DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF

Gilles Sérasset *
GETALP-LIG, UJF-Grenoble 1
BP 53, 38051 Grenoble cedex 9, France
gilles.serasset@imag.fr

Abstract. Contributive resources, such as Wikipedia, have proved to be valuable to Natural Language Processing or multilingual Information Retrieval applications. This work focusses on Wiktionary, the dictionary part of the resources sponsored by the Wikimedia foundation. In this article, we present our effort to extract multilingual lexical data from Wiktionary data and to provide it to the community as a Multilingual Lexical Linked Open Data (MLLOD). This lexical resource is structured using the LEMON Model.
This data, called *DBnary*, is registered at http://thedatahub.org/dataset/dbnary.

Keywords: Wiktionary, Multilingual, Lexical Resource, LEMON, Multilingual Linguistic LOD

1. Introduction

The GETALP (study group for speech and language translation/processing) team of the LIG (Laboratoire d'Informatique de Grenoble) is in need of multilingual lexical resources that should include language correspondences (translations) and word sense definitions. In this regard, the data included in the different Wiktionary[1] language editions is a precious mine.

Alas, many inconsistencies, errors, difference in usage do exist in the different Wiktionary language editions. Hence, we decided to provide an effort to extract precious data from this source and provide it to the community as Linked Data. After a first version that used an RDF version of the LMF model [4,3] and described in [10], we decided to adapt our extractors to LEMON model [8]. This linked dataset has won the "Monnet-Challenge" in 2012.

2. Extracting data from Wiktionary

2.1. Motivation

Errors and incoherences are inherent to a contributive resource like Wiktionary. Some language editions (like French and English) have many moderators who do limit the number of incoherences among entries of the same language. Such languages, which contain the most data, use many *templates* that simplify the extraction process. For instance, the translation section of the French Wiktionary always uses a template to identify each individual translation (e.g. `{{trad+|de|Katze}} "f"` is the wiki code used for one of the German translation of the French word *chat*).

This is not true anymore with less developed Wiktionary language editions. For instance, in the Finnish edition, some translations into French are introduced by the appropriate template (i.e. `{{fr}}`) while others are introduced by the string `ranska` which is the Finnish translation for "French". In this case the extractor needs to know the Finnish translation of all language names to cope with the second case and avoid losing almost half of the available translation data.

[*] E-mail: Gilles.Serasset@imag.fr.
[1] http://www.wiktionary.org

Such inconsistencies and some errors in the data render the development of an extractor quite tedious. As many people in NLP is trying to use this data for different applications, we decided that we would extract lexical data from as much Wiktionary language editions as we could and provide it to the community while ensuring interoperability with other lexical data.

The DBnary extractor is written in java and is open-source (LGPL licensed, available at `http://dbnary.forge.imag.fr`). Anyone may contribute to this extraction effort by taking contact with the author.

2.2. Scope of the extracted data

The main goal of our efforts is not to extensively reflect Wiktionary content, but to create a lexical resource that is structured as a set of monolingual dictionaries + bilingual translation information. This way, the structure of extracted data follows the usual structure of Machine Readable Dictionaries (MRD). We originally extracted this data as we needed translations in many languages along with textual definitions of senses that we used to compute semantic similarity between senses (using an adapted Lesk measure) for the blexisma multilingual lexical disambiguation system [9].[2] Such data is already useful for several applications, but it is merely a starting point for a future multilingual lexical database.

The monolingual data is always extracted from its own Wiktionary language edition. For instance, the French lexical data is extracted from the French language edition.[3] Hence, we completely disregard the French data that may be found in other language editions.

We also filtered out some part of speeches in order to produce a result which is closer to existing monolingual dictionaries. For instance, in French, we disregard abstract entries that are prefixes, suffixes or flexions (e.g.: we do not extract data concerning *in-* or *-al* that are prefixes/suffixes and have a dedicated page in the French language edition).

2.3. Availability of the extracted data

DBnary data is provided using Creative Commons Attribution-ShareAlike 3.0 license (🅭🅯🄎). It may be downloaded from the DBnary website[4] as a set of turtle files (one per language).

As the Wiktionary language editions constantly evolve with entry modifications and additions, the DBnary dataset also evolves. Each time the Wikimedia foundation provides a new dump[5] of a Wiktionary language edition, DBnary data is extracted with the new dump and made available online. Older versions are kept and remain available for further reference. At the time of writing, the dumps are updated about once every ten days for each language.

DBnary data is also available as Linguistic Linked Open Data (LLOD). Hence, all DBnary IRIs are dereferencable and a SPARQL endpoint is available at `http://kaiko.getalp.org/sparql`. However, as the DBnary data changes almost everyday, the data that is available this way is not necessarily up to date.

2.4. Interlinking

Our work did focus only on the lexical data. Hence, we do not provide any reference to any ontology. Moreover, in this dataset, we only try to *extract* lexical data from Wiktionary, but we do not try to *enrich* it (yet). Hence, this dataset is not (at the time of printing) linked to other lexical linked data.

Also, any interlinking with DBnary data will require that we take into account the ever evolving nature of the dataset that changes every two days on average (as Wikimedia dumps are made available). Indeed, there is a chance that URIs of lexical senses may change between two versions as word senses may be re-ordered in the original Wiktionary data.[6] We do believe that such changes are rather unfrequent, but we still have to find out a way to cope with them.

As the DBnary data is now extracted regularly for almost one year, with about 25 different versions per language, diachronic studies may now be performed to evaluate the frequency of such changes. However, such studies are not trivial to implement as a change in a definition does not necessarily imply that the lexical sense has changed.

[2] We also used this dataset to build an UIMA component for word sense disambiguation, more details are available at `http://getalp.imag.fr/WSD`

[3] We use the term “French language edition” to refer to the data available on `http://fr.wiktionary.org`

[4] `http://kaiko.getalp.org/about-dbnary`

[5] dumps are available at `http://dumps.wikimedia.org/`.

[6] Strictly speaking, even URIs of lexical entries may change, but this is even more unfrequent.

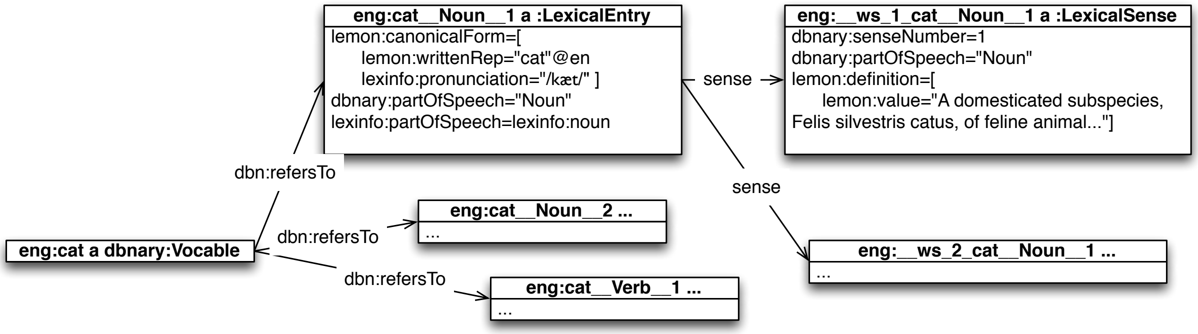
eng:cat__Noun__1 a :LexicalEntry
lemon:canonicalForm=[
lemon:writtenRep="cat"@en
lexinfo:pronunciation="/kæt/"]
dbnary:partOfSpeech="Noun"
lexinfo:partOfSpeech=lexinfo:noun
sense
eng:__ws_1_cat__Noun__1 a :LexicalSense
dbnary:senseNumber=1
dbnary:partOfSpeech="Noun"
lemon:definition=[
lemon:value="A domesticated subspecies, Felis silvestris catus, of feline animal..."]
dbn:refersTo
sense
eng:cat__Noun__2 ...
...
eng:cat a dbnary:Vocable
dbn:refersTo
eng:__ws_2_cat__Noun__1 ...
...
dbn:refersTo
eng:cat__Verb__1 ...
...

Fig. 1. An extract of the DBnary entry "cat" in English, showing the respective roles of Vocable, LexicalEntry and LexicalSense in the DBnary dataset.

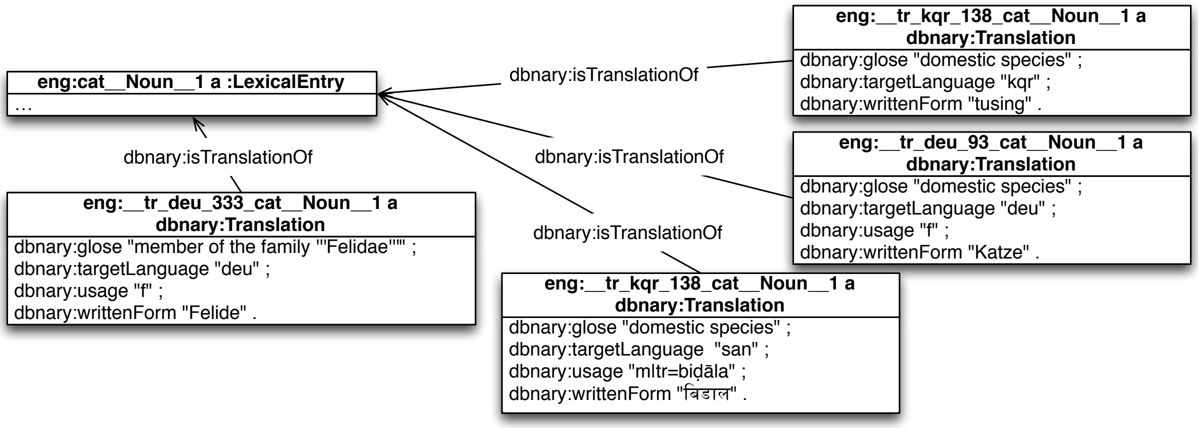
eng:__tr_kqr_138_cat__Noun__1 a dbnary:Translation
dbnary:glose "domestic species" ;
dbnary:targetLanguage "kqr" ;
dbnary:writtenForm "tusing" .
dbnary:isTranslationOf
eng:cat__Noun__1 a :LexicalEntry
…
dbnary:isTranslationOf
dbnary:isTranslationOf
eng:__tr_deu_93_cat__Noun__1 a dbnary:Translation
dbnary:glose "domestic species" ;
dbnary:targetLanguage "deu" ;
dbnary:usage "f" ;
dbnary:writtenForm "Katze" .
dbnary:isTranslationOf
eng:__tr_deu_333_cat__Noun__1 a dbnary:Translation
dbnary:glose "member of the family '''Felidae'''" ;
dbnary:targetLanguage "deu" ;
dbnary:usage "f" ;
dbnary:writtenForm "Felide" .
eng:__tr_kqr_138_cat__Noun__1 a dbnary:Translation
dbnary:glose "domestic species" ;
dbnary:targetLanguage "san" ;
dbnary:usage "mltr=biḍāla" ;
dbnary:writtenForm "बिडाल" .

Fig. 2. A subset of the translations related to the lexical entry eng:cat_Noun_1.

3. Extracted Data as a LEMON Lexical Resource

3.1. Using LEMON for legacy lexical data

LEMON model itself is not sufficient to represent lexical data that is currently available in classical monolingual and bilingual dictionaries. For instance, LEMON does not contain anything to represent translations between languages as it assumes that such a translation will be handled by the ontology description. Moreover, LEMON assumes that all data is well-formed and fully specified. As an example, synonymy relation is a property linking a LexicalSense to another LexicalSense. While this is correct to assume as a *principle*, this does not account for the huge amount of legacy data that is available in dictionaries and lexical databases.

As an example, in the English language edition, one may find a synonymy relation between *cat_n* and *bitch*. This relation links a *Lexical Entry* with a *Vocable*. Creating a corresponding lexico-semantic relation from *Lexical Sense* to *Lexical Sense* would imply: 1. detecting the correct sense of *cat_n* (here sense #4/15); 2. deciding which is the target lexical entry (here, *bitch* has 2 lexical entries, but only one is nominal); 3. deciding which lexical sense it is (#2/9). Such a process is error prone, hence, we decided to provide the data as it appears in Wiktionary and leave these decisions to further processing.

In order to cope with this legacy data, we extended the LEMON model by adding new classes and properties. However, when a piece of data is representable as a LEMON entity, then it is done so. Moreover, when possible, we did use the ISOcat registry [11] to identify standard elements in the lexical data.

3.2. DBnary extension to LEMON

The LEMON model has been extended to cope with legacy lexical data. Added classes and properties are:

Vocable: Several lexical entries may be contained in a single Wiktionary page. And most lexical relations are simply targeted to Wiktionary pages. Hence, we introduced the dbnary:Vocable[7] class to represent a Wiktionary page, as a reification of the set of Lexical entries it contains. This class is a subclass of lemon:LexicalEntry. Instances of this class are related to their lexical entries through the dbnary:refersTo property.

LexicalEntity: Lexical relations should usually link two *Lexical Senses*. However, most relations found in legacy lexical data are underspecified. Some relations link a *Lexical Sense* to a *Vocable* or to a *Lexical Entry*. Others even link two *Lexical Entries* together. In order to cope with such underspecified relations, we introduced the dbnary:LexicalEntity class which is the union of *Lexical Entry* and *Lexical Sense*.[8]

Nyms: Most Wiktionary language edition do provide “nym” relations (mainly synonym, antonym, hypernym, hyponym, meronym and holonym), which are almost always underspecified. Hence, DBnary introduces 6 new “nym” properties (in dbnary name space). These relations domains and ranges are *Lexical Entities*.

Translations: As there is no way to represent bilingual translation relation in LEMON, we introduced the dbnary:Translation class that collects translation information contained in Wiktionary. This class admits several properties:

- dbnary:isTranslationOf relates the translation to its source *lexical Entity* (i.e. either a Lexical Sense, or a Lexical Entry).
- dbnary:targetLanguage is an object property whose range is a `dcterms:LinguisticSystem`. All values are in the lexvo namespace.[9]
- dbnary:writtenForm gives the written form of the translation in the target language. Here, we decided not to relate to a vocable as some translations are not to be defined as lexical entries in the target language.
- dbnary:gloss is a string property that contains any available information used to denote the lexical sense of the source of the translation (e.g. in the English entry "cat", the German translation "Katze" appears in a box labelled with the *gloss* "domestic species", used to denote the fact that "Katze" is a translation of the *Lexical Sense* defined by "A domesticated subspecies, Felis silvestris catus, of feline animal...".[10]
- dbnary:usage is a string property that contains any available information concerning this translation object. It usually gives additional information on the target entry.

3.3. *Structure of the extracted lexical data*

Figure 1 illustrates the main elements characterizing a lexicon entry in the DBnary data. Each Wiktionary page is represented by a dbnary:Vocable element that refers to its corresponding lemon:LexicalEntry. Each lexical entry corresponds to one lemma, one etymology and one part of speech. Each lexical entry is related to its lemon:LexicalSense by the lemon:sense property. A lexical sense corresponds to one definition.

Each lexical entry is related to its canonical form and eventually to alternate forms (that are represented using lemon:LexicalForms). The part of speech is available through dbnary:partOfSpeech property that gives the part of speech as defined by the Wiktionary language edition, and through the isocat:partOfSpeech property that points to a standard ISOcat part of speech value.

Figure 2 illustrates the *DBnary* extension to LEMON that is used to represent the many translations that are available in Wiktionary. Each translation is represented as a dbnary:Translation instance that is associated to a *lexical entry* by the isTranslationOf property. The translation is given as a string through the writtenForm property, as some translations may not necessarily correspond to vocables in the target language (e.g. explanatory translations). The target language of the translation is given using ISO 639-3 3 letter language code [7]. When available, the gloss property value gives an indication concerning the lexical sense that is translated. In the current dataset, the translation is linked to

[7] The dbnary namespace used in this article resolves to `http://kaiko.getalp.org/dbnary#`

[8] Indeed, we could have simply defined the domain/range of lexical relations and translations as a owl:unionOf *Lexical Entry* and *Lexical Sense*.

[9] e.g. http://lexvo.org/id/iso639-3/deu represents the German language

[10] This means that, in an ideal situation, the translation should be linked to this *Lexical Sense*, instead of the *Lexical Entry*, however, such a linking may be error prone, hence, we decided to make such refinements afterwards.

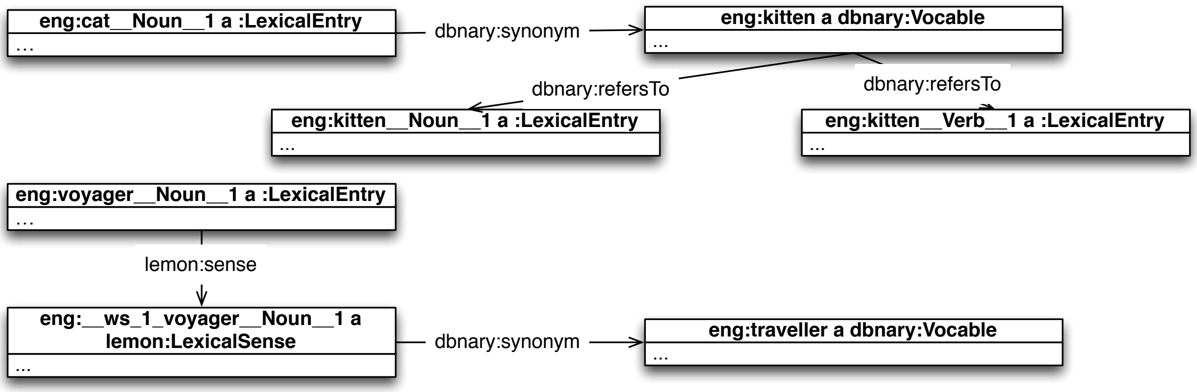
eng:cat__Noun__1 a :LexicalEntry
...
dbnary:synonym
eng:kitten a dbnary:Vocable
...
dbnary:refersTo
dbnary:refersTo
eng:kitten__Noun__1 a :LexicalEntry
...
eng:kitten__Verb__1 a :LexicalEntry
...
eng:voyager__Noun__1 a :LexicalEntry
...
lemon:sense
eng:__ws_1_voyager__Noun__1 a lemon:LexicalSense
...
dbnary:synonym
eng:traveller a dbnary:Vocable
...

Fig. 3. An example of a synonymy relation for the lexical entry eng:cat_Noun_1.

a usually ambiguous lexical entry and the *gloss* is kept for further attachment to the correct word sense.

Wiktionary contains many lexical relations (like synonymy, antonymy, etc.) that are represented using the above mentioned "nym" properties which relate a *lexical entity* with another *lexical entity*. In the current dataset, most of the property subjects are *lexical entries*. However, in case of monosemous lexical entries, the synonymy relation is attached to the unique lemon:LexicalSense. Figure 3 illustrates how DBnary encodes "nym" relations with examples of the dbnary:synonym property. In the upper part of the figure is an example of a dbnary:synonym property that is related to the ambiguous lexical entry eng:cat__Noun__1. The lower part shows that, in the monosemous lexical entry eng:voyager__Noun__1, the dbnary:synonym property is related to the unique lexical sense of the entry. The object of such properties are always dbnary:Vocable.

4. Size and quality of the involved data

All sizes indicated in this section reflect the state of the DBnary data at the time of writing (June 2013). These numbers are constantly evolving, as the original Wiktionary data is edited and as the extractor itself is improved.[11] Table 1 gives the number of resources available in DBnary for the 8 languages currently extracted.[12]

	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 1

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

Table 2 gives an overview of the number of lexico-semantic relations available in each language edition.

	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 2

Number of lexico-semantic relations. Languages are sorted according to their number of lexical entries.

Table 4 shows the number of translations available in each language edition, for the 8 currently extracted languages. It also gives the total number of translations and the number of the different target languages with translations. Not surprisingly, English language edition shows the most translations to more than 1000 different languages.

[11] Statistics concerning the latest extracts will be made available on the DBnary website.

[12] The currently extracted languages are (by increasing order of the ISO 639-3 language codes): German (deu), Greek (ell), English (eng), Finnish (fin), French (fra), Portuguese (por), Russian (rus),

Source/Target	deu	ell	eng	fin	fra	ita	por	rus	others	Total	# of target languages
eng	62501	23794	1	74938	57959	37467	30256	74837	764710	1126463	1143
fra	34608	7063	74687	7589	12	18806	17784	7783	296624	464956	952
deu	0	2675	81015	4947	67143	41485	8872	17354	248401	471892	355
rus	23056	3295	48559	3966	14776	12643	5567	0	206709	318571	490
ell	2242	2	10090	1056	8436	1470	1149	1315	29892	55652	246
fin	8046	918	30103	0	6700	3856	2196	7997	58912	118728	329
por	7000	2816	11284	4607	8720	7096	4	4396	179142	225065	695
ita	4619	506	17539	925	4461	75	1219	938	27514	57796	315

Table 4

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.

language	# of transl.
eng	5110 (.991)
fra	5799 (1.070)
deu	10287 (.992)
rus	8436 (248.117)
ell	2598 (.643)
fin	7245 (289.80)
por	17720 (.932)
ita	7855 (31.673)

Table 3

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

Asserting the quality of the extracts is difficult. We may compare the data with other Wiktionary extraction initiatives like [12] (contained in UBY [5]) or Wiktionary2RDF [6]. But this will only give informations regarding the common extracted languages.

In order to guide the extractors definition and maintenance, we compare the extracted data with the count of interwiki links[13] (available through a Wiktionary API). Table 3 gives an overview of such a comparison. Column *# of transl.* shows the number of extracted translations and the ratio of extracted translation vs interwiki links. As may be seen, the Greek extractor only gets a 0.64 ratio of extracted translations vs interwiki links. This is an indication of the fact that this extractor is in a very rough state. On the contrary, French extractor gets more translations than interwiki links, as the French edition does not generate links when the target language edition does not exist. Such a heuristic is not applicable for Finnish and Russian editions as they do not use interwiki links for their translations.

Finally, by comparing the evolution of extracted data over time, we are able to detect when an editorial decision is made in a language edition that leads to the loss of extracted data. This was the case when the French language edition decided to change the names of the macros used to represent a translation.

5. Conclusion and Perspectives

The current paper presents the DBnary dataset that is a LEMON based lexical network from different Wiktionary language editions. Such a work is interesting for many users that will be able to use the extracted data in their own NLP system. Moreover, as the extracted resource uses the Resource Description Framework (RDF) standard and the LEMON model, the extracted data is also directly usable for researchers on the Semantic Web, where it could be used to ease the ontology alignment systems when terms in different languages are used to describe ontologies of a domain.

This resource describes a significative number of entries for at least English and French languages which makes it comparable to Wordnet [2]. Moreover, it contains many translations with certain language pairs that makes it also comparable to the Open Multilingual Wordnet [1] (an aggregation of several existing multilingual Wordnets).

Our next objective is to better generalize the treatments of the current extractors, so that it will be easier to create and maintain extractors for other languages. We have recently introduced the extractor of Russian and Greek languages and are working on others. We welcome all initiatives aiming at the addition of new languages to this open-source tool.

[13] Interwiki links are links going from one Wiktionary language edition to another. This count is a rough estimate of the translations available in an entry.

We will also enhance the *DBnary* data by providing more lexico-semantic relations and translations linked on the lexical sense level.

6. Acknowledgements

The work presented in this paper was conducted in the Videosense project, funded by the French National Research Agency (ANR) under its CONTINT 2009 programme (grant ANR-09-CORD-026).

References

[1] F. Bond and R. Foster. Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362. The Association for Computer Linguistics, 2013. ISBN 978-1-937284-50-3.

[2] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press., Cambridge, MA, 1998.

[3] G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. Lexical markup framework (LMF) for NLP multilingual resources. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W06/W06-1001`.

[4] G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. Lexical Markup Framework (LMF). In *International Conference on Language Resources and Evaluation - LREC 2006*, Gênes/Italie, 2006. elra. URL `http://hal.inria.fr/inria-00121468/en/`. LIRICS.

[5] I. Gurevych, J. Eckle-Kohler, S. Hartmann, M. Matuschek, C. M. Meyer, and C. Wirth. Uby - a large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590, Apr 2012.

[6] S. Hellmann, J. Brekle, and S. Auer. Leveraging the crowdsourcing of lexical resources for bootstrapping a linguistic data cloud. In H. Takeda, Y. Qu, R. Mizoguchi, and Y. Kitamura, editors, *JIST*, volume 7774 of *Lecture Notes in Computer Science*, pages 191–206. Springer, 2012. ISBN 978-3-642-37995-6, 978-3-642-37996-3.

[7] ISO639-3. Codes for the representation of names of languages — part 3: Alpha-3 code for comprehensive coverage of languages. ISO 639-3:2007, 2007.

[8] J. McCrae, G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, May 2012. ISSN 1574-020X. . URL `http://dx.doi.org/10.1007/s10579-012-9182-3`.

[9] D. Schwab, J. Goulian, A. Tchechmedjiev, and H. Blanchon. Ant colony algorithm for the unsupervised word sense disambiguation of texts: Comparison and evaluation. In *25th International Conference on Computational Linguistics, COLING 2012*, pages 2389–2404, IIT Mumbai (India), December 2012.

[10] G. Sérasset. Dbnary: Wiktionary as a LMF based Multilingual RDF network. In *Language Resources and Evaluation Conference, LREC 2012*, Istanbul, 2012. URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/387_Paper.pdf`.

[11] M. Windhouwer and S. Wright. Linking to linguistic data categories in ISOcat. *Linked Data in Linguistics*, pages 99–107, 2012.

[12] T. Zesch, C. Müller, and I. Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *in Proceedings of LREC*, 2008. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.168.3605`.