# 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. For this reason, PanLex gathers and integrates information from thousands of linguistic resources, such as monolingual dictionaries, bilingual dictionaries, multilingual dictionaries, glossaries, standards and thesauri. In this dataset description paper we detail how we transformed this data to RDF, 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. It documents the known lexical translations (translations of lexemes) among all languages.

Although a list of all the translations of all words into all other languages would be large (amounting to trillions of translations), there are more serious obstacles. The vast majority of these translations are unknown. Those that are known are not universally agreed on. And lexical polysemy and ambiguity make the question "What is the translation of lexeme X into language Y?" underspecified for practical purposes. So the PanLex project is designed to permit the application of artificial intelligence techniques to infer and select translations.

Currently there is a growing community trying to combine linguistic knowledge with Semantic Web technologies and thereby build a Web of Linguistic Data, also known as the *Linguistic Linked Open Data (LLOD) cloud*. The PanLex project is still in its data-

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

acquisition phase, but it has provided a few web services[2] and APIs[3] to access its data. Since the project shares the idea of open access, we have undertaken the project of making PanLex a part of the LLOD cloud.

In Section 2 we introduce the PanLex dataset, present our PanLex RDF vocabulary and explain how we transformed the one into the other. 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 original PanLex dataset. Subsequently, we introduce our URI and vocabulary design and we explain the steps taken to publish the data as 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 monolingual dictionaries, bilingual dictionaries, multilingual dictionaries, glossaries, standards, and thesauri. The concrete list of used sources is available online[4]. The data derived from these sources comprises single- and multi-word expressions and meanings assigned to them. Conceptually, the PanLex database thus “represents *assertions* about the *meanings* of *expressions*”[5]. 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 their relations are depicted in Figure 1 and are explained in more the detail in the following.

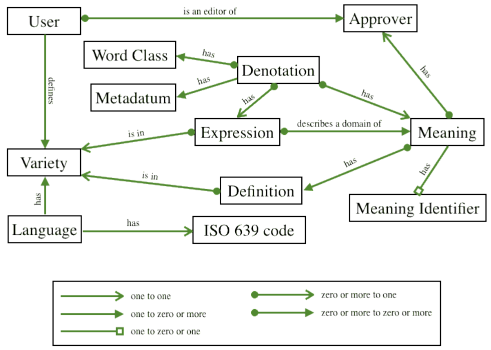

Fig. 1. The PanLex database schema

- 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[6] individual and macrolanguage codes, ISO 639-2[7] collective codes and ISO 639-5[8] 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[9] 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 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

[2] http://panlex.org/try/
[3] http://panlex.org/tech/doc/api/
[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

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

Table 1
Number of entities in the PanLex database

*"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[10]. For example, *fall* can be a verb or a noun for autumn. Homonyms are those expressions that are connected to multiple meanings.
- *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*, *request* (meaning that one has to ask the author of the resource), *GNU General Public License*, *GNU Lesser General Public License*, *GNU Free Documentation License*, *MIT License*, *copyright* (stating that there is a certain copyright holder), *other* and *unknown*.

An overview of the number of instances per entity in the current PanLex database is given in Table 1.

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.

## 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 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.

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

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

Table 2

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

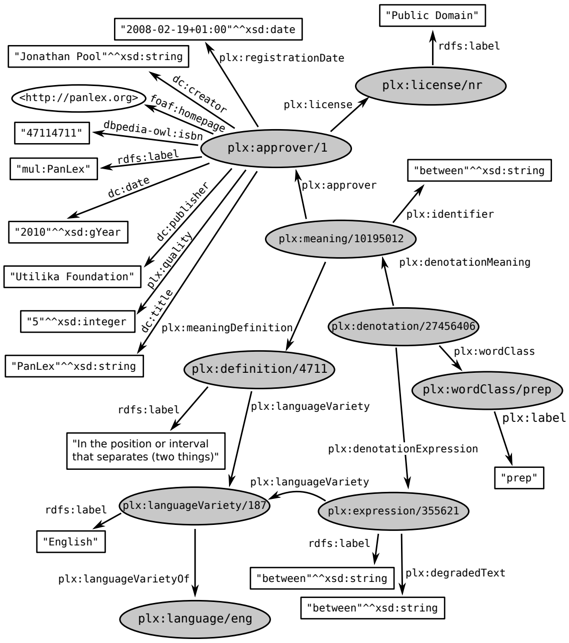

Fig. 2. Overview of the PanLex RDF vocabulary

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. Transformation workflow

New sources are added to PanLex on almost a daily basis. The size of the PanLex database (including indexes) is currently approximately 18 GB, which is big enough to make recurrent RDF conversion cumbersome: The RDF file of a full conversion of a database is much larger than the database itself. Using conventional hardware, it takes impractically long to convert all the data. This makes testing and debugging and fixing issues in the modeling or data conversion of the data very time-consuming. As the PanLex data already resides in a relational database, the use of a virtual RDB2RDF[11] mapping solution is a natural choice. The *Sparqlify system*[12] 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:

- *From:* Specifies the logical SQL table (i.e. table, view or query) underlying the SML view.
- *With:* Defines SPARQL variables by means of expressions over relational columns that yield RDF terms.
- *Construct:* The template (i.e. the set of triple patterns) to be generated in this view, based on the SPARQL variable definitions.

Figure 3 shows an example of an SML view for the languages in PanLex: From each row of the *i1* table three resources are created from 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 3) already established the interlinking of the PanLex languages with Lexvo. In this section we outline the interlinking with DBpedia.

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

```
Create View il As
  Construct {
    ?lang a plx:Language ;
          a <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
    il
```

Fig. 3. 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.

| Language | 639-1 | 639-3 | Links |
|---|---|---|---|
| English | en | eng | 1,415,241 |
| German | de | deu | 224,146 |
| French | fr | fra | 187,364 |
| Italian | it | ita | 147,485 |
| Spanish | sp | spa | 117,056 |
| Portuguese | pt | por | 112,266 |
| Polish | pl | pol | 110,974 |
| Russian | ru | rus | 68,040 |
| Czech | cs | ces | 28,767 |
| Catalan | ca | cat | 27,779 |
| Korean | ko | kor | 24,912 |
| Turkish | tr | tur | 22,258 |
| Bulgarian | bg | bul | 19,431 |
| Hungarian | hu | hun | 18,203 |
| Slovene | sl | slv | 11,981 |
| Greek | el | ell | 1,112 |
| Total | | | 2,537,015 |

Table 3
Number of DBpedia links per language

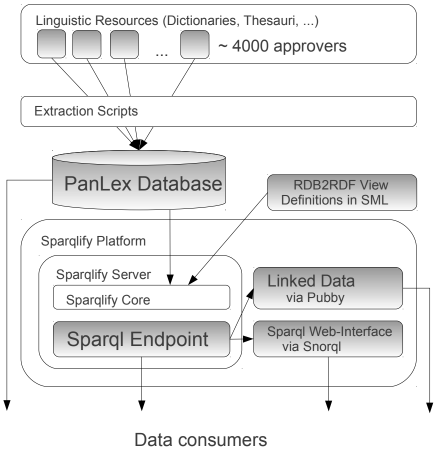

Fig. 4. PanLex architecture

For DBpedia we were interested in creating *valid* and thus *dereferenceable* links. Therefore, we iterated the *titles* datasets[13], 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[14] 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 3 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

As stated in Section 1, the PanLex project already provides several interfaces for data access. With our RDF conversion work, we complement these interfaces with Linked Data and an SPARQL endpoint[15]. An overview is shown in  Figure 4. Our SML views and the code for interlinking are hosted on GitHub[16]. The created linksets are hosted in the PanLex database and published together with the other data using Sparqlify. Downloads and additional information are available on the Panlex-RDF project page[17].

## 5. Dataset Benefits and Usage Scenarios

There are general benefits of RDF conversions, such as the paradigm shift towards thinking about what one intends to express with the data, making this meaning explicit using ontologies, enabling of interlinking, easing data integration tasks and uniform data access via Linked Data and SPARQL.

[13] http://wiki.dbpedia.org/Downloads38
[14] http://unicode.org/reports/tr15/
[15] http://ld.panlex.org/sparql and http://ld.panlex.org/snorql
[16] https://github.com/AKSW/PanLex-2-RDF
[17] http://ld.panlex.org

With the RDF version of PanLex, a large lexical resource participates in the Web of Data, making it available to a broader community. Users can easily explore the PanLex data using the services provided by SNORQL[18] and Pubby[19]. This also increases the chance of development of new interesting mashups. For example, the TeraDict translation service[20] could now be easily realized using simple SPARQL queries. Another future usage scenario is to link PanLex to Wortschatz[21] with the aim of improving the recall of word-by-word translations by considering their co-occurrences.

## 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*[22], 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. At the level of the Semantic Web, there is currently a trend in publishing linguistic resources as Linked Data and via SPARQL. These efforts are referred to as the *Linguistic Linked Open Data Cloud* (LLOD). Several (quasi-)standard ontologies have been developed for covering different aspects of linguistic resources. Examples include the *Ontologies of Linguistic Annotation* (OLiA) [1], the *Lexicon Model for Ontologies* (lemon) [5] as well as *General Ontology for Linguistic Description* (GOLD) [3] for modeling lexicon and machine-readable dictionaries, *POWLA* for modeling linguistic corpora[2] and the *Natural Language Processing Interchange Format* (NIF)[23].

## 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 virtual RDB2RDF solution *Sparqlify* 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. Although not discussed explicitly due to space limitations, we also made first steps towards enhancing the RDB2RDF view definitions as to facilitate the integration of the PanLex data with the lemon and GOLD data models.

In conclusion, we see this work as a valuable contribution to the vision of countering language attrition by making a vast amount of lexical resources available as Linked Open Data.

We identified some weaknesses in the original database design, which we intend to overcome in the future: The original dataset currently does not cleanly model information sources and approvers as distinct entities. Also, while it is possible to determine each meaning's and each denotation's approver, an approver can have multiple users with editorial privileges, so it is not possible to track each contributed, modified, and deleted meaning and denotation by editor. Retaining this information seems beneficial for translation approaches, as this enables for example the attribution of qualities and relevances to specific editors.

[18] https://github.com/kurtjx/SNORQL

[19] http://wifo5-03.informatik.uni-mannheim.de/pubby/

[20] http://panlex.org/teradict/?lg=eng

[21] http://wortschatz.uni-leipzig.de/

[22] http://www.globalwordnet.org/

[23] http://nlp2rdf.org/nif-1-0

## 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.