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.
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 – ALMAnaCH, Paris – Centre Marc Bloch, Berlin – Université Paris Diderot, Paris
Berlin-Brandenburgische Akademie der Wissenschaften, Berlin

- 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.

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

WEB site
Targeting both scientific (linguists, historians, lexicographers, social scientists) and general public.
Simple and advanced research interface, 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.

Language section (1906):
- 1066 pages in-8°
- 44876 entries
- 2532 illustrations
- ca 1.06 million words

Some word datations:
1908: autobus, griviste (stricker)
1911: antimilitarisme
1912: purecultivate, technicien
1914: tract
1915: copyright
1916: scientisme, syndicalisme
1918: boche (pej. for German)
1920: alerter
1921: bolchevik, bolschivisme
1922: mazout, pogrom, soviet

References:

Info: projet-nenufar@univ-montp3.fr
Data will be published under CC-BY license.

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 consortia CORLI and CAHIER.