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] *Long Now Foundation, San Francisco, pool@panlex.org*

Abstract. At present, there are approximately 7,000 living languages in the world. However, some experts claim that the process of globalization may eventually lead to the world losing this linguistic diversity. The vision of the PanLex project is to help save these languages, especially low-density ones, by allowing them to be intertranslatable and thus to be a part of the Information Age. Semantic Web technologies can support achieving this goal, for reasons such as their capabilities of flexibly representing, interlinking and reasoning with data, in our case particularly linguistic resources and annotations. Conversely, an RDF version of PanLex makes a significant contribution towards improving the coverage of the Linguistic Web of Data, as to the best of our knowledge there exists no large scale Linked Data data set for panlingual translation of non-mainstream languages. In this dataset description paper we detail how we transformed the data of the PanLex project 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: Multi-lingual Linked Open Data, PanLex, Lexcial Resource, RDF, RDB2RDF, Sparqlify

1. Introduction

At present, there are about 7,000 living languages in the world[1]. Nonetheless, some experts claim that processes such as nation-state consolidation and globalization are producing language attrition so rapidly that up to 90% of all languages alive today will be extinct within a century [7]. Theorists of biolinguistic diversity argue that the loss of language diversity, the loss of human biological knowledge, and the loss of species diversity are mutually supportive and thus that language preservation and revitalization are essential to the preservation of biological diversity [6]. The vision of the PanLex project is to help save these thousands of languages, especially those low-density ones that are threatened by extinction, by supporting their use in global communication. This requires panlingual translation: translation from any language of the world into any other. One of the crucial components of panlingual translation of discourses is panlingual lexical translation. PanLex is designed to support that component by making use of several thousand sources and artificial intelligence approaches.

Apart from using PanLex's data for inferring lexical translations, a further direction of work is to generate added value by enriching this data with knowledge from other (linguistic) sources. Grounding on the idea of a Semantic Web, a complete stack of technologies has been devised and standardized by the *World Wide Web Consortium*[2] supporting the definition of a machine readable and interpretable Linked Data network, which has led to emergence of the *Web of Data*. Currently there is a growing community working on lever-

[1] See `http://www.sil.org/iso639-3/download.asp` for a list of registered languages

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

aging these Semantic Web technologies for linguistic knowledge and thereby building a Web of Linguistic Data, also known as the *Linguistic Linked Open Data (LLOD) cloud*.

Our steps to connect PanLex to this Linked Data network are as follows. In Section 2 we introduce the PanLex dataset, present our PanLex RDF vocabulary, explain how we transformed the one into the other, and how we established conformance with additional data models. Section 3 is about how we linked to other datasets of the LLOD cloud, whereas Section 4 is about the actual dataset publishing. Usage scenarios are given in Section 5. In Section 6 we discuss related work, and finally, in Section 7 we conclude our approach and give some hints to future work.

2. Triplification of the Raw Data

In this section, we first provide an analysis of the PanLex dataset. Subsequently, we introduce our URI and vocabulary design which closely resembles PanLex's original conceptual model. Afterwards, we briefly describe how we classified PanLex's instance data using additional data models. Finally, we explain the steps taken to transform the data to RDF.

2.1. Analysis of the Original Dataset

At the core of the PanLex *project*, there is the PanLex *database* which is created from the imports of thousands of lexical resources, such as mono- and multilingual dictionaries, glossaries, standards, and thesauri. The concrete list of used sources is available online[3]. The data derived from these sources comprise single- and multi-word expressions and meanings assigned to them. Conceptually, the PanLex database thus "represents *assertions* about the *meanings* of *expressions*"[4]. As of now, the database contains about 20 million meanings and 19 million expressions extracted from about 2,000 sources. The most important entities and relations of PanLex's conceptual model are depicted in Figure 1 and are explained in more the detail in the following.

- The starting point of the data acquisition is the *approver* entity: An editor processes the content of a certain mono- or multilingual source as mentioned above and adds it to the PanLex dataset. This combination of a user and a source is referred to as an approver.
- *Expressions* are lemmas, i.e. dictionary entries. For example, *"go"* is a valid expression, whereas *"went"* is not. Expressions are always given in a language variety and can only be given once per language variety.
- *Languages* in PanLex are identified using ISO 639-3[5] individual and macrolanguage codes, ISO 639-2[6] collective codes and ISO 639-5[7] codes.
- *Language varieties* allow one to make more fine grained distinctions within a language. Their codes are composed of the language code combined with a PanLex specific identifier. For example, "eng" is the ISO 639-3 code for the English language. Panlex defines various varieties of it, including "English" (eng-000), "Simple English" (eng-001) and "British English" (eng-005). Their labels, when possible, are autonyms, written in the native writing system. So, in contrast to mechanisms like IETF BCP 47[8] there is no need for a transcription.
- *Meanings* in PanLex are entities of which each represents a unique possible sense of an expression. Meanings are assigned by editors based on their interpretation of expressions. Usually this assignment is done on a per-source basis so that identical meanings across multiple sources are not resolved. This means that if there is e.g. a translation of the fruit *"apple"* in an English to

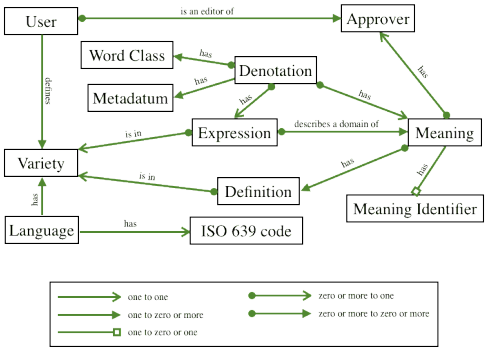

Fig. 1. The PanLex database schema

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

License	Count	License	Count	License	Count
unknown	1886	*PD*	123	*LGPL*	9
copyright	1212	*other*	104	*request*	5
CC	335	*MIT*	32		
GPL	172	*FDL*	24		

Table 1
Number of approvers using a certain license

Entity	Instances	Entity	Instances
Denotations	50,803,243	Language Varieties	7,248
Meanings	20,023,427	Approvers	3,905
Expressions	18,580,594	Licenses	10
Definitions	2,522,605	Users	7
Languages	7,839		

Table 2
Number of entities in the PanLex database

German dictionary and another translation from English to French, these do not necessarily result in a single meaning entity linking to all involved languages. Instead, there could be *"apple"* and *"Apfel"* sharing one meaning entity and *"apple"* and *"pomme"* sharing another.

- *Definitions* are optional descriptions of a meaning. They are given in a certain language variety. A description of the verb *browse* for example could be *"move or surf through various files on a computer, the Internet, etc."*, marked as a definition in the "English" language variety.
- *Denotations* are entities that relate expressions to meanings and may optionally carry annotations in form of sets of key value pairs. For instance, an English expression *pig*, when referring to *police officer*, could be annotated with *pragmatics=vulgar*. Furthermore, denotations can be tagged with part-of-speech tags, such as word classes, selected from a closed list based on the Open Lexicon Interchange Format (OLIF) standard[9].
- *Users* have editorial privileges over the language varieties and the approvers that they define.
- *Licenses* are also considered by the PanLex project. At present there are ten different license categories an approver can be annotated with. They are *public domain*, *Creative Commons (CC)*, *request* (meaning that one has to ask the author of the resource), *GNU General Public License (GPL)*, *GNU Lesser General Public License (LGPL)*, *GNU Free Documentation License (FDL)*, *MIT License*, *copyright* (stating that there is a certain copyright holder), *other* and *unknown*. The license distribution is shown in Table 1.

An overview of the number of instances per entity in the current PanLex database is given in Table 2.

Note that *approver* combines a user with an *information source*, however the *information source* is not modeled as a distinct entity. Also, the information of whether or not two meanings with different approvers are the same is not being captured.

[9] `http://www.olif.net/`

2.2. The PanLex vocabulary

The entities and relations of the schema described in the previous section serve as the base for the development of 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 normalized 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`.
- 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`).
- Following the semantics of the PanLex database, 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

Class	Properties
plx:Approver	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.

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 achieved with the denotation's plx:wordClass property pointing to a plx:WordClass entity.

– All approvers share the plx:Approver class. The characteristics of an approver 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.

Since the PanLex project compiled its database by extracting data from different sources, the licenses of these sources were also considered. At present, we support the different license categories given 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 be directly exposed as RDF. Additionally, we also reuse existing vocuabularies, namely the *Lexicon Model for Ontologies* (lemon) [5] as well as the *General Ontology for Linguistic Description* (GOLD) [3]. Since these models differ from the PanLex one to some extent, 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.

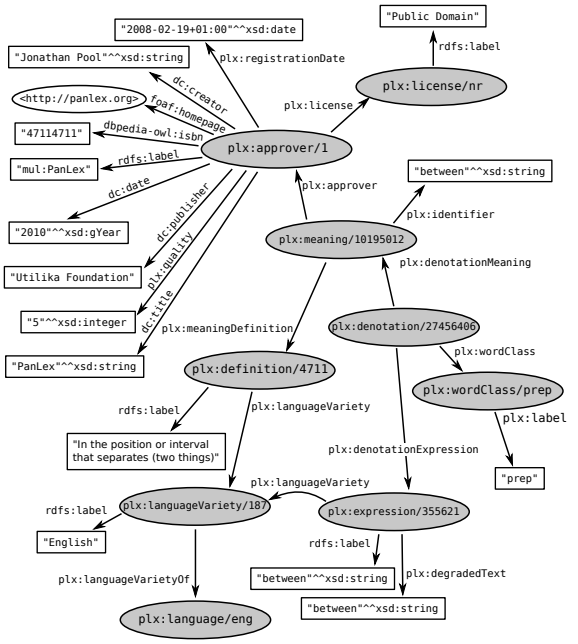

Fig. 2. Overview of the PanLex RDF vocabulary

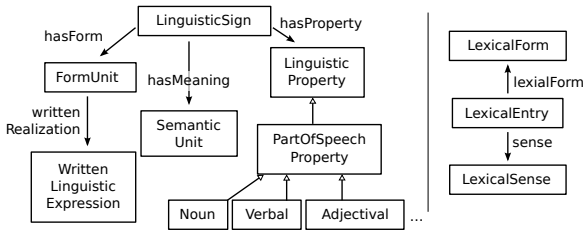

Fig. 3. Parts of the GOLD (left) and lemon model (right) re-used in PanLex (URI prefixes are omitted for brevity)

2.4. Transformation workflow

Since new sources are added to the PanLex database on almost a daily basis and because of its current size (~18 GB) an approach of recurrent conversions is impractical. As the PanLex data already resides in a relational database, the use of a virtual RDB2RDF[10]

[10] http://www.w3.org/2001/sw/wiki/RDB2RDF

```
Create View i1 As Construct {
    ?lang a plx:Language, <http://schema.org/Language> ;
          plx:iso639-3Code ?iso3 .
    ?iso3 a plx:Iso639-3Code ;
          owl:sameAs ?lexvo3 .  }
  With
      ?lang = uri(plx:language, '/', ?iso3)
      ?iso3 = uri(plx:iso639-3, '/', ?iso3)
      ?lexvo3 = uri('http://lexvo.org/id/iso639-3/', ?iso3)
  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.

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 also demonstrates that changing the vocabulary or adding support for new ones does not require an extract transform load (ETL) process, and can therefore be done with little effort.

3. Linking

The SML view in the previous section (Figure 4) already established the interlinking of the PanLex languages with Lexvo. In this section we outline the interlinking with DBpedia. For DBpedia, we were interested in creating *valid* and thus *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.

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

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], powered by Sparqlify and Virtuoso, respectively. 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 the Sparqlify RDB2RDF tool. Finally, we offer downloads tagged with timestamps of their creation[20].

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 pos-

[11] https://github.com/AKSW/Sparqlify
[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/

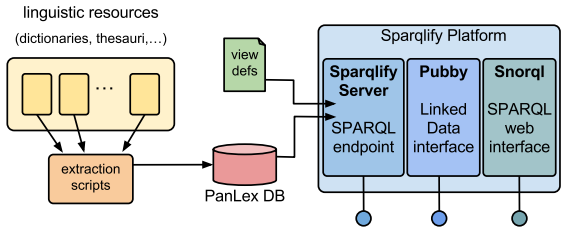

Fig. 5. PanLex architecture

sibility 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[23], seems worthwhile.

6. Related Work

PanLex is an integration project of many existing lexical resources. The extraction of information from linguistic sources, and techniques for automatically inferring translations, are relevant work discussed in [4]. An important initiative is the Global Wordnet Association[24], which offers a platform for sharing wordnets and defines several goals. These include setting forth standards for uniformly representing wordnets of different languages and establishing a universal index of meaning. Wordnets are usually focused on the definition of *synsets* and relations between them in a single language, whereas PanLex's focus is on capturing lexical translations between languages. Hence, these efforts are complementary. In the Semantic Web context, several (quasi-)standard vocabularies and ontologies have been developed with the rise of the *Linguistic Linked Open Data* 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)[25].

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 the data access via external linguistic ontologies.

There exist some shortcomings which we intend to overcome in the future: Modeling information sources and users as distinct entities would enable one to unequivocally relate meanings and denotations to them, which in turn would allow for a more fine grained attribution of qualities and relevances. This could be beneficial for translation approaches. Another important aspect is, that each of PanLex's information sources can be seen as a dataset on its own, and thus relations between them could be modeled with the VoID vocabulary[26].

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] S. Farrar and D. T. Langendoen. A linguistic ontology for the semantic web. *GLOT International*, 7(3):97–100, 2003.
[4] 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.
[5] 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.
[6] D. Nettle and S. Romaine. *The Extinction of the World's Languages*. Oxford University Press, 2000.
[7] A. C. Woodbury. What is an endangered language? `http://www.linguisticsociety.org/content/what-endangered-language`, 2006.

[21] `http://panlex.org/teradict/?lg=eng`
[22] `http://ld.panlex.org`
[23] `http://spotlight.dbpedia.org`
[24] `http://www.globalwordnet.org/`
[25] `http://nlp2rdf.org/nif-1-0`
[26] `http://rdfs.org/ns/void`