Historical Dictionaries as Digital Editions and Connected Graphs: the Example of Le Petit Larousse Illustré

Anas Fahad Khan, Hervé Bohbot, Francesca Frontini, Mohamed Khemakhem, Laurent Romary

Historical Dictionaries as Digital Editions and Connected Graphs: the Example of Le Petit Larousse Illustré

Anas Fahad Khan (anasfkhan81@gmail.com), Istituto di Linguistica Computazionale "A. Zampolli", Italy
Hervé Bohbot (herve.bohbot@cnrs.fr), Praxiling UMR 5267 CNRS — Université Paul-Valéry — Montpellier 3
Francesca Frontini (francesca.frontini@univ-montp3.fr), Praxiling UMR 5267 CNRS — Université Paul-Valéry — Montpellier 3
Mohamed Khemakhem (mohamed.khemakhem@inria.fr), Inria ALMAnaCH, Paris; Centre Marc Bloch, Berlin; Paris Diderot University, Paris
Laurent Romary (laurent.romary@inria.fr), Inria ALMAnaCH, Paris; Centre Marc Bloch, Berlin; BBAW - Berlin-Brandenburgische Akademie der Wissenschaften, Berlin

Introduction: Publishing Print Dictionaries as TEI-XML and RDF

The trend towards the retrodigization of print dictionaries -- especially those considered to have a historical as well as a scientific importance -- has been given a new impetus in recent years thanks to improvements in optical character recognition as well as to developments in the creation and promotion of standards and technologies (Khemakhem et al., 2017; Khemakhem et al., 2018) which enable machine actionable versions of such texts to be published and shared much more easily than before. Traditionally the Text Encoding Initiative (TEI) (Budin et al., 2012) has been the favoured standard for the encoding of retrodigitised print dictionaries, but there is now starting to be a move towards the publication of electronic editions of print dictionaries as Linked Data[1] (LD).

The prior popularity of TEI for this task reflects the dual nature that digital editions of print dictionaries can often have: namely, both as representations of dictionaries as texts, that is as printed artifacts that follow particular typographical conventions and have specific styles of page layout, etc; and at the same time, as computational resources that serve to make the lexicographic and, more broadly speaking linguistic, information contained in the original texts more accessible for querying and further machine processing.

On the other hand one of the greatest strengths of Linked Data (LD) lies in its core emphasis on interoperability between datasets, heterogeneous as to subject area and type, through the use of a common semantic model, that of the Resource Description Framework (RDF), as well as the use of common vocabularies across datasets. Not only does LD encourage the mutual enrichment of individual datasets by facilitating the creation of extensive links between them, it also gives modellers access to a whole ecosystem of Semantic Web technologies and standards, including several out-of-the-box tools for manipulating and visualising structured data. In addition, formal languages such as RDF, RDFS and OWL, which make up part of the Semantic Web stack, allow us to specify and to elaborate on the semantics of the classes and properties used to structure dictionary data.

Modelling a dictionary using RDF requires us to represent the information contained within it as a series of subject-predicate-object statements, which taken together describe a formal graph structure. As a consequence it is much less successful -- which in this case means much less verbose -- than TEI at representing things like the layout and formatting of the original text, or properties relating to the status of the text as a series of tokens, as well as in encoding certain kinds of deeply nested structures. This might suggest that RDF is better

[1] A recent project, ELEXIS (**https://elex.is/**), which aims to create a platform for accessing and working with dictionary data, and linking senses together across dictionaries in different languages, works with both TEI-XML and RDF versions of editions of dictionaries.

suited to describing the more abstract linguistic content of retrodigitised dictionaries, e.g., describing grammatical and semantic information for each entry (along with dictionary metadata), and for embellishing this content through links to other salient datasets and vocabularies. However as we will see, there often exist aspects of print dictionaries that although strictly speaking they concern *how* information is presented in the text -- and relate, for example, to the dictionary as a historical artifact -- and not the lexical information contained in the text itself, are still worth explicitly encoding in RDF. This is both because RDF allows us to make this extra-lexical information more accessible and usable and because it helps to ensure that each RDF version of a dictionary is a more self-contained resource.

Modelling Le Petit Larousse illustré

In order to flesh out some of the issues outlined in the preceding section, especially from the point of view of elucidating the potential benefits of using RDF as a means of publishing historic dictionaries, we will focus on a particular case study which concerns the French national project, Nénufar[2] (Bohbot et al., 2018). One of the main goals of Nénufar is to make different consecutive editions of the illustrated French language dictionary *Le Petit Larousse illustré* (PLI), published during the first half of the 20th century, publically available both via an online interface as well as as downloadable TEI-XML digital editions and as a linked data dataset. So far all of the editions of PLI published between 1906 and 1924 have been digitised, encoded and made searchable. The native digitisation format is TEI, although the encoding adheres as much as possible to the TEI-Lex0 format (Bański et al., 2017; Romary and Tasovac, 2018); the conversion of entries into RDF is currently ongoing.

The PLI frequently includes encyclopedic information pertaining to its lexical entries along with the more typical kinds of lexicographic data which means that it possesses a strong socio-cultural and historic interest in addition to its significance as a legacy lexicographic resource; indeed, to some extent it can be considered a hybrid resource, dictionary-encyclopedic. Take, for instance, the entry for the word *aviation* from three different editions of PLI, from 1906, 1912 and 1915[3]. It's interesting to track how changes in the entry reflect contemporary discoveries that were taking place in the field of aeronautics at the time.

Here the three successive versions of the entry each contains slightly different encyclopedic glosses. Note also that in the last of the three versions of the entry a reference appears to the lexical entry for the word *aéroplane*. In this case two aspects of the same entry have changed over the course of different editions: the textual content of the encyclopedic gloss and the appearance of a new reference to another entry.

As regards the linked data version of the PLI, we made the decision to include as much of the encyclopedic and bibliographic information from the original dataset as possible and to model the evolution of entries across editions since, as we mentioned in the last section, this helps to ensure that the dataset is relatively self-contained -- and prevents users of the RDF version of the PLI having to constantly refer back to the TEI encoding[4], something which would go against the universalising spirit of the Semantic Web -- and also because some of this information is well suited to being represented in RDF. By explicitly modelling the editions in RDF using bibliographic and temporal vocabularies and associating each with a specific year, we can query the data for date-specific information.

[2] http://nenufar.huma-num.fr
[3] See the different versions of this entry at http://nenufar.huma-num.fr/?article=3807
[4] Go tho the "resources" section of each Nénufar entry to inspect the TEI xml.

Furthermore we used the well-known Ontolex-Lemon vocabulary (McCrae et al., 2017) for publishing lexicon-ontologies as linked data as the basis of our encoding in addition to making extensive use of other standards and vocabularies such as the *lexinfo* registry, SKOS and DEO[5]. However these did not always provide the properties and classes we needed and so in several cases we decided to create our own. Note that although at the time of writing the Ontolex-Lemon lexicography module is still in the process of being finalised for publication (Bosque-Gil et al., 2016) we have tried our best to make sure that we don't define any new classes or properties similar to those likely to be in the former. In order to model links between entries we utilised the already existing class *Reference* from the DEO vocabulary, and defined a new class *DictionaryGloss* to represent any written explanation of a lexical element in a dictionary.

To reiterate, our intention was to model the changes between PLI editions, and indeed in some cases between reprints of the same edition. We decided to model all the separate editions in one graph, since individual changes between entries in different editions usually weren't comprehensive enough to warrant a separate graph per edition, and in addition there were also differences between reprintings of the same edition and we wanted to avoid creating too many different graphs. To this end we created a class *PLIDictionary* to represent *separatePLI* editions, along with the object properties, *appearsIn*, *firstAppearsIn* and *lastAppearsIn* to allow us to associate elements with different editions. In our RDF encoding of the PLI, then, we model changes within entries by creating an individual entry component, whether this is a form, sense or gloss, etc, for every change and associating it with one or more PLI editions using *appearsIn*, *firstAppearsIn* and *lastAppearsIn*.

We will explain the strategy which was followed using the example of the RDF encoding of *aviation.* In Figure 1 we represent the entry for *aviation* and its relationships with its senses. Note that we have added information to the entry regarding its first appearance in the PLI by associating it with the individual *1906_001,* which represents the 1906 edition of the work, using the property *firstAppearsIn.* Each of the two senses of *aviation* has a gloss apiece neither of which changes over different editions in the example.

In Figure 2 we show the three notes associated with the *aviation* entry each of which has been modelled as an individual of the class *DictionaryGloss* and each of which is associated with a different edition of the dictionary. We are still looking into the best, read most efficient, way of adding information about what is contained in each edition. The simplest way would be to link each lexical element to each of the editions in which it appears, but this would obviously lead to an explosion of triples. Our provisional solution is to focus on adding version information to the elements that change between versions and linking them together using the Dublin Core relation *isVersionOf*.

[5] http://www.sparontologies.net/ontologies/deo

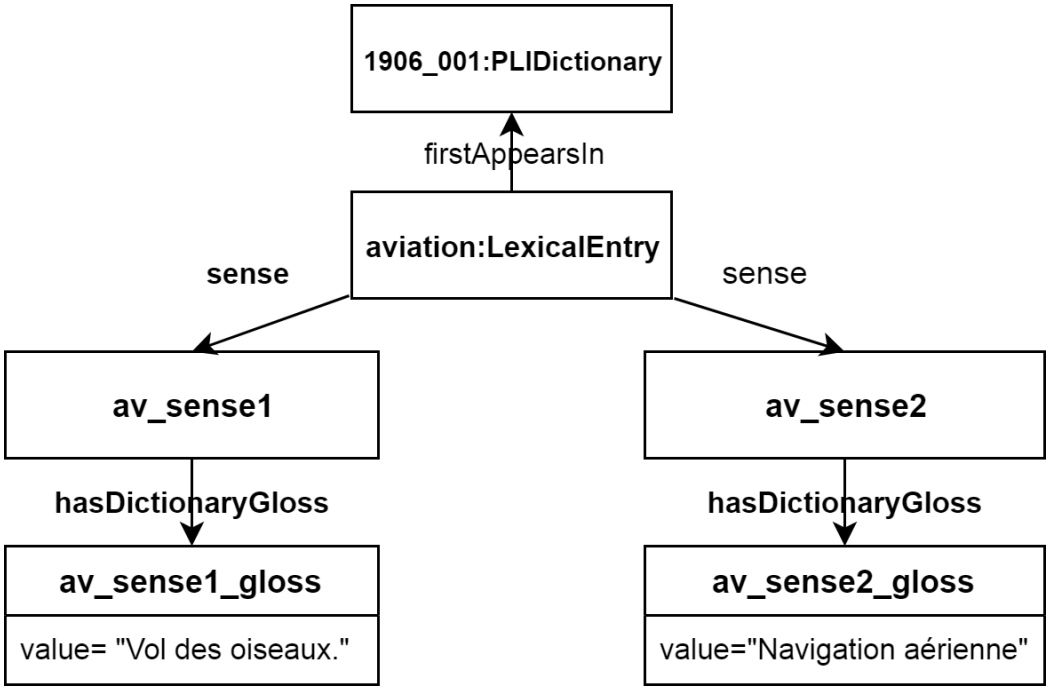

Figure 1. The PLI entry for aviation and its senses and their gloss.

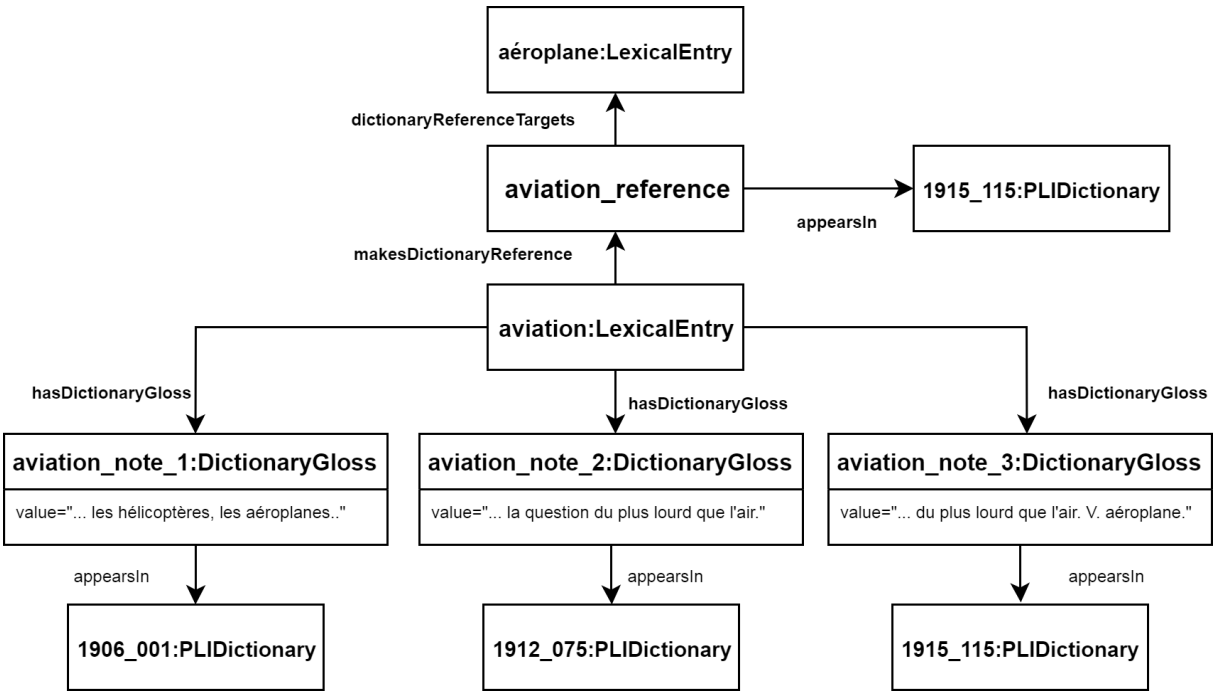

Figure 2. The PLI entry for aviation and the various versions of the entry note associated with it.

Future Work

At the time of writing we are carrying out the conversion of the dataset into RDF using the approach outlined above. By the middle of 2019 we plan to make whole of the dataset available both via a SPARQL endpoint and downloadable both in RDF and TEI-XML formats.

In the final version we also plan to add links to external conceptual/ontological resources (such as DBpedia and WordNets) using the Ontolex *reference* property.

References

Bański, P., Bowers, J. and Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. *ELex2017*.

Bohbot, H., Frontini, F., Luxardo, G., Khemakhem, M. and Romary, L. (2018). Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré. *GLOBALEX 2018 - Globalex Workshop at LREC2018*. Miyazaki, Japan, pp. 1–6 https://hal.archives-ouvertes.fr/hal-01728328 (accessed 21 April 2018).

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E. and Aguado-de-Cea, G. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries case. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf (accessed 23 April 2017).

Budin, G., Majewski, S. and Mörth, K. (2012). Creating Lexical Resources in TEI P5. A Schema for Multi-purpose Digital Dictionaries. *Journal of the Text Encoding Initiative*(Issue 3) doi:10.4000/jtei.522. http://journals.openedition.org/jtei/522 (accessed 23 April 2019).

Khan, F., Frontini, F., Boschetti, F. and Monachini, M. (2016). Converting the Liddell Scott Greek-English Lexicon into Linked Open Data using lemon. *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, pp. 593–96 http://dh2016.adho.org/abstracts/236.

Khemakhem, M., Foppiano, L. and Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. *Electronic Lexicography, ELex 2017*. Leiden, Netherlands https://hal.archives-ouvertes.fr/hal-01508868 (accessed 28 February 2018).

Khemakhem, M., Herold, A. and Romary, L. (2018). Enhancing Usability for Automatically Structuring Digitised Dictionaries. *GLOBALEX Workshop at LREC 2018*. Miyazaki, Japan https://hal.archives-ouvertes.fr/hal-01708137 (accessed 21 November 2018).

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. *ELex2017*.

Romary, L. and Tasovac, T. (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. *JADH 2018 'Leveraging Open Data'*. Tokyo, pp. 274–75 https://tei2018.dhii.asia/AbstractsBook_TEI_0907.pdf.