Dbnary: wiktionary et le web lexical

Wiktionnaire, RDF, Linked Data: une colonne vertébrale pour le lexique ?

Gilles Sérasset

GETALP-LIG, Université Joseph Fourier, Grenoble, France

2nd July 2013

Propos liminaires

Cette présentation à été préparée avec une méthode ancestrale...

Propos liminaires

Cette présentation à été préparée avec une méthode ancestrale...

C'est la méthode ALARACHE !!

Cela signifie que je compte sur vous! Rendez la vivante et complète en m'interrompant!
Ou nous nous ferons tous Magn(ennuyer)

Propos liminaires

Cette présentation à été préparée avec une méthode ancestrale...

C'est la méthode ALARACHE !!

Cela signifie que je compte sur vous! Rendez la vivante et complète en m'interrompant!
Ou nous nous ferons tous ch... comme la mort (bored to death)

Qui suis je ?

Je travaille sur la structuration/gestion/... des bases lexicales multilingues (MLDB) depuis 1990.

Thèse: définition d'une approche pivot pour les MLDB et d'un système générique de gestion des MLDB (1994)

Responsable projet Papillon depuis 2000.

Mes contributions principales ?

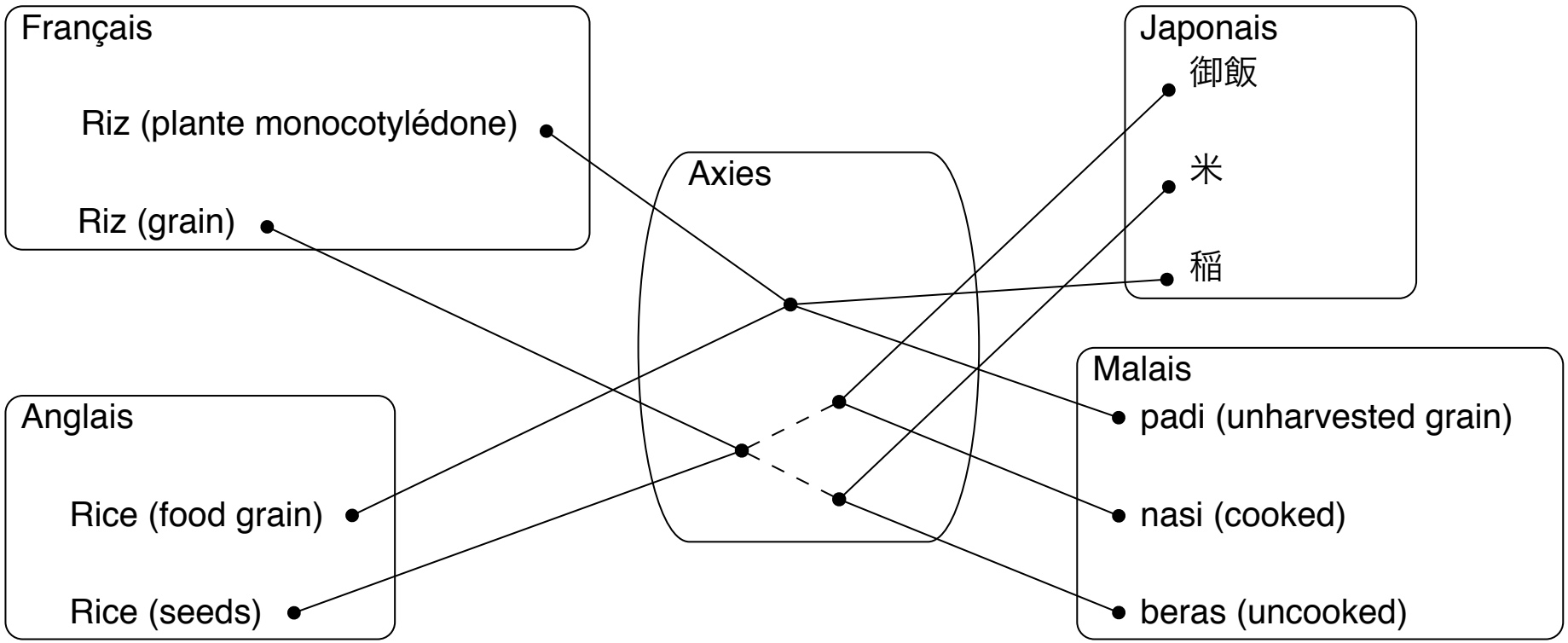
- Approche par acceptions interlingues (Axies)
  - évite les biais de l'utilisation d'une langue naturelle comme pivot
  - Reprise dans le standard Lexicon Markup Framework
- Outils génériques pour les MLDB
  - plateforme JIBIKI
  - utilisée pour le projet Papillon
  - réutilisée (ainsi que l'approche par acception) pour le projet LexALP
  - ... GDEF, DILAF (Mathieu Mangeot)
- Réflexion sur les lexiques, leur acquisition
  - Modèle de graphes: Lexical systems,
  - corpus vu comme un graphe
  - extraction d'informations lexicales vu comme des opérations sur des graphes
  - ... Thèse de Vincent Archer

Overview of the Papillon Project

- Initialy launch by a French Japanese consortium (2000)
- Quickly extended itself to a larger consortium
  - with partners in Australia, Canada, China, France, Germany, Japan, Malaysia, Thailand, Vietnam...
  - ...and accepting any new partner motivated enough to join the adventure...
- Aims at the development of a rich, "open source", multilingual lexical database

Architecture of the Data

Macrostructure: An acception based multilingual lexical database



Architecture of the Data

A microstructure inspired by Mel'čuk's ECD and Polguère's DICO

regretter ,

v.tr.

sentiment LA personne X ~ SON action Y

GOVERNMENT PATTERN

X = I Y = II

1 . N
1 . N
2 . de V-inf

LEXICAL FUNCTIONS

QSyn : se repentir

S0 : regret#1

Able2 : (*Que l'on peut R.*) regrettable

Magn : (*Intensément*) beaucoup

Y étant grave, Magn : amèrement , cruellement ; _se mordre les doigts_

EXAMPLES

1 . *C'est une décision qu'il va regretter cruellement.*

2 . *Il ne regrette pas d'avoir investi 4 000 F dans ce nouveau programme.*

Mes réflexions actuelles

- Interopérabilité "syntaxique"
  - Biais d'XML dans jibiki
  - ... Passage en RDF...
  - ... Lexical Linked Data.
- Interopérabilité "sémantique"
  - Relier/aligner les sens de différentes resources
  - techniques proches du WSD
  - Besoin de sens décrits, avec des définitions
  - ... difficulté d'en construire.
- Multilinguisme
  - modèle par acceptions toujours d'actualité
  - besoin de bases lexicales multilingues encore d'actualité (même si on est passé d'une disette à une sur-abondance d'information lexicale)
  - importance des informations adjectivales, verbales, adverbiales
  - focalisation du domaine sur le nominal (wikipedia)

Dbnary: une première action

- Un grand dictionnaire, gratuit, avec des définitions ?
  - Wiktionary
- Intéropérabilité "à grain fin"
  - RDF
- Un modèle standard
  - LEMON (inspiré de LMF)
- Multilingue
  - extraction de plusieurs "language edition" de Wiktionary
- Disponible, utilisable et utilisé
  - Linked Data

Dbnary: Wiktionary comme graphe lexical

English [edit]

Pronunciation [edit]

- (*UK*) IPA: /kæt/, [kʰæt]
- (*US*) IPA: /kæt/, [kʰæʔ(t̚)], [kʰe̞ə̯t]
- Audio (UK) 0:00 MENU
- Audio (US) 0:00 MENU
- Audio (US-Inland North) 0:00 MENU
- Rhymes: -æt

Etymology 1 [edit]

From Middle English *cat*, *catte*, from Old English *catt* ("male cat") and *catte* ("female cat"), from Late Latin *cattus* ("domestic cat"), from Latin *catta* (c.75 B.C., Martial)[1], from Afro-Asiatic (compare Nubian *kadís*, Berber *kaddîska* 'wildcat'), from Late Egyptian *čaute*,[2] feminine of *čaus* 'jungle cat, African wildcat', from earlier Egyptian *tešau* 'female cat'. Cognate with Scots *cat* ("cat"), Welsh *cath* ("cat"), West Frisian *kat* ("cat"), North Frisian *kåt* ("cat"), Dutch *kat* ("cat"), Low German *Katt, Katte* ("cat"), German *Katze* ("cat"), Danish *kat* ("cat"), Swedish *katt* ("cat"), Icelandic *köttur* ("cat"), Armenian կատու (*katu*, "cat").

Noun [edit]

cat (*plural* **cats**)

1. A domesticated subspecies, *Felis silvestris catus*, of feline animal, commonly kept as a house pet. [from 8th c.]
2. Any similar animal of the family Felidae, which includes lions, tigers, bobcats, etc.
3. A catfish. [quotations ▼]
4. (*offensive*) A spiteful or angry woman. [from earlier 13th c.]
5. An enthusiast or player of jazz.
6. (*slang*) A person (usually male).
7. (*nautical*) A strong tackle used to hoist an anchor to the cathead of a ship.
8. (*chiefly nautical*) *Short form of* cat-o'-nine-tails. [quotations ▼]
9. (*slang*) Any of a variety of earth-moving machines. (from their manufacturer Caterpillar Inc.)
10. (*archaic*) A sturdy merchant sailing vessel (*now only in "catboat"*).
11. (*archaic*, *uncountable*) The game of "trap and ball" (also called "cat and dog").
12. (*archaic*, *uncountable*) The trap of the game of "trap and ball".
13. (*slang*) Prostitute. [from at least early 15th c.]
14. (*slang*, *vulgar*, *African American Vernacular*) A vagina; female external genitalia [quotations ▼]
15. A double tripod (for holding a plate, etc.) with six feet, of which three rest on the ground, in whatever position it is placed.

Wikipedia has an article on:
Cat

A domestic cat (1)

Dbnary: Wiktionary comme graphe lexical

Synonyms [edit]

- (*any member of the suborder (sometimes superfamily) Feliformia or* Feloidea): feliform ("cat-like" carnivoran), feloid (compare Caniformia, Canoidea)
- (*any member of the family Felidae*): felid
- (*any member of the subfamily Felinae, genera* Puma, Acinonyx, Lynx, Leopardus, *and* Felis*)*): feline cat, a feline
- (*any member of the subfamily Pantherinae, genera* Panthera, Uncia *and* Neofelis): pantherine cat, a pantherine
- (*technically, all members of the genus* Panthera): panther (i.e. tiger, lion, jaguar, leopard), (*narrow sense*) panther (i.e. black panther)
- (*any member of the extinct subfamily* Machairodontinae, *genera* Smilodon, Homotherium, Miomachairodus, *etc.*): Smilodontini, Machairodontini (Homotherini), Metailurini, "saber-toothed cat" (saber-tooth)
- (*domestic species*): housecat, puss, pussy, malkin, kitten, kitty, pussy-cat, mouser, tomcat, grimalkin
- (*man*): bloke (*UK*), chap (*British*), cove (*UK*), dude, fellow, fella, guy
- (*spiteful woman*): bitch
- See also Wikisaurus:cat
- See also Wikisaurus:man

Derived terms [edit]

Terms derived from *cat* in the above senses [show ▼]

Dbnary: Wiktionary comme graphe lexical

Translations [edit]

domestic species	[show ▼]
member of the suborder (or superfamily) Feliformia (Feloidea), "cat-like" carnivorans	[show ▼]
member of the family Felidae	[show ▼]
member of the subfamily Felinae	[show ▼]
member of the subfamily Pantherinae	[show ▼]
member of the extinct subfamily Machairodontinae	[hide ▲]

Select targeted languages

- Esperanto: maĥairodeno (eo)
- French: machairodontiné (fr) *m*, machairodontinés (fr) *pl*
- German: Säbelzahnkatze (de) *f*, Machairodontine (de) *m*, Machairodontine (de) *f*, Machairodontinen (de) *pl*, Machairodontinae (de) *pl*
- Add translation : Preview translation More

Script template: (e.g. Cyrl for Cyrillic, Latn for Latin)

type of fish — *see* **catfish**	
spiteful woman — *see* **bitch**	
jazz enthusiast	[show ▼]
guy, fellow	[show ▼]
strong tackle used to hoist an anchor to the cathead of a ship	[show ▼]
cat-o'-nine-tails — *see* **cat-o'-nine-tails**	
type of boat — *see* **catboat**	
game of "trap and ball" (or "cat and dog")	[show ▼]
the trap in the game of "trap and ball"	[show ▼]

Dbnary: Wiktionary comme graphe lexical

```
==English==
[[Category:English three-letter words]]{{rfc-auto}}
{{wikipedia}}
[[Image:Cat03.jpg|thumb|A domestic cat (1)]]

===Pronunciation===
* {{a|UK}} {{IPA|/kæt/|[kʰæt]}}
* {{a|US}} {{IPA|/kæt/|[kʰæʔ(t̚)]|[kʰeə̯t]}}
* {{audio|En-uk-a cat.ogg|Audio (UK)}}
* {{audio|En-us-cat.ogg|Audio (US)}}
* {{audio|En-us-inlandnorth-cat.ogg|Audio (US-Inland North)}}
* {{rhymes|æt}}

===Etymology 1===
From {{etyl|enm|en}} {{term|cat|lang=enm}}, {{term|catte|lang=enm}}, from {{etyl|ang|en}} {{term|catt||male cat|lang=ang}} and

====Noun====
{{en-noun}}

# A domesticated [[subspecies]], ''[[Felis silvestris catus]]'', of [[feline]] animal, commonly kept as a house [[pet]]. {{defdate|fro
# Any similar animal of the family [[Felidae]], which includes [[lion]]s, [[tiger]]s, bobcats, etc.
# A [[catfish]].
#* '''1913''', [[w:Willa Cather|Willa Cather]], ''[[s:O Pioneers!|O Pioneers!]]'', [[s:O Pioneers!/The Wild Land, II|chapter 2]]:
#*: She missed the fish diet of her own country, and twice every summer she sent the boys to the river, twenty miles to the sou
# {{context|offensive|lang=en}} A spiteful or angry [[woman]]. {{defdate|from earlier 13th c.}}
# An enthusiast or player of [[jazz]].
```

Dbnary: Wiktionary comme graphe lexical

=====Synonyms=====
* {{sense|any member of the [[suborder]] (sometimes [[superfamily]]) [[Feliformia]] or {{taxlink|Feloidea|suborder}}}} [[feliform]
* {{sense|any member of the [[family]] [[Felidae]]}} [[felid]]
* {{sense|any member of the [[subfamily]] [[Felinae]], genera ''[[Puma]]'', ''[[Acinonyx]]'', ''[[Lynx]]'', ''[[Leopardus]]'', and ''[[Fe
* {{sense|any member of the subfamily [[Pantherinae]], genera ''[[Panthera]], [[Uncia]]'' and ''[[Neofelis]]''}} [[pantherine cat]],
* {{sense|technically, all members of the genus ''Panthera''}} [[panther]] (i.e. [[tiger]], [[lion]], [[jaguar]], [[leopard]]), {{qualifier|
* {{sense|any member of the [[extinct]] subfamily ''{{taxlink|Machairodontinae|subfamily}}'', genera {{taxlink|Smilodon|genus|r
* {{sense|domestic species}} [[housecat]], [[puss]], [[pussy]], [[malkin]], [[kitten]], [[kitty]], [[pussy-cat]], [[mouser]], [[tomcat]],
* {{sense|man}} [[bloke]] {{qualifier|UK}}, [[chap]] {{qualifier|British}}, [[cove]] {{qualifier|UK}}, [[dude]], [[fellow]], [[fella]], [[guy]
* {{sense|spiteful woman}} [[bitch]]
* See also [[Wikisaurus:cat]]
* See also [[Wikisaurus:man]]

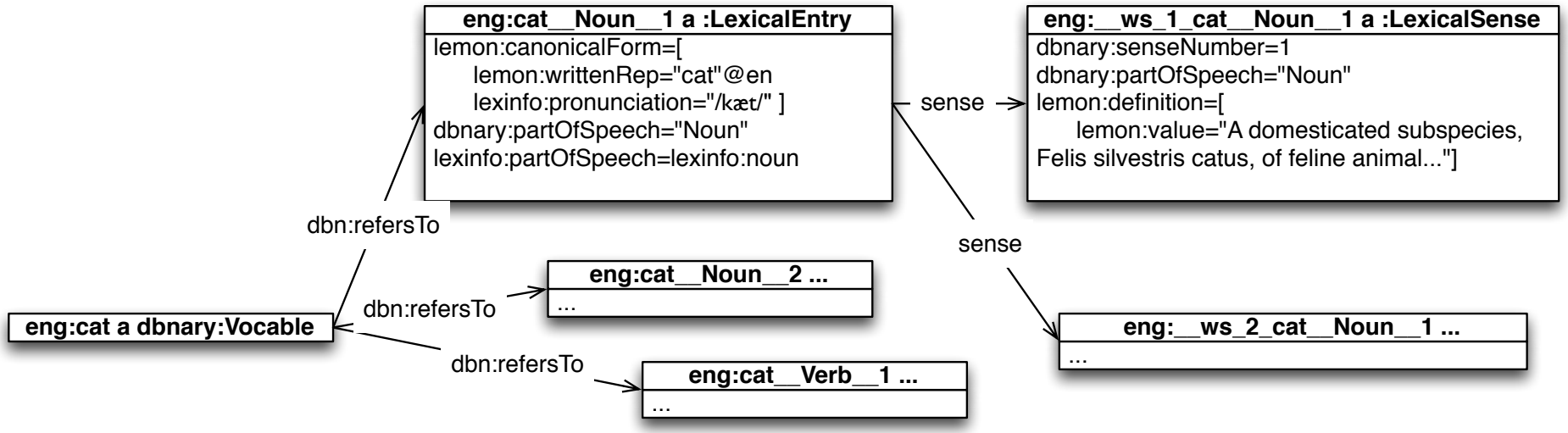
Dbnary: Wiktionary comme graphe lexical

=====Translations=====

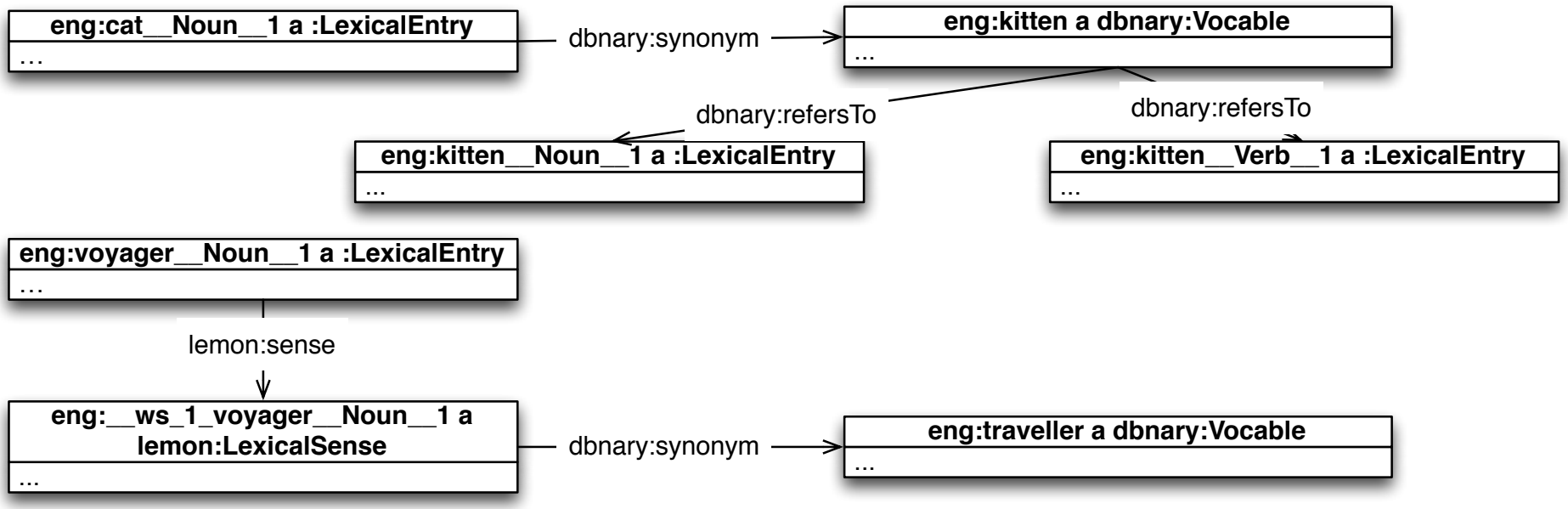
{{trans-top|domestic species}}

* {{trreq|ab}}
* Acehnese: {{tø|ace|mië}}
* Adyghe: {{tø|ady|кIэтыу|sc=Cyrl}}
* Afrikaans: {{t+|af|kat}}
* Ainu: {{tø|ain|チャペ|tr=cape}}
* Akan: [[agyinamoa]] {{n}}
* Albanian: {{t+|sq|mace|f}}
* Alemannic German: {{tø|gsw|Chàtz}}
* Amharic: {{t-|am|ድመት|tr=dəmät|sc=Ethi}}
* Apache:
*: Western Apache: {{tø|apw|gídí}}
* Arabic: {{t+|ar|قط|m|tr=qiṭṭ|sc=Arab}}, {{t+|ar|قطة|f|tr=qíṭṭa|sc=Arab}}
*: Egyptian: {{tø|arz|قط|m|tr='uṭṭ}}, {{tø|arz|قطة|f|tr='uṭṭɑ}}
*: Libyan: {{t-|ar|قطوس|m|tr=gɑṭṭūs|sc=Arab}}, {{t-|ar|قطوسة|f|tr=gɑṭṭūsa|sc=Arab}}
*: Moroccan Arabic: {{tø|ary|مش|tr=mešš}}, {{tø|ary|مشة|f|tr=mešša}}
* Aramaic:
*: Syriac: [[ܩܛܘܠ]] (šūnārā') {{m}}, [[ܩܛܘܠܬܐ]] (šūnārtā') {{f}}

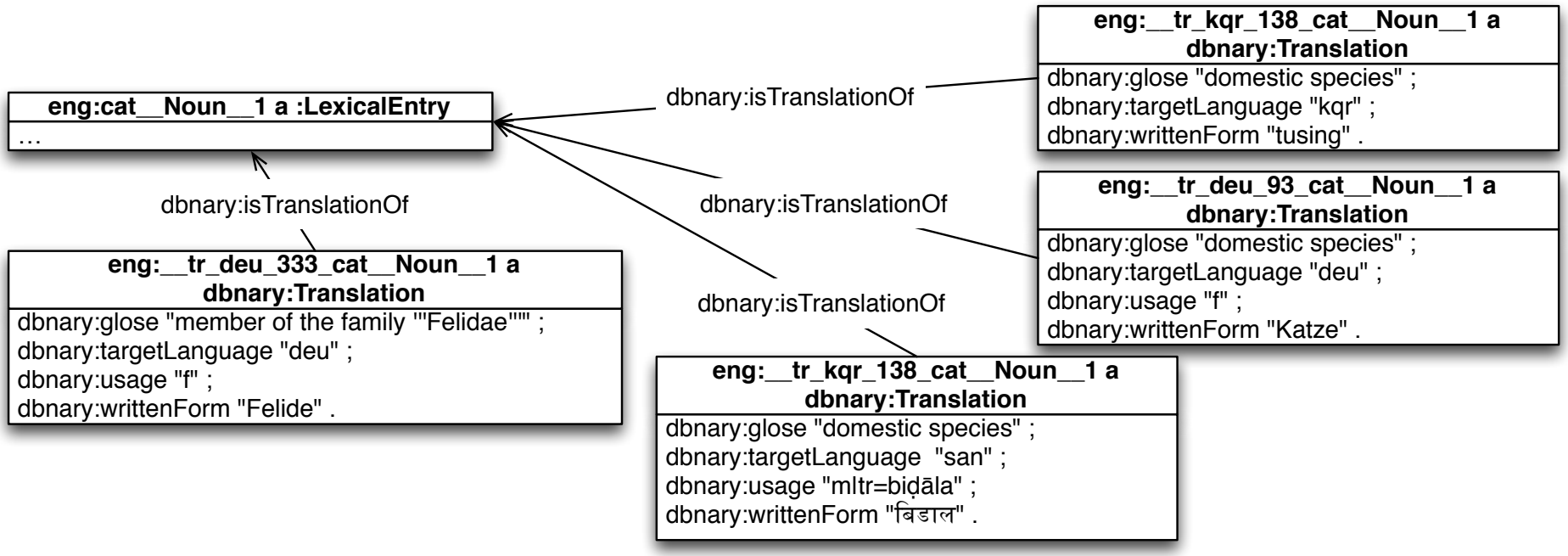
Dbnary: Wiktionary comme graphe lexical



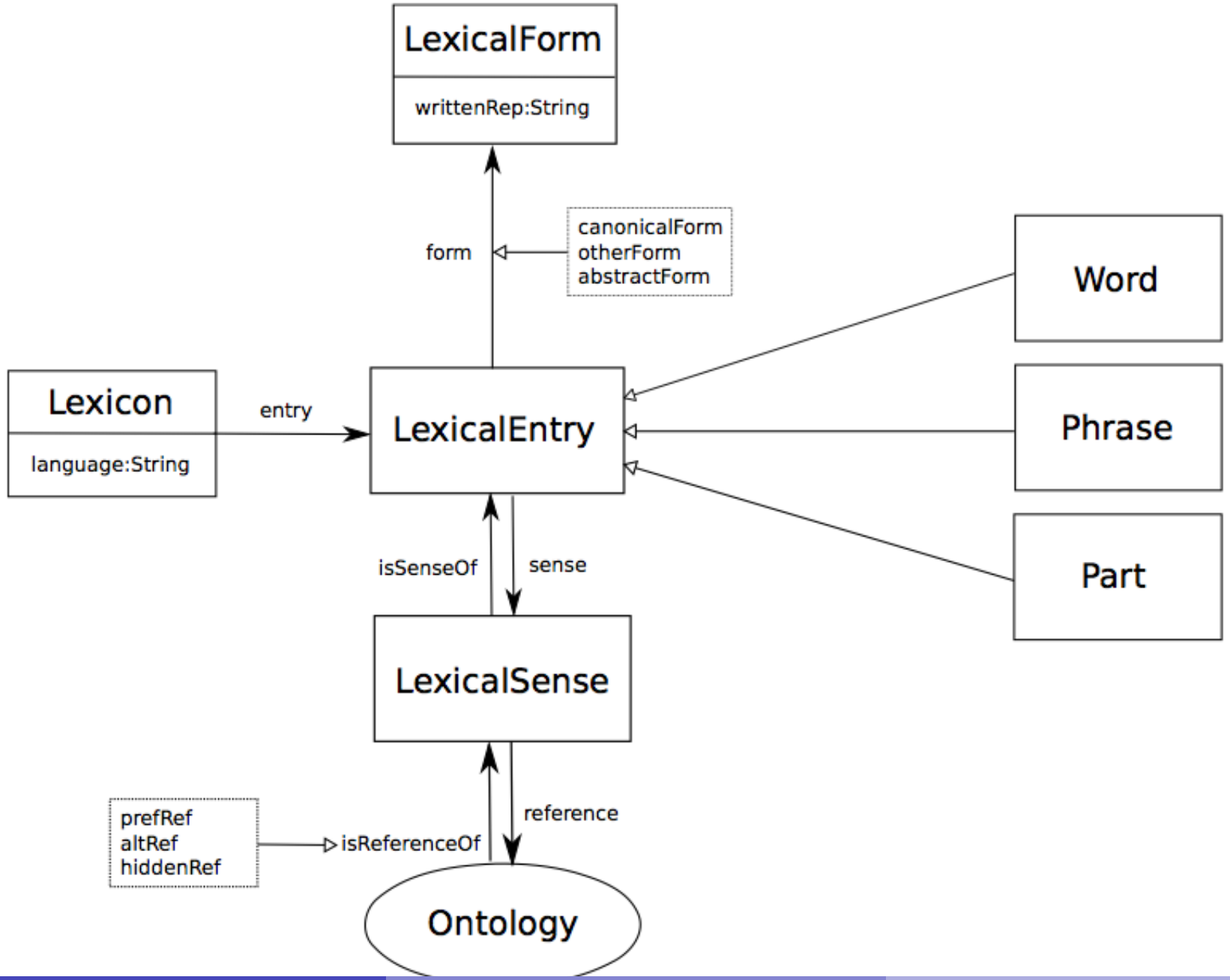
Dbnary: Wiktionary comme graphe lexical



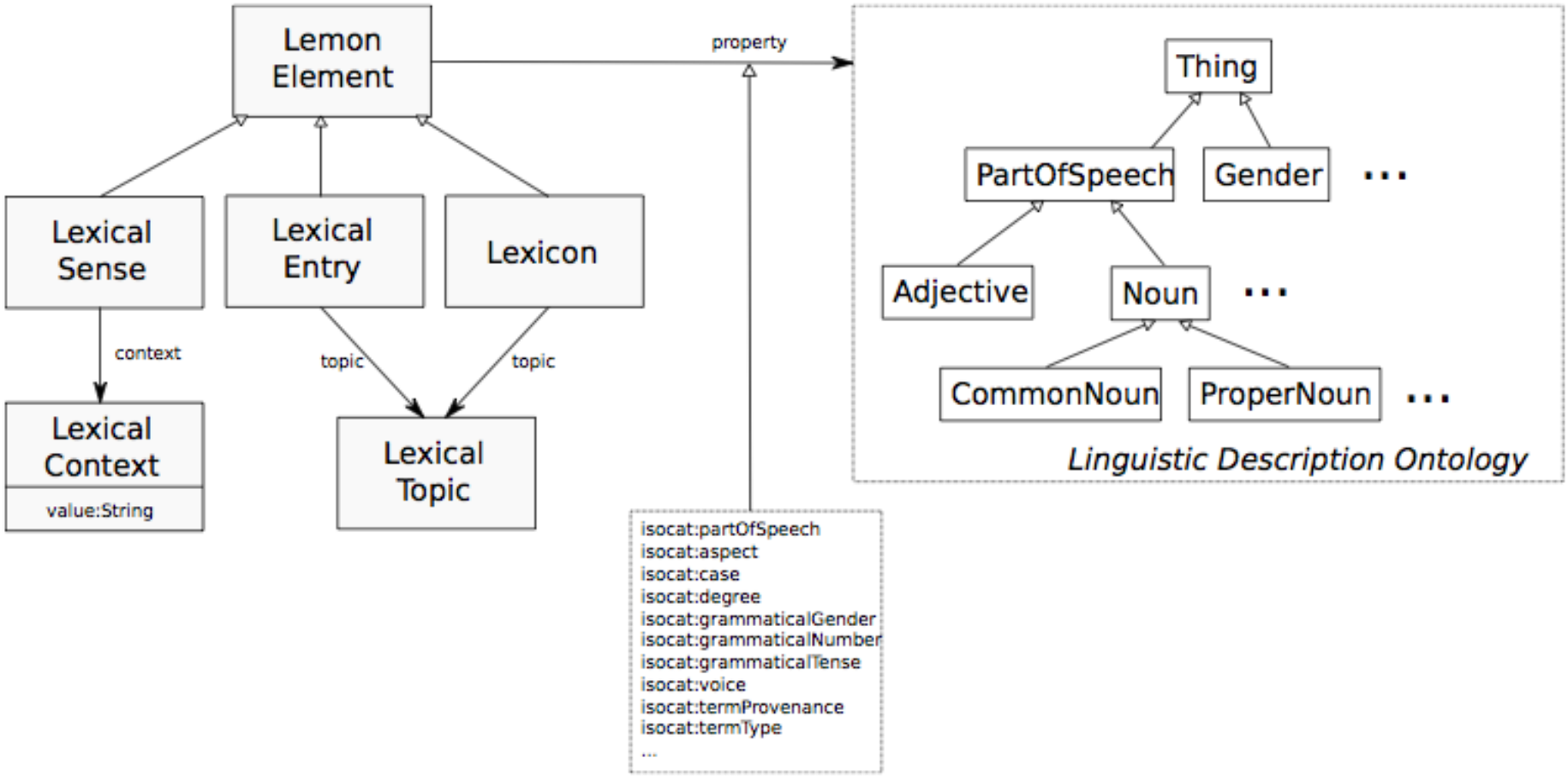
Dbnary: Wiktionary comme graphe lexical



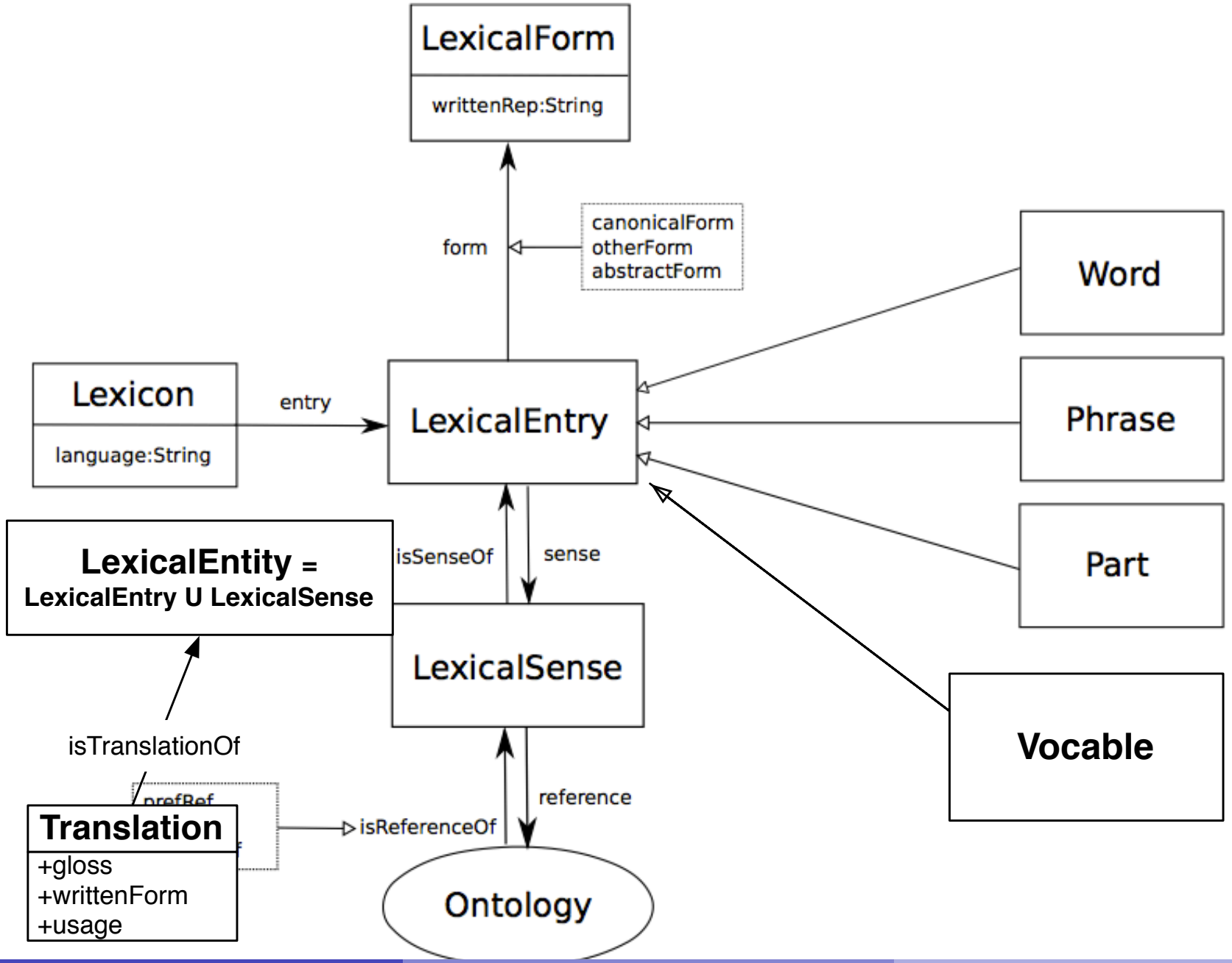
LEMON


LexicalForm
writtenRep:String
form
canonicalForm
otherForm
abstractForm
Word
Lexicon
language:String
entry
LexicalEntry
Phrase
Part
isSenseOf
sense
LexicalSense
prefRef
altRef
hiddenRef
isReferenceOf
reference
Ontology

LEMON


Lemon Element
property
Lexical Sense
Lexical Entry
Lexicon
context
topic
topic
Lexical Context
value:String
Lexical Topic
isocat:partOfSpeech
isocat:aspect
isocat:case
isocat:degree
isocat:grammaticalGender
isocat:grammaticalNumber
isocat:grammaticalTense
isocat:voice
isocat:termProvenance
isocat:termType
...
Thing
PartOfSpeech
Gender
Adjective
Noun
CommonNoun
ProperNoun
Linguistic Description Ontology

LEMON


LexicalForm
writtenRep:String
form
canonicalForm
otherForm
abstractForm
Word
Lexicon
language:String
entry
LexicalEntry
Phrase
Part
LexicalEntity =
LexicalEntry U LexicalSense
isSenseOf
sense
LexicalSense
Vocable
isTranslationOf
Translation
+gloss
+writtenForm
+usage
isReferenceOf
reference
Ontology

Principe d'extraction

- 8 éditions extraites: en, fr, de, ru, it, pt, fi, el
- Extraction à partir des dumps de wikimedia
- Chaque édition met à jour son dump tous les 10-15j
- l'extraction est faite en direct, les fichiers sont accessibles en ligne immédiatement
- le serveur "Linked Data" contient des données plus anciennes

Taille des données (aujourd'hui)

	Entries	Vocables	Senses	Translations
eng	527067	504594	421232	1126463
fra	273822	283847	358921	464956
deu	135103	201736	95593	471892
rus	127271	139235	99243	325345
ell	74056	74800	34932	55652
fin	48164	48050	56559	118728
por	43042	44061	77631	225065
ita	25279	31935	35061	57796

Table : Number of resources by type and language, sorted by number of lexical entries.

Taille des données (aujourd'hui)

	syn	ant	hyper	hypo	mero	holo
eng	31461	6877	959	1103	114	0
fra	30088	6735	8215	3557	943	1847
deu	27516	14315	30202	9509	0	0
rus	22631	9204	21028	4756	0	0
ell	3975	1116	0	0	0	0
fin	2255	0	0	0	0	0
por	3527	575	6	3	0	0
ita	7091	2337	0	0	0	0

Table : Number of lexicon-semantic relations. Languages are sorted according to their number of lexical entries.

Taille des données (aujourd'hui)

Source/Target	deu	ell	eng	fin	fra	ita	por
eng	62501	23794	1	74938	57959	37467	30256
fra	34608	7063	74687	7589	12	18806	17784
deu	0	2675	81015	4947	67143	41485	8872
rus	23056	3295	48559	3966	14776	12643	5567
ell	2242	2	10090	1056	8436	1470	1149
fin	8046	918	30103	0	6700	3856	2196
por	7000	2816	11284	4607	8720	7096	4
ita	4619	506	17539	925	4461	75	1219

Table : Number of translations from/to the 8 currently extracted languages. Source languages are sorted according to their number of lexical entries. Target languages are sorted by their ISO 639-3 language code. The number of different target languages is also given.

Taille des données (aujourd'hui)

Source/Target	por	rus	others	Total	# of lang
eng	30256	74837	764710	1126463	1143
fra	17784	7783	296624	464956	952
deu	8872	17354	248401	471892	355
rus	5567	0	206709	318571	490
ell	1149	1315	29892	55652	246
fin	2196	7997	58912	118728	329
por	4	4396	179142	225065	695
ita	1219	938	27514	57796	315

Table : Number of translations from/to the 8 currently extracted languages. Source languages are sorted according to their number of lexical entries. Target languages are sorted by their ISO 639-3 language code. The number of different target languages is also given.

Qualité des données

- Évaluer la qualité des données est difficile
- Qualité de l'extraction $\neq$ qualité des données
- Pas d'application particulière actuellement
- Pas encore d'alignement avec d'autres ressources (wordnet, jeuxdemots, ...)
- Mais quelques indices...

Qualité des données (hier)

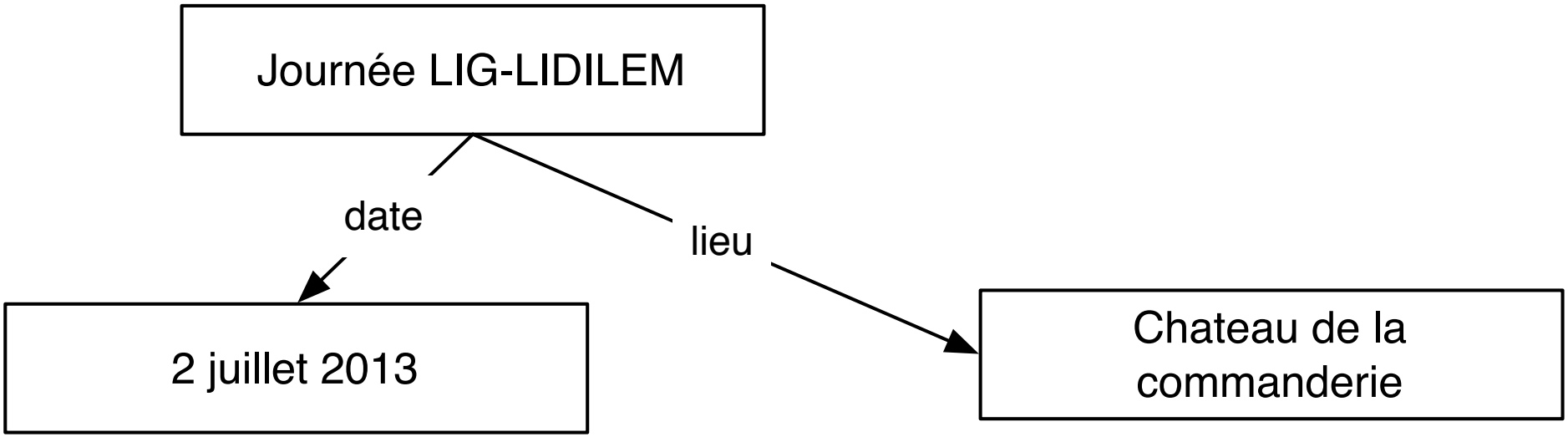
language	# of transl.
eng	5110 (99.1 %)
fra	5799 (107.0 %)
deu	10287 (99.2 %)
rus	8436 (24811.7 %)
ell	2598 (64.3 %)
fin	7245 (28980 %)
por	17720 (93.2 %)
ita	7855 (3167.3 %)

Table : Extracted translations vs interwiki links, on a random sample of 1000 entries.

RDF: Resource Description Framework

Un graphe = un ensemble de Statements

un statement = un triplet: (S, P, O)



RDF: Resource Description Framework

Un graphe = un ensemble de Statements

un statement = un triplet: (S, P, O)

S: forcément une URI; P: une URI; O: une URI ou un littéral

RDF: Resource Description Framework

Un graphe = un ensemble de Statements

un statement = un triplet: (S, P, O)

Attention: une URI n'est pas forcément une document URL

`http://getalp.imag.fr/Journée_LIG-LIDILEM` : EROR 404: Not found !

RDF: Resource Description Framework

Comment représenter un document RDF ?

RDF est un modèle, pas une syntaxe...

On le confond souvent avec RDF+XML une syntaxe absconse...

Mais il y a des tas de manière de transférer des données RDF: Turtle, N3, HTML+RDFa...
Les outils les comprennent souvent toutes...

RDF: Resource Description Framework

Exemple en turtle...

```
@prefix getalp: <http://getalp.imag.fr/>
@prefix sem: <http://seminaires.fr/38/>
@prefix dcterm: <http://purl.com/dcterm>

getalp:Journée_LIG-LIDILEM
    dcterm:date "2 juillet 2013"@fr
    dcterm:lieu sem:chateau_commanderie .
```

RDF: Resource Description Framework

Quoi d'autre ?

On peut aussi décrire les meta-donnée en RDF (OWL)

par exemple:
l'objet d'une relation "lieu" est forcément un lieu géographique.
un "dcterm:event" a forcément une date, un lieu et au moins un participant...

Un raisonnement est possible:
`getalp:Journée_LIG-LIDILEM a dcterm:event`
implique qu'il existe un participant...

RDF: Resource Description Framework

Quoi d'autre ?

De nombreux outils existent

par exemple:
Bases de données spécialisées (virtuoso, JENA, Sesame, etc...)
Langage de requêtes: SPARQL
Raisonneurs

Et le "Linked Data" ? Koikesse ?

En gros (très gros), les URIs deviennent des URLs

On peut donc "résoudre" (déréférencer) les noeuds de l'arc

Tout se passe comme si chaque noeud était à la foi
un objet du modèle... (comme en RDF)
une page web décrivant l'objet pour un utilisateur...
un document RDF décrivant l'objet pour un outil du web sémantique

C'est le serveur web qui "négocie" et redirige vers l'info adéquate.

Et le "Linked Data" ? safékoi ?

Si les données sont disponibles sur les serveurs, il est inutile de les avoir "en local"...

Elles sont dispersées

On peut donc utiliser des données qu'on n'a pas "chez soi".

Conclusion

- Il reste des données à construire (langues peu dotées, collocations, etc)
- Mais pour les grandes langues, c'est l'abondance
- Besoin fort actuel: l'interopérabilité
- Une réponse: le Lexical Linked Data
- Une ressource à laquelle "attacher" les autres données: dbnary
- Suite: désambiguïser pour "monter" les relations au niveau des sens
- Suite: faire émerger un pivot lexical non linguo-centré
- Suite: aligner aver d'autres resources RDF (Wordnet, JeuxDeMots (trivial à RDFiser), ...)