Projet HimalCo : LMF et dictionnaires

Céline Buret

July 24, 2015

1 Qu'est-ce que LMF ?

LMF est une norme ISO (*International Standard Organisation*) du comité technique 37 et sous-comité 4 : ISO-TC37/SC4 24613.
Cette norme est adaptée aux dictionnaires généraux et spécialisés, monolingues et multilingues. Elle décrit une structure générique formelle indépendante des supports de publication : à partir d'une source lexicographique unique bien structurée, on peut obtenir une forme imprimée et une forme électronique des données.
LMF suit une approche lexicographique centrée sur le lemme. C'est un modèle à deux couches : morphologique et sémantique.

Le modèle LMF est divisé en deux grandes parties : ce qu'ils appellent le *core package*, un squelette simple, rigide, et obligatoire, qui est le cœur du modèle ; et des extensions.

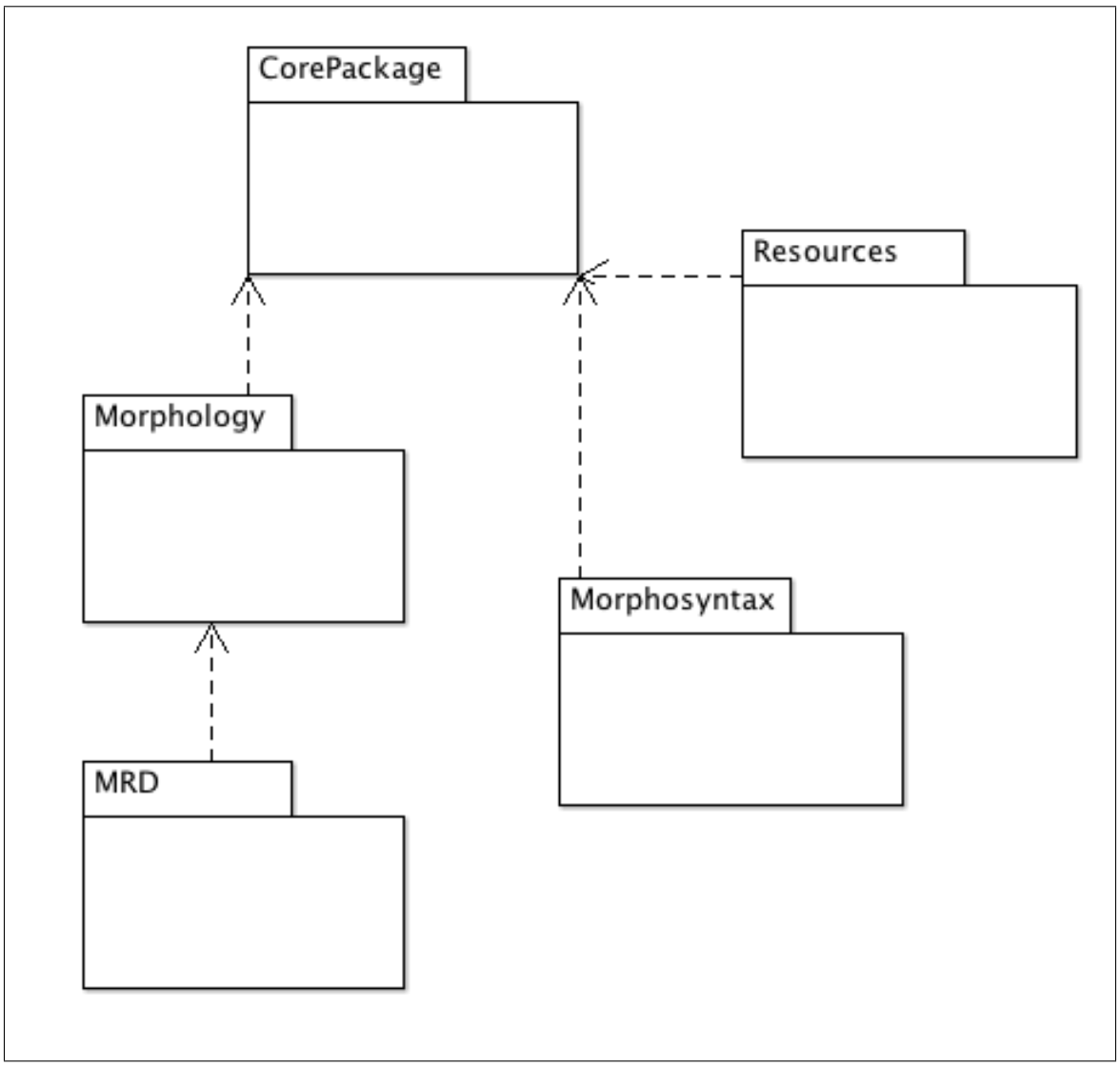

Figure 1: LMF packages

Le *core package* est divisé en deux sous-systèmes :

- l'entrée lexicale, *Lexical Entry*, et ses différentes formes, *Form* (signifiant) ;
- le ou les sens, *Sense* (signifié).

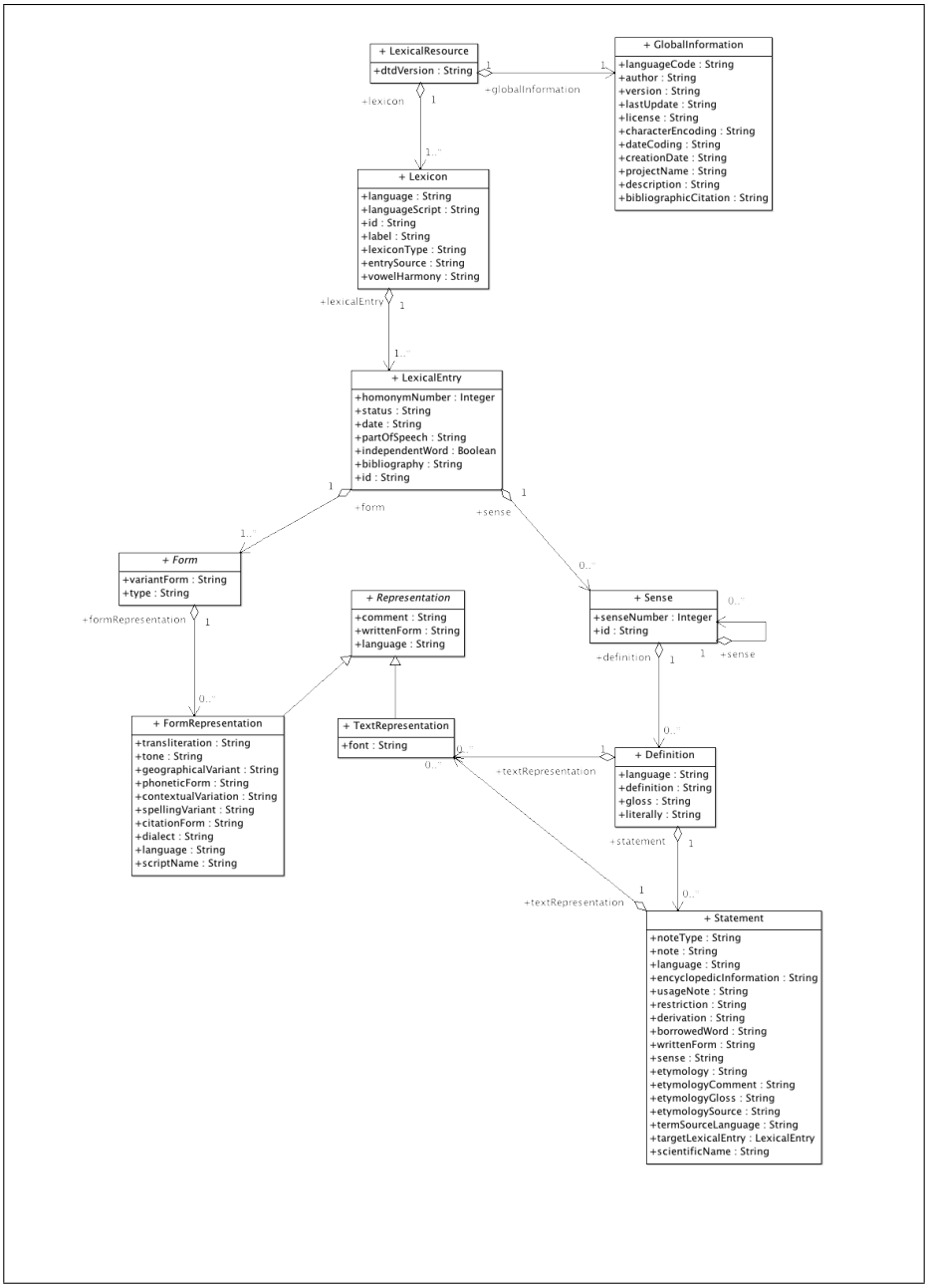

Figure 2: Core Package

Les systèmes périphériques (extensions) sont souples, optionnels, mais puissants. Parmi les 8 extensions proposées, j'en ai sélectionnées certaines qui me paraissaient pertinentes pour nos besoins.

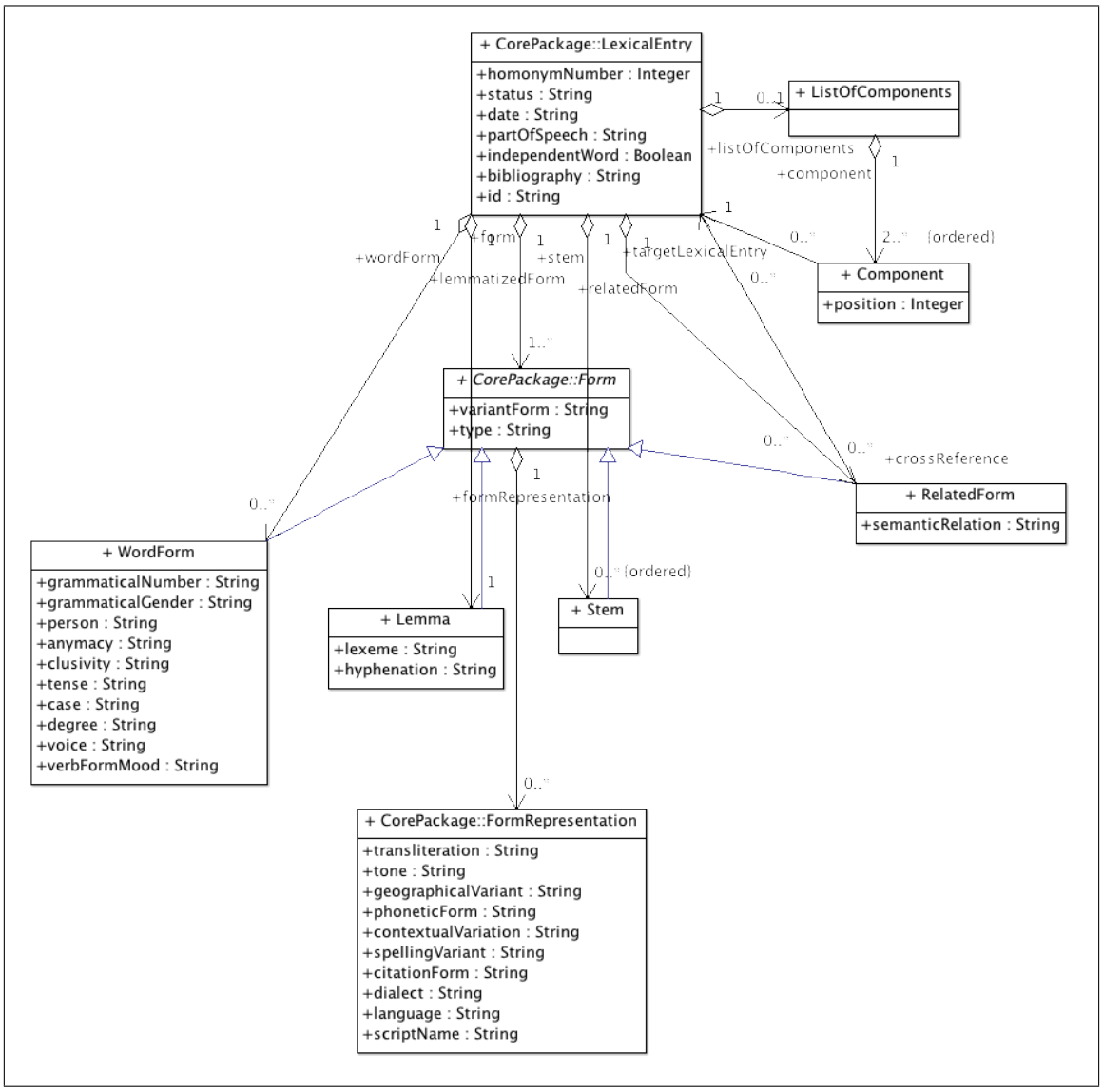

Figure 3: Morphology

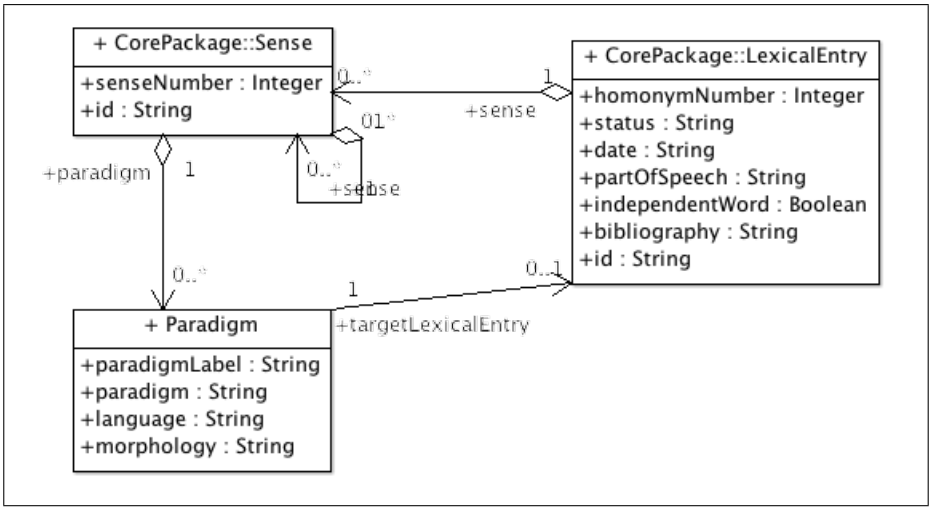

Figure 4: Morphosyntax

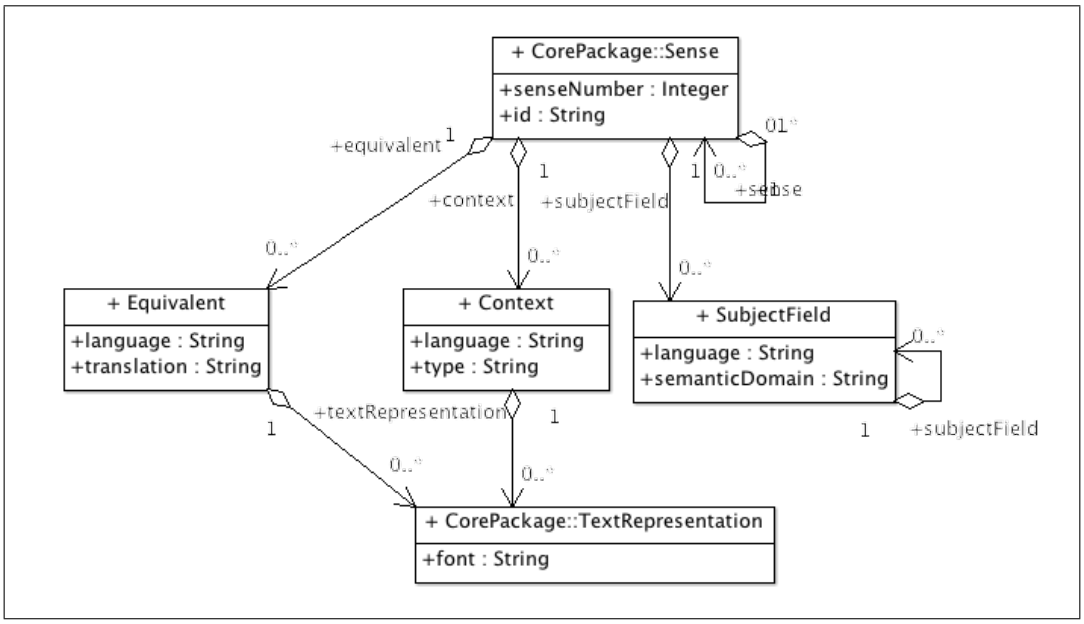

Figure 5: MRD (Machine Readable Dictionary)

En plus des extensions existantes, on peut en créer. C'est ce que je propose de faire pour la gestion des ressources audio et des locuteurs.

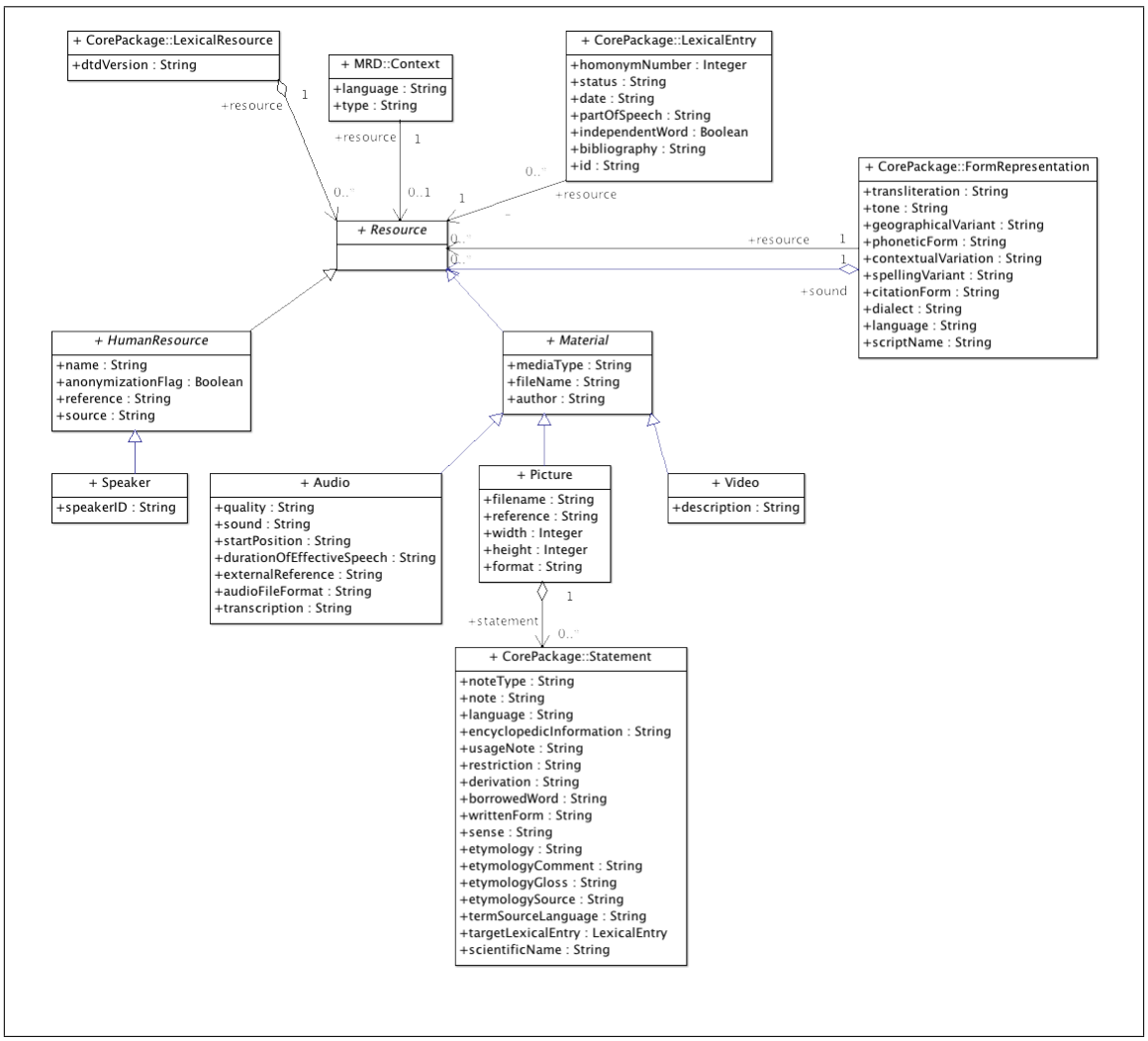

Figure 6: Resources

2 Classes et attributs

Dans cette section, je vais m'attarder sur ce que sont une classe et ses attributs - de manière simplifiée rassurez-vous. Pourquoi ? Car il y a en fait une correspondance directe entre l'architecture logicielle utilisée et le format XML LMF choisi.

2.1 Correspondance entre UML et XML

Un petit exemple afin de visualiser : prenons la classe *Statement* du *Core Package* (en bas à droite de la figure). Cette classe comporte de nombreux attributs, dont les 2 suivants :

- *borrowed word*
- *written form*

En suivant les recommandations LMF, si l'on souhaite par exemple représenter un emprunt de l'anglais du mot *cool*, on obtient les lignes XML suivantes :

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

Plusieurs commentaires sur cet exemple :

- En LMF, les attributs de classe sont struturés en tant que paire d'attributs de la balise spécifique *feat*.
  - Le nom de l'attribut est en fait la valeur de l'attribut *att* de la balise *feat* ;
  - La valeur donnée à cet attribut est la valeur de l'attribut *val* de la balise *feat*.
- Dans cet exemple, il est à noter que conformémement à LMF (et d'ailleurs aussi MDF), la langue d'emprunt doit être renseignée dans l'attribut *borrowed word*, tandis que le mot d'emprunt, lui, est renseigné dans l'attribut *written form*.

3 Pour les novices : qu'est-ce qu'une classe ?

Une classe est une entité abstraite qui représente un objet, par exemple une voiture, et qui comporte certains attributs, par exemple la marque ou la couleur de la voiture. Une classe a également ce qu'on appelle des méthodes, c'est-à-dire des fonctions qu'elle implémente : pour une voiture, ce serait par exemple démarrer, accélérer, etc. Alors que les attributs sont en règle générale matérialisés par des noms communs, les méthodes, elles, sont nommées par des verbes d'action.
D'autre part, une classe peut dériver d'une autre classe, c'est-à-dire en simplifiant qu'elle hérite des attributs et des méthodes de sa classe mère. Cet héritage est représenté sur les schémas UML qui précèdent par une flèche pleine. Par exemple, on pourrait imaginer une classe véhicule, de laquelle dériveraient les classes voiture, moto, etc. Ils auraient tous des attributs en commun (nombre de roues, de portes, marque, couleur du véhicule, etc.) qui seraient donc des attributs de la classe véhicule, et des attributs spécifiques

comme par exemple la béquille pour une moto ou un vélo.
Une classe peut avoir une relation d'agrégation ou de composition avec une autre classe, c'est-à-dire qu'elle en fait partie. Si l'on reprend l'exemple classique de la voiture, et si l'on crée une classe roue, on pourrait dire que la voiture est composée, entre autres, de 4 roues. Cette relation est schématisée par un losange en UML.
Une dernière relation utilisée dans les schémas UML de la section précédente est une simple flèche, qui signifie qu'une classe en référence une autre. Par exemple, une voiture et son propriétaire sont deux entités bien distinctes qui existent indépendamment l'une de l'autre. Cependant, il existe bien un lien entre ces deux entités, représenté par une association.
Enfin, en UML, les classes abstraites sont notées en italique.

3.1 Classes et attributs définis par LMF

Pour chacun des *packages* décrits dans la section précédente, des classes et des relations entre ces classes sont définies et non modifiables (à noter que certains projets existants dévient un peu du standard en proposant des améliorations). Par contre, nous sommes (plus ou moins) libres de définir les attributs que nous souhaitons pour chacune de ces classes. Cependant, chaque attribut doit être référencé dans le DCR (*Data Category Registry*). On peut utiliser les éléments existants, ou bien en proposer de nouveaux le cas échéant. Il s'agit d'une base de données ouverte, accessible sur le site http://www.isocat.org.
Une difficulté que j'ai rencontrée avec cette base de données, c'est qu'il y a beaucoup de redondance, de doublons : beaucoup de termes quasi-identiques sont définis 2 ou 3 fois. Dans ces cas-là, lequel choisir ? Sur quels critères ? J'ai tenté de privilégier la définition qui se rapprochait le plus du besoin, et à définition à peu près similaire, j'ai privilégié les termes issus de MDF, ou créés par Gil Francopoulo (l'auteur du livre sur LMF). Cependant, plutôt que suivre le principe de MDF concernant les marqueurs associés spécifiquement aux langages vernaculaire, régional et national, j'ai choisi de laisser davantage de liberté en définissant un attribut général associé à un attribut de langue (exemple : définition dans la langue 'xxx' plutôt que 'dn' qui impose une définition dans une langue nationale prédéfinie). Ce qui en outre évite d'avoir à définir 'df' par exemple pour le français.

Dans le tableau ci-dessous, je n'ai listé que les attributs de chaque classe, et non les méthodes, car cela alourdirait les spécifications sans apporter d'informations pertinentes. J'ai également noté les marqueurs MDF auxquels les attributs font référence s'ils existent. Quant à l'extension LMF concernée, elle se trouve dans la colonne *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
Core Package	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	-	

Table 1: (continued)

LMF package	Class name	Attribute	Attribute type or example value	DCR PID and type	MDF marker	Comment
		character encoding	“UTF-8”	2564 open	-	
	Lexicon (no subclass)	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	-
		id	“toto_1”	6196 open	lx <id>, se <id>	unique identifier or key form is an XML attribute
	Lexical Entry (no subclass)	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)					

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	
		anony-mization 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
- affix 1234
- bitransitive verb 1275
- classifier 2345
- conjunction 1260
- determiner 1272
- ideophone 4192
- impersonal verb 1306
- interjection 1318

- intransitive verb 1322
- negation 2313
- noun 1333
- numeral 1334
- particle 3372
- possessive pronouns 1359
- preposition 1366
- pronoun 1370
- reflexive verb 5592
- transitive verb 1405
- verb 1424

Valeurs non trouvées dans la 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 Remarques et limitations

1. Les sous-entrées de Toolbox sont codées comme des *Lexical Entry* dont la principale a des liens vers les autres.

2. Avec le modèle proposé, on ne peut pas mettre de référence ('cf') d'un sens vers un autre sens. C'est au niveau de l'entrée lexicale que l'on peut référencer une autre entrée lexicale en tant que synonyme par exemple. Est-ce qu'il y a un besoin de faire ça au niveau des 'sn' (*sense number*) ? Cela ajoute de la complexité au niveau du modèle, mais c'est une amélioration possible. On peut aussi simplifier le modèle si vous pensez que certains attributs voire certaines classes ne sont pas nécessaires.

3. Cas des prédicats complexes VV ou NV : prenons le cas du prédicat complexe NV. En suivant ce modèle LMF, on aurait 3 entrées lexicales :

   - V avec l'attribut *independent word* = *no* ;
   - N avec l'attribut *independent word* = *no* ;
   - NV avec l'attribut *independent word* = *yes*, ayant comme liste de composants (*List Of Components*) un lien vers les 2 entrées lexicales définies ci-dessus.

4 Exemples

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˥, / ɦwɤ˥, / ʝi˧, / la˧, / tʰo˧li˧, / ɱv˧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"/>
```

```
42                  <feat att="language" dcr:datcat="http://www.
                        isocat.org/datcat/DC-2482" val="fra"/>
43                  <feat att="note" dcr:datcat="http://www.isocat.
                        org/datcat/DC-382" val="LM confirmé type "
                        porc""/>
44              </Statement>
45              <Statement>
46                  <feat att="noteType" val="tone"/>
47                  <feat att="note" val="LM"/>
48              </Statement>
49          </Definition>
50          <Definition>
51              <feat att="language" dcr:datcat="http://www.isocat.
                    org/datcat/DC-2482" val="eng"/>
52              <feat att="definition" dcr:datcat="http://www.isocat
                    .org/datcat/DC-1972" val="chicken"/>
53              <feat att="gloss" dcr:datcat="http://www.isocat.org/
                    datcat/DC-244" val="chicken"/>
54          </Definition>
55          <Definition>
56              <feat att="language" val="chn"/>
57              <feat att="definition" val=" "/>
58              <feat att="gloss" val=" "/>
59          </Definition>
60          <Definition>
61              <feat att="language" val="fra"/>
62              <feat att="definition" val="poulet, poule"/>
63              <feat att="gloss" val="poulet"/>
64          </Definition>
65          <Paradigm targets="mi1">
66              <feat att="paradigmLabel" dcr:datcat="http://www.
                    isocat.org/datcat/DC-3741" val="classifier"/>
67              <feat att="paradigm" dcr:datcat="http://www.isocat.
                    org/datcat/DC-3736" val="mi"/>
68          </Paradigm>
69          <Context>
70              <TextRepresentation>
71                  <feat att="language" dcr:datcat="http://www.
                        isocat.org/datcat/DC-2482" val="na?"/>
72                  <feat att="writtenForm" dcr:datcat="http://www.
                        isocat.org/datcat/DC-1836" val="æ dz-ze"/>
73              </TextRepresentation>
74              <TextRepresentation>
75                  <feat att="language" val="eng"/>
76                  <feat att="writtenForm" val="...has eaten (a/
                        some) chicken"/>
77              </TextRepresentation>
78              <TextRepresentation>
79                  <feat att="language" val="chn"/>
80                  <feat att="writtenForm" val=" "/>
81              </TextRepresentation>
82              <TextRepresentation>
83                  <feat att="language" val="fra"/>
84                  <feat att="writtenForm" val="...a mangé (un/du)
                        poulet"/>
```

```
85                          <feat att="comment" dcr:datcat="http://www.
                                isocat.org/datcat/DC-1846" val="PHONO"/>
86                      </TextRepresentation>
87                  </Context>
88                  <Context>
89                      <TextRepresentation>
90                          <feat att="language" val="na?"/>
91                          <feat att="writtenForm" val="æ hwæ- ze"/>
92                      </TextRepresentation>
93                      <TextRepresentation>
94                          <feat att="language" val="eng"/>
95                          <feat att="writtenForm" val="...has bought (a)
                                chicken"/>
96                      </TextRepresentation>
97                      <TextRepresentation>
98                          <feat att="language" val="chn"/>
99                          <feat att="writtenForm" val="  "/>
100                     </TextRepresentation>
101                     <TextRepresentation>
102                         <feat att="language" val="fra"/>
103                         <feat att="writtenForm" val="...a acheté (un/du)
                                poulet"/>
104                         <feat att="comment" val="PHONO"/>
105                     </TextRepresentation>
106                 </Context>
107                 <Context>
108                     <TextRepresentation>
109                         <feat att="language" val="na?"/>
110                         <feat att="writtenForm" val="æ, | hkv, | bo,
                                | hw, | i, | l, | h toli, | mvgv, | bvv
                                , | wæ, | jo, | i"/>
111                     </TextRepresentation>
112                     <TextRepresentation>
113                         <feat att="language" val="eng"/>
114                         <feat att="writtenForm" val="the twelve years of
                                the duodenary cycle"/>
115                     </TextRepresentation>
116                     <TextRepresentation>
117                         <feat att="language" val="chn"/>
118                         <feat att="writtenForm" val="  "/>
119                     </TextRepresentation>
120                     <TextRepresentation>
121                         <feat att="language" val="fra"/>
122                         <feat att="writtenForm" val="les douze signes
                                astrologiques"/>
123                     </TextRepresentation>
124                 </Context>
125             </Sense>
126         </LexicalEntry>
127         <LexicalEntry id="mi_1">
128             <Lemma>
129                 <feat att="lexeme" val="mi"/>
130             </Lemma>
131         </LexicalEntry>
132     </Lexicon>
```

```
133 </LexicalResource>
```

A noter que les attributs *dcr:datcat* peuvent être définis dans la DTD, afin d'alléger le XML.

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-ʁrtɯm, kɯ-ɾŋʲi tsa ci ŋu, ɯ-di mnɤm, ɯ-mɯntoʁ kɯ-ɣɯrni ŋgɯ kɯ-wɣrum tsa ci ŋu, ɯ-zrɤm kɯ-xtɕɯ-xtɕi ma me, wʑo smɤn ɯ-ŋgɯ kɤ-lʁt 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 mnm, -mnto k-rni ŋg 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

Dans les dictionnaires d'Alexandre François, des marqueurs spécifiques ont été utilisés. En voici une liste, ainsi que les équivalences proposées en LMF.

Table 4: Mowtlap dictionary: matching between MDF and LMF

MDF	**Purpose**	**LMF**
wr	*word reference* pour avoir plusieurs 'ps' différents dans la même entrée 'lx', à ne pas confondre avec les sous-entrées	plusieurs *Lexical Entry*
we	détourné pour restriction syntaxique : contexte syntaxique ; notes grammaticales qui spécifient plus précisément le sens en particulier	équivalent : 'ov'
wn	même chose en anglais	équivalent : 'oe'
he	étiquette sémantique pour qualifier le type de relation sémantique : métaphoriquement, sens figuré, etc.	*Related Form semantic relation* : ajouter "métaphore" et "sens figuré"
hn	'he' en anglais	'he' uniquement en anglais
ll	équivalent de 'lt' en anglais	Definition literally with language = "eng"
oe	note sur un exemple	équivalent : 'xc'
on	'oe' en anglais	Text Representation comment with language = "eng"
ur (regional = bislama)	sujet ou possesseur typique ; pour un sens donné, de quel type de sujet c'est le prédicat	Statement usage note
se	peut aussi indiquer la forme préfixé du nom	Form variant form : ajouter type = "prefix"
el	langue de l'étymologie	Statement term source language
dc	date de création	ajouter *creation date* dans *Lexical Entry*
la	forme préfixée pour une entrée, comme 'se' suivi de 'wr'	Form variant form : ajouter type = "prefix"
lg	légende de la photo	Picture statement with note type = "legend"

Table 4: (continued)

ce	glose de ‘cf’ en français	Statement etymology gloss
u	*underlined form* correspondant à ‘a’, destiné au *parser*	Form Representation spelling variant
xm	exemple caché	ajouter un type “exemple caché”
rm	référence d’un exemple caché	Context resource reference
xa	version anglaise d’un exemple caché	Context text representation with language = “eng”
mr	morpho	Paradigm morphology
ue	label	fichier de configuration
un	label en anglais	fichier de configuration
tb	encadré de liste de mots en français	Table written form with type = “word list” and language = “fra” (à ajouter)
ta	équivalent de ‘tb’ en anglais	Table written form with type = “word list” and language = “eng” (à ajouter)
tl	encadré en prose	Table written form with type = “text” and language = “fra” (à ajouter)
tn	équivalent anglais de ’tl’	Table written form with type = “text” and language = “eng” (à ajouter)

Syntaxe spécifique utilisée :

- “ax:” pour un texte en italique : à remplacer par “fi:”
- mini-chevrons pour indiquer l’objet syntaxique : *Statement usage note*

4.4 Tamang

Il s'agit du dictionnaire de Martine Mazaudon, écrit dans Word et basé sur le format LEXWARE. Voici une liste (qui se veut exhaustive) des marqueurs qui ont été utilisés, ainsi que leurs équivalents en MDF ou 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	si plusieurs sens	sn
ton	de 0 à 5 ; notés x,x si hésitation	np
dff		df
dfe		de
dfn	nepali (langue nationale)	dn
dfzoo	définition zoologique	sc
dfbot	définition botanique	sc
nbbot	remarques sur le champ botanique	Definition statement
nag	translittération nagari (écriture locale)	Form Representation transliteration with script name = "nagari"
phr	phrase : exemple de phrases incomplètes	Context with type = "'incomplete' (à ajouter)
il	illustration : exemple	xv
ilnep		xn
gram		ng
rec	enregistrements	sf
xr	cross-reference	cf
nb	nota bene	nt
nbi	'i' pour interne	nq
emp	langue d'emprunt	bw
check	note personnelle	status
sem	champ sémantique	sd
enc	notes encyclopédiques	ee
inf	informateurs	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

Table 5: (continued)

morph		Paradigm morphology
var		va
niv	niveau de langue ?	à ajouter ?
ps		ps
so		so
cons	?	
comp	?	
conj	?	
stedt	?	

Syntaxe spécifique utilisée :

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

4.5 Limbu

Il s'agit du dictionnaire de Boyd Michailovsky, préalablement converti de LEXWARE en XML, dont la structure est décrite ci-dessous.

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/>
```

```
47              <usg/>
48          </hom>
49      </entry>
50 </DICO>
```

Syntaxe spécifique :

Listing 4: Limbu syntax

```
1 <foreign xml:lang=""lif …></foreign>
2 <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

5 A venir

- Mettre à jour la DTD, et la convertir en schéma XSD.