical Entry id
sf (wav)	Material file name
sf <qual> (wav or wav8)	Audio quality
bb or hbf	Lexical Entry bibliography
hm	Lexical Entry homonym number
dt	Lexical Entry date
dt <print>	-
ph	Form Representation romanization
ph <print>	-
ph <lang>	Form Representation script name
bw	Borrowed Word borrowed word / written form
et	Etymology etymology
ec	Etymology etymology comment
ec <lang>	Etymology language
ps	Lexical Entry part of speech
sn	Sense sense number
sy	Related Form cross reference with Related Form semantic relation = synonym
an	Related Form cross reference with Related Form semantic relation = antonym
cf	Related Form cross reference
cf <type>	Related Form semantic relation
sd	Subject Field semantic domain
sd <lang>	Subject Field language
nt	Statement note
nt <print>	-
nt <lang>	Statement language
nt <code>	Text Representation font
nt <type>	Statement note type
np	Statement note
np <type>	Statement note type
ng	Statement note

Table 3: (continued)

ng <type>	Statement note type
nd	Statement note
nq	Statement note
nq <print>	-
mr or ms	Paradigm paradigm
mr or ms <lang>	Paradigm language
mr or ms <type>	Paradigm paradigm label
pd etc.	Word Form grammatical number / grammatical gender / person / anymacy / clusivity
pdl or comit or constr	Paradigm paradigm label
pdv	Paradigm paradigm with language = “jya”
pde	Paradigm paradigm with language = “eng”
pdf	Paradigm paradigm with language = “fra”
de	Definition definition with Definition language = “eng”
ge	Definition gloss with Definition language = “eng”
dn	Definition definition with Definition language = “chn”
gn	Definition gloss with Definition language = “chn”
dr	Definition definition with Definition language = “nep”
gr	Definition gloss with Definition language = “nep”
df	Definition definition with Definition language = “fra”
gf	Definition gloss with Definition language = “fra”
uv	Statement usage note with language = “jya”
ue	Statement usage note with language = “eng”
un	Statement usage note with language = “chn”
ur	Statement usage note with language = “nep”
ev	Statement encyclopedic information with language = “jya”
ee	Statement encyclopedic information with language = “eng”
en	Statement encyclopedic information with language = “chn”
er	Statement encyclopedic information with language = “nep”
xv	Representation written form with Representation language = “jya”
xe	Representation written form with Representation language = “eng”
xn	Representation written form with Representation language = “chn”
xr	Representation written form with Representation language = “...”
xf	Representation written form with Representation language = “fra”
xc	Representation comment
dt	Lexical Entry date

\lx *akarɯ*
\ps N
\ge origan
\gn 牛至
\hbf plante
\xv *akarɯ nɯ sɯjno kɯ-xtɕi ci ŋu, ɯ-ru kɯ-xtshɯ-xtshɯm kɯ-ɣɯrni ci ŋu, ʁnɯ-tɣa jamar ma mɤ-mbro, ɯ-jwaʁ kɯ-ɤrtɯm, kɯ-ɽɲɟi tsa ci ŋu, ɯ-di mnɤm, ɯ-mɯntoʁ kɯ-ɣɯrni ŋgɯ kɯ-wɣrum tsa ci ŋu, ɯ-zrɤm kɯ-xtɕɯ-xtɕi ma me, ɯʑo smɤn ɯ-ŋgɯ kɤ-lɤt ɲɯ-sna.*
\xn 牛至是一种小植物，茎非常细，呈红色，只有两乍高，有椭圆形的小叶花是红里透白 有香味，只有小小的根。可以放在药里。
\dt 03/Jul/2014

Listing 2: Japhug example

```
<?xml version=”1.0” encoding=”UTF-8”?>

<!DOCTYPE LexicalResource SYSTEM ”DTD_LMF_REV_16.dtd”>
<LexicalResource dtdVersion=”16”>
    <GlobalInformation>
        <feat att=”languageCode” val=”ISO-639-3”/>
    </GlobalInformation>
    <Lexicon>
        <LexicalEntry id=” akar_1”>
            <feat att=”partOfSpeech” val=”noun”/>
            <feat att=”bibliography” dcr:datcat=”http://www.isocat.org/
                datcat/DC-3687” val=”plante”/>
            <feat att=”date” val=”2014-07-03”/>
            <Lemma>
                <feat att=”lexeme” val=” akar”/>
            </Lemma>
            <Sense id=” akar_1-0”>
                <Definition>
                    <feat att=”language” val=”eng”/>
                    <feat att=”gloss” val=”origan”/>
                </Definition>
                <Definition>
                    <feat att=”language” val=”chn”/>
                    <feat att=”gloss” val=” ”/>
                </Definition>
                <Context>
                    <TextRepresentation>
                        <feat att=”language” val=”jya”/>
                        <feat att=”writtenForm” val=” akar n sjno k-
                            xti ci ŋu, -ru k- xtsh- xtshm k -rni ci ŋu,
                            n- ta jamar ma m-mbro, - jwa k - rtm, k- ri
                           tsa ci ŋu, -di mmm, - mnto k -rni ŋ g k-
                           wrum tsa ci ŋu, - zrm k- xt- xti ma me, o
                           smn ŋ- g k- lt -sna.”/>
                    </TextRepresentation>
                    <TextRepresentation>
                        <feat att=”language” val=”chn”/>
                        <feat att=”writtenForm” val=”
                              , , , , , , ,        ”/>
                    </TextRepresentation>
                </Context>
            </Sense>
        </LexicalEntry>
    </Lexicon>
</LexicalResource>
```

4.3 Mwotlap, Araki, Lo, Teanu

In dictionaries from Alexandre François, specific markers have been used. Here is a list and proposed equivalences in LMF.

Table 4: Mowtlap dictionary: matching between MDF and LMF

MDF	Purpose	LMF
wr	*word reference* to have several different ‘ps’ for the same ‘lx’ entry, not to be confused with sub-entries	several *Lexical Entry*
we	diverted for syntactic restriction: syntactic context ; grammatical notes that specify more precisely the sense in particular	equivalent: ‘ov’
wn	same thing in English	equivalent: ‘oe’
he	semantic label to qualify the type of semantic relation: metaphorically, figuratively, etc.	*Related Form semantic relation*: add “metaphor” and “figuratively”
hn	‘he’ in English	‘he’ only in English
ll	equivalent of ‘lt’ in English	Definition literally with language = “eng”
oe	note on an example	equivalent: ‘xc’
on	‘oe’ in English	Text Representation comment with language = “eng”
ur (regional = bislama)	subject or typical possessor ; for a given sense, which type of subject it is the predicate of	Statement usage note
se	can also indicate the prefixed form of the noun	Form variant form: add type = “prefix”
el	language of etymology	Statement term source language
dc	creation date	add *creation date* in *Lexical Entry*
la	prefixed form for an entry, as ‘se’ followed by ‘wr’	Form variant form: add type = “prefix”
lg	legend of the picture	Picture statement with note type = “legend”
ce	gloss of ‘cf’ in French	Statement etymology gloss
u	*underlined form* corresponding to ‘a’, destinated to the *parser*	Form Representation spelling variant
xm	hidden example	add a type “hidden example”

Table 4: (continued)

rm	reference of a hidden example	Context resource reference
xa	English version of a hidden example	Context text representation with language = “eng”
mr	morpho	Paradigm morphology
ue	label	configuration file
un	label in English	configuration file
tb	frame of list of words in French	Table written form with type = “word list” and language = “fra” (to add)
ta	equivalent of ‘tb’ in English	Table written form with type = “word list” and language = “eng” (to add)
tl	frame in prose	Table written form with type = “text” and language = “fra” (to add)
tn	English equivalent of ’tl’	Table written form with type = “text” and language = “eng” (to add)

Specific used syntax:

- “ax:” for a text in italics: to replace by “fi:”
- small angle brackets to indicate the syntactic object: *Statement usage note*

4.4 Tamang

It is the dictionary of Martine Mazaudon, written in Word and based on the LEXWARE format. Here is an exhaustive list of used markers and their equivalents in MDF or LMF.

Table 5: Tamang dictionary: matching between Word and MDF or LMF

Word	**Purpose**	**MDF or LMF**
hdr	header	Lexicon label
hw	headword	lx
...X	if several senses	sn
ton	from 0 to 5 ; noted x,x if hesitation	np
dff		df
dfe		de
dfn	nepali (national language)	dn
dfzoo	zoological definition	sc
dfbot	botanical definition	sc
nbbot	remarks on the botanic field	Definition statement
nag	nagari transliteration (local writing)	Form Representation transliteration with script name = “nagari”
phr	*phrase*: example of incomplete sentences	Context with type = “’incomplete’ (to add)
il	illustration: example	xv
ilnep		xn
gram		ng
rec	records	sf
xr	cross-reference	cf
nb	nota bene	nt
nbi	‘i’ for internal	nq
emp	borrowing language	bw
check	personal note	status
sem	semantic field	sd
enc	encyclopedic notes	ee
inf	informers	rf
cf		Related Form with semantic relation = “simple link”
syn		Related Form with semantic relation = “synonym”
anton		Related Form with semantic relation = “synonym”
etym		et
morph		Paradigm morphology
var		va

Table 5: (continued)

niv	language level?	to add?
ps		ps
so		so
cons	?	
comp	?	
conj	?	
stedt	?	

Specific used syntax:

- *old = don't print*
- mm = Martine Mazaudon

4.5 Limbu

It is the dictionary of Boyd Michailovsky, previously converted from LEXWARE to XML, which structure is described below.

Listing 3: Limbu XML format

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE DICO
   SYSTEM "dicoLimbu.dtd">

<DICO>
   <entry id“=”xxx_1>
      <form>
         <pron type="headword|var|pastem|prstem|pa|pask|fem|poss|root|
             neg|allom" valid“=”doubt>xxx</pron>
         <note type=''ph|rem|comm|gram|stem'' valid=''doubt''>...</note>
      </form>
      <gramGrp>
         <pos valid“=”doubt class="v|vprefix|vsuffix|preverb|"misc…></
             pos>
         <note/>
      </gramGrp>
      <sense>
         <def type="binom|"par xml:lang“…”= valid“=”doubt>…</def>
         <invertkey>…</invertkey>
         <sem>…</sem>
         <xptr target“…”= valid“=”doubt>...</xptr>
         <eg type“=”hidden>
            <q>…</q>
            <xptr>…</xptr>
            <link xmlns:xlink“…”= xlink:type“…”= xlink:actuate“…”=
                xlink:show“…”= xlink:href="…"…></link>
            <trans>
               <tr xml:lang=""...>…</tr>
            </trans>
         </eg>
         <note/>
      </sense>
      <xr type="herbier>
         <ptr type“…”= target“=”yyy_2 valid“…”=>yyy</ptr>
         <xptr/>
         <lexx/>
         <ref valid“=”doubt/>
         <wordFamily type“…”= family“…”= valid“=”doubt/>
         <note/>
      </xr>
      <usg>
         <dial>…</dial>
         <note/>
      </usg>
      <hom n="3">
         <form/>
         <gramGrp/>
         <sense/>
         <xr/>
```

```
            <usg/>
        </hom>
    </entry>
</DICO>
```

Specific syntax:

Listing 4: Limbu syntax

```
<foreign xml:lang=""lif …></foreign>
<family name“…”…=></family>
```

Table 6: Limbu dictionary: matching between XML and LMF

TEI-based XML	Purpose	LMF
entry	main entry	Lexical Entry
form	spoken and morphophonemic forms ; orthography if available	Lemma lexeme, Form Representation, Word Form
pron	phonological transcription	Form Representation phonetic form
usg	usage: dialect, level of language, etc.	Statement usage note
dial	dialect	Form Representation dialect
gramGrp	grammatical information (part of speech, etc.)	Word Form
pos	part of speech	Lexical Entry part of speech
sense	definitions, keys for inverting the dictionary, example sentences, encyclopedic information, certain semantic categories...	Sense
def	definition	Definition
invertedkey	the key under which the definition appears in the English index	Equivalent translation
sem	semantic class, a limited inventory for certain domains only	Subject Field semantic domain
eg	illustrative example	Context
q	citation	Context text representation
trans / tr	translation	Context text representation
xr	internal and external references	Related Form

Table 6: (continued)

ptr	cross-reference to another entry in the dictionary	Related Form cross reference
xptr	reference to an external item, in this case a printed document	Lexical Entry bibliography
wordFamily	a word-family of roots to which the entry belongs	Stem