A Diachronic Digital Edition of the Petit Larousse illustré
Herve Bohbot, Alexandre Fauchere, Francesca Frontini, Agata Jackiewicz, Giancarlo Luxardo, Agnès Steuckardt, Mohamed Khemakhem, Laurent Romary

To cite this version:

HAL Id: hal-01873805
https://hal.archives-ouvertes.fr/hal-01873805
Submitted on 13 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
The Petit Larousse illustré is the only French dictionary to be updated every year, since it was first published in 1905 (coined 1906).

Handy and affordable, with each edition selling several hundreds of thousands copies (Pruvost, 2002), one Petit Larousse may virtually be found in every French household.

Editions published prior to 1948 are public domain according to French laws, however, no open full text resource is currently available.

WEB site

Targeting both scientific (linguists, historians, lexicographers, social scientists) and general public. Simple and advanced research criteria, user friendly interface, responsive design.

Public release in July 2018 with 1906 to 1924 editions, then new data will be regularly added.

TEI XML

Reference digital edition format. For long-term preservation and expert querying.

Users: digital humanists, computational linguists.

The use of TEI-LEX0 format (Banski et al., 2017) will allow for a predictable structure and for the exploitation of ad hoc digitisation technologies such as GROBID-Dictionaries (Khemakhem et al., 2017)

Ontolex-Lemon RDF

Natural language processing applications and linking to other resources.

Users: language technologists, computer scientists.

Presenting the Nénufar project:
A Diachronic Digital Edition of the Petit Larousse illustré

Hervé Bobbot, Alexandre Faucher, Francesca Frontini, Agata Jackiewicz, Giancarlo Luxardo, Agnès Steuckardt

Praxiling UMR 5267, CNRS – Univ. Paul-Valéry Montpellier 3, Montpellier, France

Mohamed Khemakhem, Laurent Romary

Inria – ALMAAnaCH, Paris – Centre Marc Bloch, Berlin – Université Paris Diderot, Paris

Berlin-Brandenburgische Akademie der Wissenschaften, Berlin

Nénufar project: aiming to a diachronic corpus of 1906 to 1948 editions

Nénufar project name stands for Nouvelle édition numérique de fac-similés de référence. This name was chosen after a typically French polemic between supporters and opponents of some orthographic modifications of the French language. Amongst others, it was proposed that the waterlily (Nymphaea L.), should be spelled nénufar instead of nénuphar, which was perceived by some conservative people as an aggression towards the language integrity… However, it was actually the original, forgotten spelling, as it could be seen in not so old and common dictionaries like Petit Larousse (see below).

There was clearly a gap to be filled, in order to bring to light recent evolutions of the language. Amongst others, it was proposed that the waterlily (Nymphaea L.), was actually nénufar. This name was chosen after a typically French polemic between supporters and opponents of some orthographic modifications of the French language.

The use of TEI-LEX0 format (Banski et al., 2017) will allow for a predictable structure and for the exploitation of ad hoc digitisation technologies such as GROBID-Dictionaries (Khemakhem et al., 2017)

Ontolex-Lemon RDF

Natural language processing applications and linking to other resources.

Users: language technologists, computer scientists.

A new resource for studying recent history of language, culture and techniques

An example with the entry AVIATION between 1906 and 1912 editions

WEB site

Targeting both scientific (linguists, historians, lexicographers, social scientists) and general public. Simple and advanced research criteria, user friendly interface, responsive design.

Public release in July 2018 with 1906 to 1924 editions, then new data will be regularly added.

TEI XML

Reference digital edition format. For long-term preservation and expert querying.

Users: digital humanists, computational linguists.

The use of TEI-LEX0 format (Banski et al., 2017) will allow for a predictable structure and for the exploitation of ad hoc digitisation technologies such as GROBID-Dictionaries (Khemakhem et al., 2017)

Ontolex-Lemon RDF

Natural language processing applications and linking to other resources.

Users: language technologists, computer scientists.

References:


Nénufar project is headed by laboratoire Praxiling at Paul-Valéry University of Montpellier in collaboration with Inria. It is supported by the Délégation Générale à la Langue Française et aux Langues de France (an agency of French Culture and Communication Ministry) and the Huma-Num consortium CORLI and CAHIER.