Semantic Web 1 (2012) 1–5
IOS Press

Countering language attrition with PanLex and the Web of Data

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Patrick Westphal [a], Claus Stadler [a], Jonathan Pool [b]
[a] *University of Leipzig, {pwestphal, cstadler}@informatik.uni-leipzig.de*
[b] *The Long Now Foundation, San Francisco, pool@panlex.org*

Abstract. The world is losing some of its 7,000 languages. Hypothesizing that language attrition might subside if all languages were intertranslatable, the PanLex project supports panlingual lexical translation by integrating all known lexical translations. Semantic Web technologies can flexibly represent and reason with the content of its database and interlink it with linguistic and other resources and annotations. Conversely, PanLex, with its collection of translation links between more than a billion pairs of lexemes from more than 9,000 language varieties, can improve the coverage of the Linguistic Web of Data. We detail how we transformed the content of the PanLex database to RDF, established conformance with the lemon and GOLD data models, interlinked it with Lexvo and DBpedia, and published it as Linked Data and via SPARQL.

Keywords: Multilingual Linked Open Data, LLOD Cloud, PanLex, Lexical Resource, RDF, RDB2RDF, SPARQL, Sparqlify

1. Introduction

There are about 7,000 living languages[1], but language attrition has extinguished or threatened from 10% to over 75% of all languages in the last 60 years in various regions [10]. This attrition arguably imperils human biological knowledge and species diversity [9]. Hypothetically, panlingual intertranslatability would make all languages more useful and incentivize their preservation and revitalization.

The PanLex project is making all languages' lexicons intertranslatable. In some contexts (e.g., profiles, catalogs, tags, search, and web navigation), lexical translation can be most of the translation load. PanLex systematically integrates lexical translations, found in diverse sources, into a database for research, applications, and public use. The content can be interpreted as a graph linking lexemes in "is-a-translation-of" relations, and permitting automated inference to additional, unattested relations.

The Semantic Web initiative has led to the development of standards and technologies supporting a machine-readable and -interpretable Linked Data network, known as the *Web of Data*.[2] From these efforts the *Linked Open Data (LOD) cloud*[3] emerged. A growing community is leveraging Semantic Web technologies for linguistic knowledge, building a *Linguistic LOD (LLOD) cloud*.

Here we describe how we connected PanLex to this Linked Data network. In Section 2 we introduce the dataset, present a PanLex RDF vocabulary, and explain how we transformed the one into the other and established conformance with additional data models. Section 3 shows how we linked to other datasets of the LLOD cloud, and Section 4 is about the publication of the dataset. Usage scenarios are given in Section 5, and

[1] http://www-01.sil.org/iso639-3/iso-639-3.tab

[2] http://www.w3.org/standards/semanticweb/

[3] http://lod-cloud.net/

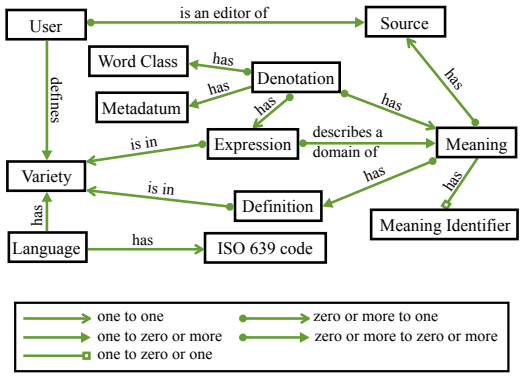

Fig. 1. The PanLex database schema

related work is discussed in Section 6. Finally, Section 7 concludes this paper.

2. Triplification of the Raw Data

In this section, we analyze the PanLex dataset, introduce our URI and vocabulary design, which resemble PanLex's conceptual model, summarize how we classified PanLex's instance data with additional data models and explain our transformation of the data to RDF.

2.1. Analysis of the Original Dataset

The PanLex *database* is created by editors who consult *information sources*[4], such as mono- and multilingual dictionaries, glossaries, standards, and thesauri. The data include single- and multi-word expressions, corresponding meanings assigned to them, and related information. PanLex data constitute editors' *interpretations* of sources' *assertions* that two or more *expressions* share a *meaning*[5]. The most important entities and relations of PanLex's conceptual model are depicted in Figure 1.

- The *source* entity is the authority to which an editor attributes assertions about lexical translations.
- *Expressions* are lexical entities, each uniquely identified with a text, i.e. a string of (Unicode) characters, and a variety of a language. Expressions resemble lemmas or dictionary-entry headwords, but differ from them in at least two ways. (1) Homographs, such as the verb "hide" (conceal) and the noun "hide" (animal skin) in English, are treated as a single expression in PanLex. (2) Multiword expressions, such as "fall in love", traditionally found in an entry headed by one of their words, such as "fall" or "love", are treated as independent expressions in PanLex.
- *Languages* in PanLex are identified using ISO 639-3[6] individual and macrolanguage codes, ISO 639-2[7] collective codes, and ISO 639-5[8] codes.
- *Language varieties* are collections of expressions. Each has a unique identifier: a language code and a distinguishing integer. For example, six dialects of Ahtna are identified as "aht-000" through "aht-005". These labels are, themselves, treated as a (controlled) language variety, whose expressions (i.e. the labels) are translated into natural languages and other controlled languages (such as the IETF standard BCP 47[9]).
- *Meanings* are entities assigned to expressions, thereby identifying expressions as translations or synonyms. For example, a source's translation of the German expression "klingen" into English as "ring, sound, seem" can be interpreted as (1) the assignment of a single meaning to all four expressions or (2) the assignment of two or three meanings to "klingen" and of one of those to each of the English expressions. Meanings are source-specific. The identification and consolidation of equivalent meanings of distinct sources is a research topic, not a database feature. Meanings can have properties of three types. (1) *Definitions* are descriptions of a meaning, consisting of text strings annotated as being in particular language varieties. (2) *Domains* are expressions (e.g., "medicine") that characterize a meaning, but do not express it. (3) *Meaning identifiers* are strings acting as references to identifiers in a source.
- *Denotations* are assignments of meanings to expressions. A denotation may have one or more *word classes* (a closed set based on OLIF, the Open Lexicon Interchange Format) and/or *metadata* (arbitrary strings paired as keys and values).
- *Users* may define sources, attribute data to them, and define language varieties.

[4] http://panlex.org/tech/plrefs.shtml
[5] http://panlex.org/tech/doc/design/panlex-db-design.pdf
[6] http://www.sil.org/iso639-3/codes.asp
[7] http://www.loc.gov/standards/iso639-2/
[8] http://www.loc.gov/standards/iso639-5/
[9] http://tools.ietf.org/html/bcp47

License	Count	License	Count	License	Count
copyright	1343	*LGPL*	9	*PD*	149
CC	387	*MIT*	32	*other*	106
FDL	24	*PanLex*	7	*unknown*	1958
GPL	172	*request*	5		

Table 1

Number of sources using a certain license

Entity	Instances
Denotations	54,278,860
Meanings	20,773,371
Expressions	19,790,453
Definitions	2,747,892
Language Varieties identified	9,310
Language Varieties with data	9,239
Languages	7,843
Sources being consulted	4,190
Sources already consulted	1,453
Users	23

Table 2

Number of instances of main entities in the PanLex database

– Among the properties of sources are *licenses*. Some of the license categories are *public domain*, *request* (author invites inquiries), *GNU Free Documentation License (FDL)*, and *PanLex Use Permission* (specific permission for use in PanLex). The distribution of licenses is shown in Table 1.

Table 2 gives entity counts as of January 2014.

2.2. The PanLex Vocabulary

The entities and relations described above are the base for the PanLex RDF vocabulary. In general, all PanLex RDF resources reside in the namespace `http://ld.panlex.org/plx/`, abbreviated with `plx`. An example of the resulting ontology is depicted in Figure 2 and summarized as follows. Unless otherwise noted, the URIs of instances of PanLex classes follow the pattern *plx:{className}/{id}*, where {className} is spelled in lower camel case and the {id} is the primary key of the corresponding database table.

– Expressions are modeled as instances of the class `plx:Expression`. Their original and degraded textual representations become the values of the properties `rdfs:label` and `plx:degradedText`, respectively. Their corresponding language variety is stated using `plx:languageVariety`.
– For language and language varieties the classes `plx:Language` and `plx:LanguageVariety` are introduced. *ISO 639-1* and *ISO 639-3 codes* become instances of the classes `plx:Iso639-1Code` and `plx:Iso639-3Code`.

Class	Properties
`plx:Source`	`plx:registrationDate`, `rdfs:label`, `dc:title`, `dc:creator`, `plx:license`, `dc:date`, `plx:quality`, `foaf:homepage`, `dc:publisher`, `dbpedia-owl:isbn`
`plx:Language`	`plx:iso639-3Code`, `plx:iso639-1Code`
`plx:LanguageVariety`	`plx:languageVarietyOf`, `rdfs:label`
`plx:Iso639-1Code`	
`plx:Iso639-3Code`	
`plx:Expression`	`plx:languageVariety`, `plx:degradedText`, `rdfs:label`
`plx:Meaning`	`plx:approver`, `plx:identifier`, `plx:meaningDefinition`
`plx:Definition`	`plx:languageVariety`, `rdfs:label`
`plx:Denotation`	`plx:denotationMeaning`, `plx:denotationExpression`, `plx:wordClass`
`plx:WordClass`	`rdfs:label`
`plx:License`	`rdfs:label`

Table 3

Classes and properties used in the PanLex RDF vocabulary. Note that all `rdf:type` properties are omitted for brevity.

– The RDF analog of the PanLex *meaning* is the `plx:Meaning`. Entities of this class may have an identifier assigned with the `plx:identifier` property pointing to an `xsd:string` literal. Meanings may also have *definitions*, entities of the `plx:Definition` class, giving a textual representation (`rdfs:label`) in a certain language variety (`plx:languageVariety`).
– Meanings and expressions are linked via *denotations*. These are entities of the `plx:Denotation` class pointing to meanings and expressions via the properties `plx:denotationMeaning` and `plx:denotationExpression`. Denotations may also have a word class assigned to them. This can be achieved with the denotation's `plx:wordClass` property pointing to a `plx:WordClass` entity.
– All sources share the `plx:Source` class. The characteristics of a source are described using mainly triples with literal objects. These are for example `dc:title` to assign the title of a source, `dc:creator` to give an `xsd:string` containing the author's name. At present, we support the different license categories recognized in the database by creating resources of the `plx:License` class.

2.3. Vocabulary Reuse

The PanLex vocabulary is based on PanLex's conceptual schema and enables all of PanLex's data to

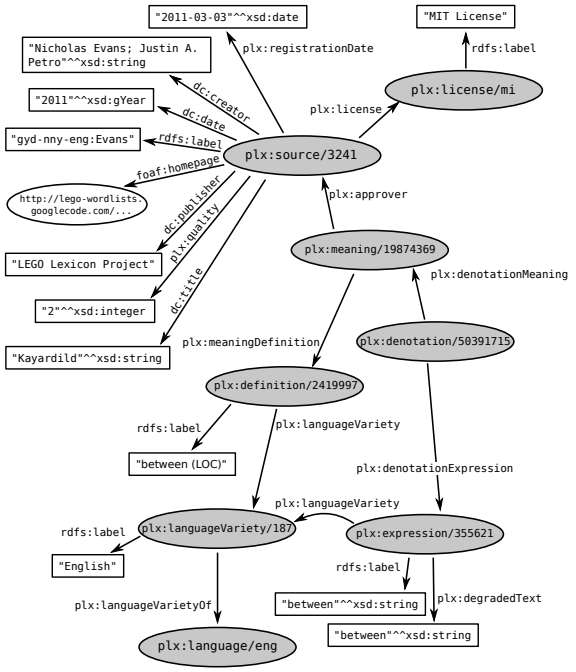

Fig. 2. Example of the PanLex RDF vocabulary showing one meaning of the expression 'between' and the corresponding source and definition

Panlex	lemon	GOLD
`plx:Denotation`	–	`gold:LinguisticSign`
`plx:Meaning`	`lemon:LexicalSense`	`gold:SemanticUnit`
`plx:Expression`	`lemon:LexicalEntry`	`gold:FormUnit`

Table 4

Classes considered to be similar across the re-used vocabulary

be directly exposed as RDF. Additionally, we also re-use existing vocuabularies, namely the *Lexicon Model for Ontologies* (lemon) [7] as well as the *General Ontology for Linguistic Description* (GOLD) [4]. Since these models differ from the PanLex one, we follow an incremental approach of aligning the PanLex data with them. Table 4 shows PanLex classes with their current counterparts in lemon and GOLD respectively. The parts implemented in our RDF conversion are displayed in Figure 3.

2.4. RDF Transformation Workflow

Since new sources are added to the PanLex database on almost a daily basis and because of its current size (~18 GB), the recurrent conversion of the database to capture changes in it is impractical. As the PanLex data already reside in a relational database, the use

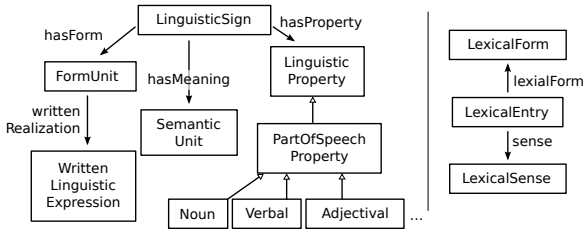

Fig. 3. Parts of the GOLD (left) and lemon model (right) re-used in PanLex (URI prefixes are omitted for brevity)

```
1  Create View i1 As Construct {
2      ?lang a plx:Language, <http://schema.org/Language> ;
3            plx:iso639-3Code ?iso3 .
4      ?iso3 a plx:Iso639-3Code ;
5            owl:sameAs ?lexvo3 .  }
6    With
7        ?lang = uri(plx:language, '/', ?iso3)
8        ?iso3 = uri(plx:iso639-3, '/', ?iso3)
9        ?lexvo3 = uri('http://lexvo.org/id/iso639-3/', ?iso3)
10   From [[SELECT iso3 FROM i1]]
```

Fig. 4. An excerpt of an SML view definition for PanLex's languages. This example also demonstrates how "is-a" relations to schema.org and links to Lexvo are established.

of a virtual RDB2RDF[10] mapping solution is a natural choice. The *Sparqlify system*[11] offers, besides an efficient query rewriting engine, also a very easy-to-use mapping language, called *Sparqlification Mapping Language* (SML). Essentially, these mappings consist of three clauses: The *From* clause specifies the logical SQL table (i.e. table, view, or query) to be used in the SML view. The *With* clause binds a set of SPARQL variables to expressions that yield RDF terms from relational columns. Finally, the *Construct* clause holds a set of triple patterns. Figure 4 shows an example of an SML view for the languages in PanLex: From each row of the table *i1* three resources are created based on the *iso3* column and bound to the variable names *?lang*, *?iso3* and *?lexvo3*. Resources for *?lang* become typed as a *Language* in the PanLex and the *schema.org* namespace. This view-based approach makes it easy to perform future revisions of RDF mapping, such as adding support for new vocabularies.

3. Linking

The SML view in (Figure 4) establishes the interlinking of languages in PanLex with Lexvo [3]. Here

[10] `http://www.w3.org/2001/sw/wiki/RDB2RDF`
[11] `https://github.com/AKSW/Sparqlify`

Language	Links	Language	Links
English	1,415,241	Catalan	27,779
German	224,146	Korean	24,912
French	187,364	Turkish	22,258
Italian	147,485	Bulgarian	19,431
Spanish	117,056	Hungarian	18,203
Portuguese	112,266	Slovene	11,981
Polish	110,974	Greek	1,112
Russian	68,040		
Czech	28,767	Total	2,537,015

Table 5
Number of DBpedia links per language

we outline the interlinking with DBpedia [5], where we were interested in creating *valid* and *dereferenceable* links. Therefore, we iterated the *titles* datasets[12], which map (non-localized) DBpedia URIs to their page titles in the respective language. For each language version we normalized the labels by applying Unicode NFKD[13] normalization and removal of punctuation characters. Each DBpedia resource was then mapped to the PanLex expression that was equal to the resource's normalized label in the respective language. Table 5 summarizes the number of links obtained.

In total, about 2.5 million links were obtained for approx. 20 million expressions. This relatively low coverage can be attributed to frequently appearing multi-word expressions that do not match the DBpedia titles well, and the fact that in this work we yet only considered DBpedia datasets for mainstream languages, whereas PanLex focuses on low-density ones.

4. Publishing

With our RDF conversion work, we complement existing APIs[14] with Linked Data, powered by Pubby[15], and two SPARQL endpoints[16,17], ran by Sparqlify and Virtuoso. An overview is shown in Figure 5. The SPARQL browser *SNORQL*[18] can be accessed by replacing *sparql* with *snorql* in the respective links. Our SML views and the interlinking code are hosted on GitHub[19]. The created linksets are hosted in the PanLex database and are published together with the other data using Sparqlify. Finally, we offer downloads tagged with timestamps of their creation[20].

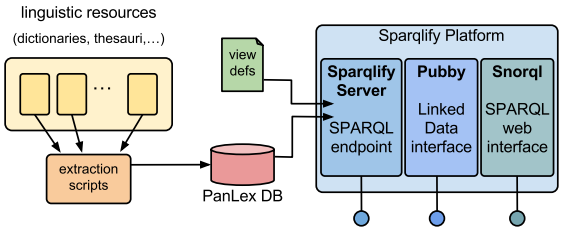

Fig. 5. PanLex architecture

5. Dataset Benefits and Usage Scenarios

There are general benefits of using Semantic Web technologies, such as the potential for simplified data integration due to RDF and vocabulary reuse, the possibility of enriching data based on interlinking, drawing advantage from reasoning and the exploration of the data through the use of generic Semantic Web tools. Moreover, some applications, like the TeraDict translation lookup service[21], can now be realized using SPARQL queries and so easily integrated in other applications. Due to space considerations, we refer the reader to the PanLex Linked Data landing page[22], where a collection of SPARQL queries is maintained. Also, since PanLex covers a niche of providing linguistic data for non-mainstream languages, investigation of its fitness for use in cross language information retrieval, as well as annotation projects, like DBpedia Spotlight [8] seems worthwhile.

6. Related Work

PanLex is a project whose editors integrate information discovered from many lexical resources. The extraction of information from linguistic sources, and techniques for automatically inferring translations, are relevant work discussed in [6]. An important related initiative is the Global Wordnet Association (GWA)[23], which offers a platform for sharing wordnets and, among other goals, uniformly representing wordnets of different languages and establishing a universal in-

[12] http://wiki.dbpedia.org/Downloads38
[13] http://unicode.org/reports/tr15/
[14] http://panlex.org/try/
[15] http://wifo5-03.informatik.uni-mannheim.de/pubby/
[16] http://ld.panlex.org/vsparql
[17] http://ld.panlex.org/sparql
[18] https://github.com/kurtjx/SNORQL
[19] https://github.com/AKSW/PanLex-2-RDF
[20] http://ld.panlex.org/downloads/releases/
[21] http://panlex.org/teradict/?lg=eng
[22] http://ld.panlex.org
[23] http://www.globalwordnet.org/

dex of meaning. Wordnets are usually focused on the definition of *synsets* and relations between them in a single language; GWA is helping to transform these single-language synonym silos into a virtual multilingual translation resource. PanLex is approximating that, in a different way, by integrating data from numerous wordnets along with translingual sources into a single graph. In the Semantic Web context, several standard or quasi-standard vocabularies and ontologies have been developed with the rise of the Linguistic LOD movement. Examples include the *Ontologies of Linguistic Annotation* (OLiA) [1] for modeling lexicon and machine-readable dictionaries, *POWLA* for modeling linguistic corpora [2] and the *Natural Language Processing Interchange Format* (NIF)[24].

7. Conclusions and Future Work

In this dataset description we detailed the PanLex database and its conversion to RDF. Based on our URI and vocabulary design, we created appropriate view definitions for the *Sparqlify* system, which carries out the actual RDF transformation. Furthermore, we interlinked the languages in PanLex with Lexvo, and created about 2.5 million links to DBpedia for expressions in 16 languages. With the integration of lemon and GOLD we also support data access via external linguistic ontologies.

We intend to address some limitations in the future: The relations among PanLex's information sources, if treated as distinct datasets, could be modeled with the VoID vocabulary[25]. The source entity should be refactored to reference users and information sources as distinct entities. Metadata attached to PanLex denotations are currently limited to arbitrary pairs of strings, but this sacrifices discovery possibilities when the metadata describe facts that can again be expressed with PanLex expressions. Finally, new collaborations between PanLex and related fields (e.g. as language identification, language geolocation, lemmatization, transliteration, localization, etc.) are promising areas for development.

[24] `http://nlp2rdf.org/nif-1-0`
[25] `http://rdfs.org/ns/void`

References

[1] C. Chiarcos. Grounding an ontology of linguistic annotations in the data category registry. *Workshop on Language Resource and Language Technology Standards (LR&LTS 2010)*, 2010.

[2] C. Chiarcos. Powla: Modeling linguistic corpora in owl/dl. In *ESWC*, volume 7295 of *LNCS*, pages 225–239. Springer, 2012.

[3] G. de Melo and G. Weikum. Language as a foundation of the Semantic Web. In C. Bizer and A. Joshi, editors, *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC 2008)*, volume 401 of *CEUR WS*, Karlsruhe, Germany, 2008. CEUR.

[4] S. Farrar and D. T. Langendoen. A linguistic ontology for the semantic web. *GLOT International*, 7(3):97–100, 2003.

[5] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.

[6] Mausam, S. Soderland, O. Etzioni, D. S. Weld, K. Reiter, M. Skinner, M. Sammer, and J. Bilmes. Panlingual lexical translation via probabilistic inference. *Artif. Intell.*, 174(9-10):619–637, 2010.

[7] J. McCrae, D. Spohr, and P. Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In *ESWC*, volume 6643 of *LNCS*, pages 245–259. Springer, 2011.

[8] P. N. Mendes, M. Jakob, A. Garcia-Silva, and C. Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.

[9] D. Nettle and S. Romaine. *The Extinction of the World's Languages*. Oxford University Press, 2000.

[10] G. F. Simons and M. P. Lewis. The world's languages in crisis: A 20-year update. In *Responses to Language Endangerment: In honor of Mickey Noonan*, Studies in Language Companion Series, pages 3–20. John Benjamins, 2013.