Connecting a French Dictionary from the Beginning of the 20th Century to Wikidata

Pierre Nugues

Pierre.Nugues@cs.lth.se

LREC 2022, Marseilles, June 21-23, 2022

Petit Larousse illustré (1905)

- The *Petit Larousse illustré* is a one-volume dictionary of French, first published in 1905.
- Definitions of words, things, and people complemented with encyclopedic developments
- Numerous illustrations
- Very popular and far-reaching cultural influence.
- Updated yearly

Source of image: Wikimedia

Dictionary Division

The dictionary consists of three main parts:

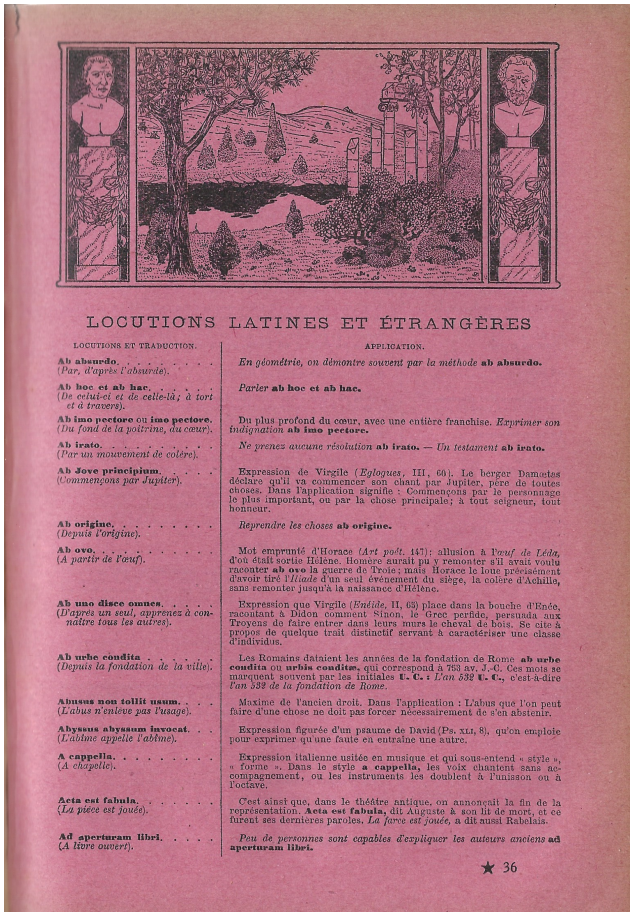
1. The first one, *langue française*, is a dictionary of French with some encyclopedic content. It is restricted to the common nouns, verbs, adjectives, adverbs, and grammatical words (1066 pages);

LANGUE FRANÇAISE

4th edition from 1925

Dictionary Division

2. The second part, *locutions*, contains quotes, mostly from classical Latin authors. It is much smaller than the two other parts (32 pages); and



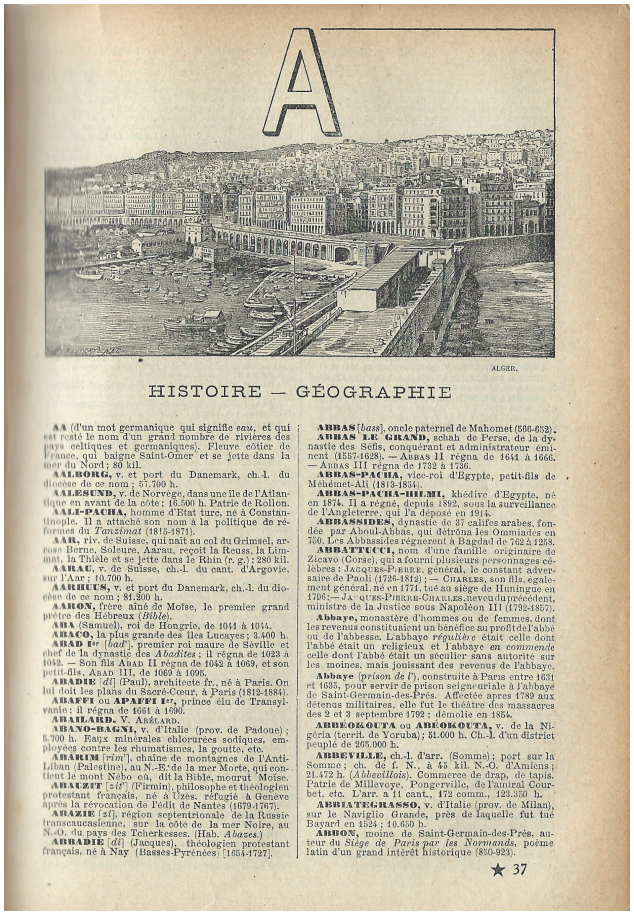
LOCUTIONS LATINES ET ÉTRANGÈRES

LOCUTIONS ET TRADUCTION.	APPLICATION.
Ab absurdo. (*Par, d'après l'absurde*).	*En géométrie, on démontre souvent par la méthode* **ab absurdo.**
Ab hoc et ab hac. (*De celui-ci et de celle-là; à tort et à travers*).	*Parler* **ab hoc et ab hac.**
Ab imo pectore ou **imo pectore.** (*Du fond de la poitrine, du cœur*).	Du plus profond du cœur, avec une entière franchise. *Exprimer son indignation* **ab imo pectore.**
Ab irato. (*Par un mouvement de colère*).	*Ne prenez aucune résolution* **ab irato.** — *Un testament* **ab irato.**
Ab Jove principium. (*Commençons par Jupiter*).	Expression de Virgile (*Eglogues*, III, 60). Le berger Damœtas déclare qu'il va commencer son chant par Jupiter, père de toutes choses. Dans l'application signifie : Commençons par le personnage le plus important, ou par la chose principale; à tout seigneur, tout honneur.
Ab origine. (*Depuis l'origine*).	*Reprendre les choses* **ab origine.**
Ab ovo. (*A partir de l'œuf*).	Mot emprunté d'Horace (*Art poét.* 147) : allusion à *l'œuf de Léda*, d'où était sortie Hélène. Homère aurait pu y remonter s'il avait voulu raconter **ab ovo** la guerre de Troie ; mais Horace le loue précisément d'avoir tiré l'*Iliade* d'un seul événement du siège, la colère d'Achille, sans remonter jusqu'à la naissance d'Hélène.
Ab uno disce omnes. (*D'après un seul, apprenez à connaître tous les autres*).	Expression que Virgile (*Enéide*, II, 65) place dans la bouche d'Enée, racontant à Didon comment Sinon, le Grec perfide, persuada aux Troyens de faire entrer dans leurs murs le cheval de bois. Se cite à propos de quelque trait distinctif servant à caractériser une classe d'individus.
Ab urbe condita (*Depuis la fondation de la ville*).	Les Romains dataient les années de la fondation de Rome **ab urbe condita** ou **urbis conditæ**, qui correspond à 753 av. J.-C. Ces mots se marquent souvent par les initiales **U. C.** : *L'an 532* **U. C.**, c'est-à-dire *l'an 532 de la fondation de Rome.*
Abusus non tollit usum. (*L'abus n'enlève pas l'usage*).	Maxime de l'ancien droit. Dans l'application : L'abus que l'on peut faire d'une chose ne doit pas forcer nécessairement de s'en abstenir.
Abyssus abyssum invocat. (*L'abîme appelle l'abîme*).	Expression figurée d'un psaume de David (Ps. XLII, 8), qu'on emploie pour exprimer qu'une faute en entraîne une autre.
A cappella. (*A chapelle*).	Expression italienne usitée en musique et qui sous-entend « style », « forme ». Dans le style **a cappella**, les voix chantent sans accompagnement, ou les instruments les doublent à l'unisson ou à l'octave.
Acta est fabula. (*La pièce est jouée*).	C'est ainsi que, dans le théâtre antique, on annonçait la fin de la représentation. **Acta est fabula,** dit Auguste à son lit de mort, et ce furent ses dernières paroles. *La farce est jouée*, a dit aussi Rabelais.
Ad aperturam libri. (*A livre ouvert*).	*Peu de personnes sont capables d'expliquer les auteurs anciens* **ad aperturam libri.**

4th edition from 1925

Dictionary Division

3. The third part, *histoire et géographie*, contains short encyclopedic descriptions of people, countries, locations, intellectual or art works (660 pages).



HISTOIRE — GÉOGRAPHIE

AA (d'un mot germanique qui signifie *eau*, et qui est resté le nom d'un grand nombre de rivières des pays celtiques et germaniques). Fleuve côtier de France, qui baigne Saint-Omer et se jette dans la mer du Nord ; 80 kil.

AALBORG, v. et port du Danemark, ch.-l. du diocèse de ce nom ; 57.700 h.

AALESUND, v. de Norvège, dans une île de l'Atlantique en avant de la côte ; 16.500 h. Patrie de Rollon.

AALI-PACHA, homme d'État turc, né à Constantinople. Il a attaché son nom à la politique de réformes du *Tanzimat* (1815-1871).

AAR, riv. de Suisse, qui naît au col du Grimsel, arrose Berne, Soleure, Aarau, reçoit la Reuss, la Limmat, la Thièle et se jette dans le Rhin (r. g.) ; 280 kil.

AARAU, v. de Suisse, ch.-l. du cant. d'Argovie, sur l'Aar ; 10.700 h.

AARHUUS, v. et port du Danemark, ch.-l. du diocèse de ce nom ; 81.200 h.

AARON, frère aîné de Moïse, le premier grand prêtre des Hébreux (*Bible*).

ABA (Samuel), roi de Hongrie, de 1044 à 1044.

ABACO, la plus grande des îles Lucayes ; 3.400 h.

ABAD Ier (*Ibad'*), premier roi maure de Séville et chef de la dynastie des *Abadites* ; il régna de 1023 à 1042. — Son fils ABAD II régna de 1042 à 1069, et son petit-fils, ABAD III, de 1069 à 1095.

ABADIE (*di*) (Paul), architecte fr., né à Paris. On lui doit les plans du Sacré-Cœur, à Paris (1812-1884).

ABAFFI ou **APAFFI Ier**, prince élu de Transylvanie ; il régna de 1661 à 1690.

ABAILARD. V. ABÉLARD.

ABANO-BAGNI, v. d'Italie (prov. de Padoue) ; 5.700 h. Eaux minérales chlorurées sodiques, employées contre les rhumatismes, la goutte, etc.

ABARIM (*rim'*), chaîne de montagnes de l'Anti-Liban (Palestine), au N.-E. de la mer Morte, qui contient le mont Nébo où, dit la Bible, mourut Moïse.

ABAUZIT (*zi'*) (Firmin), philosophe et théologien protestant français, né à Uzès, réfugié à Genève après la révocation de l'édit de Nantes (1679-1767).

ABAZIE (*zi*), région septentrionale de la Russie transcaucasienne, sur la côte de la mer Noire, au N.-O. du pays des Tcherkesses. (Hab. *Abazes.*)

ABBADIE (*di*) (Jacques), théologien protestant français, né à Nay (Basses-Pyrénées) (1654-1727).

ABBAS (*bass*), oncle paternel de Mahomet (566-653).

ABBAS LE GRAND, schah de Perse, de la dynastie des Sofis, conquérant et administrateur éminent (1557-1628). — ABBAS II régna de 1641 à 1666. — ABBAS III régna de 1732 à 1736.

ABBAS-PACHA, vice-roi d'Égypte, petit-fils de Méhémet-Ali (1813-1854).

ABBAS-PACHA-HILMI, khédive d'Égypte, né en 1874. Il a régné, depuis 1892, sous la surveillance de l'Angleterre, qui l'a déposé en 1914.

ABBASSIDES, dynastie de 37 califes arabes, fondée par Aboul-Abbas, qui détrôna les Ommiades en 750. Les Abbassides régnèrent à Bagdad de 762 à 1258.

ABBATTUCCI, nom d'une famille originaire de Zicavo (Corse), qui a fourni plusieurs personnages célèbres : JACQUES-PIERRE, général, le constant adversaire de Paoli (1726-1812) ; — CHARLES, son fils, également général, né en 1774, tué au siège de Huningue en 1796 ; — JACQUES-PIERRE-CHARLES, neveu du précédent, ministre de la Justice sous Napoléon III (1792-1857).

Abbaye, monastère d'hommes ou de femmes, dont les revenus constituaient un bénéfice au profit de l'abbé ou de l'abbesse. L'abbaye *régulière* était celle dont l'abbé était un religieux et l'abbaye *en commende* celle dont l'abbé était un séculier sans autorité sur les moines, mais jouissant des revenus de l'abbaye.

Abbaye (*prison de l'*), construite à Paris entre 1631 et 1635, pour servir de prison seigneuriale à l'abbaye de Saint-Germain-des-Prés. Affectée après 1789 aux détenus militaires, elle fut le théâtre des massacres des 2 et 3 septembre 1792 ; démolie en 1854.

ABBEOKOUTA ou **ABEOKOUTA**, v. de la Nigéria (territ. de Yoruba) ; 51.000 h. Ch.-l. d'un district peuplé de 265.000 h.

ABBEVILLE, ch.-l. d'arr. (Somme) ; port sur la Somme ; ch. de f. N., à 45 kil. N.-O. d'Amiens ; 21.472 h. (*Abbevillois*). Commerce de drap, de tapis. Patrie de Millevoye, Pongerville, de l'amiral Courbet, etc. L'arr. a 11 cant., 172 comm., 123.350 h.

ABBIATEGRASSO, v. d'Italie (prov. de Milan), sur le Naviglio Grande, près de laquelle fut tué Bayard en 1524 ; 10.850 h.

ABBON, moine de Saint-Germain-des-Prés, auteur du *Siège de Paris par les Normands*, poème latin d'un grand intérêt historique (860-923).

★ 37

ALGER.

4th edition from 1925

Core Cultural Knowledge

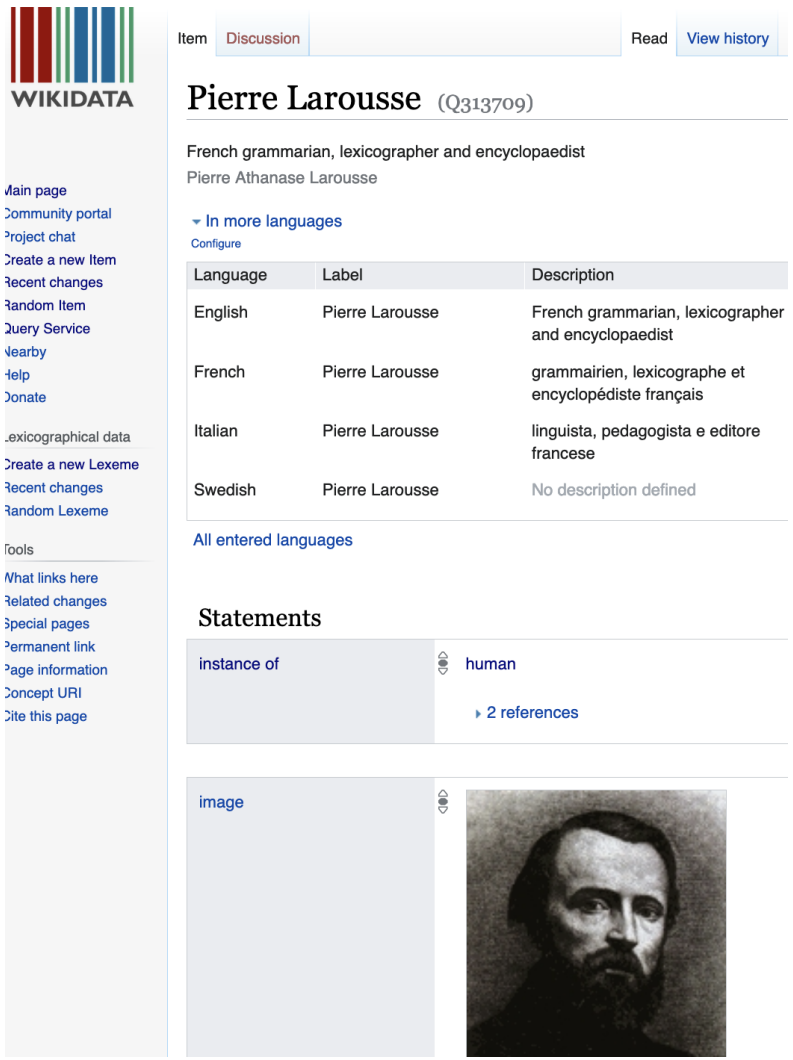
- The *Larousse* content is highly didactic and corresponds to a core cultural knowledge in France at the beginning of the 20th century
- It is a snapshot of the world's view and a repository of the cultural values at that time.
- The printed content of such a dictionary is then highly informative
- It is nonetheless limited and static by nature.
- Elementary information is sometimes missing or incomplete:
  - Places of birth and death of persons;
  - Gender and occupations;
  - Exact geographical coordinates of locations; etc.
- Makes the (automatic) analysis sometimes difficult.

Entity Linking

- In this work, we connected manually all the 20,245 entries from the *history and geography* part to wikipedia information (wikidata items).
- Enables us to:
  - Extract more data on the entities;
  - Verify information;
  - Apply some automatic analyses;
  - Compare historically-situated representations with currently available descriptions;
- The annotated dataset is available in the JSON format from `https://github.com/pnugues/petit_larousse_1905/`.

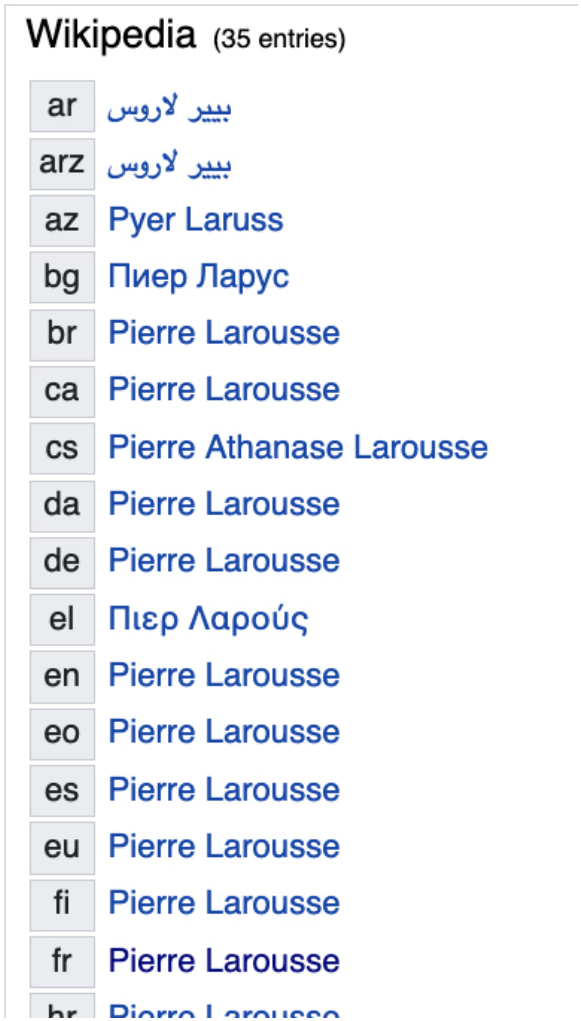
Wikidata

- As reference points and identifiers, we used the wikidata items.
- Wikidata assigns a unique identification number to each thing in wikipedia
- Pierre Larousse has identifier Q313709,
- Wikidata was designed to handle the multilingual versions of wikipedia.
- Contains structured information



Wikidata

- As reference points and identifiers, we used the wikidata items.
- Wikidata assigns a unique identification number to each thing in wikipedia
- Pierre Larousse has identifier Q313709,
- Wikidata was designed to handle the multilingual versions of wikipedia.
- Contains structured information



Wikidata

- As reference points and identifiers, we used the wikidata items.
- Wikidata assigns a unique identification number to each thing in wikipedia
- Pierre Larousse has identifier Q313709,
- Wikidata was designed to handle the multilingual versions of wikipedia.
- Contains structured information

Method

- We manually connected each entry of the *Larousse* to one or more wikidata items.
- We started from the OCRed and corrected version from the Nénufar website.

- Nénufar provides an XML-TEI structure of the entries, but we only used the raw text

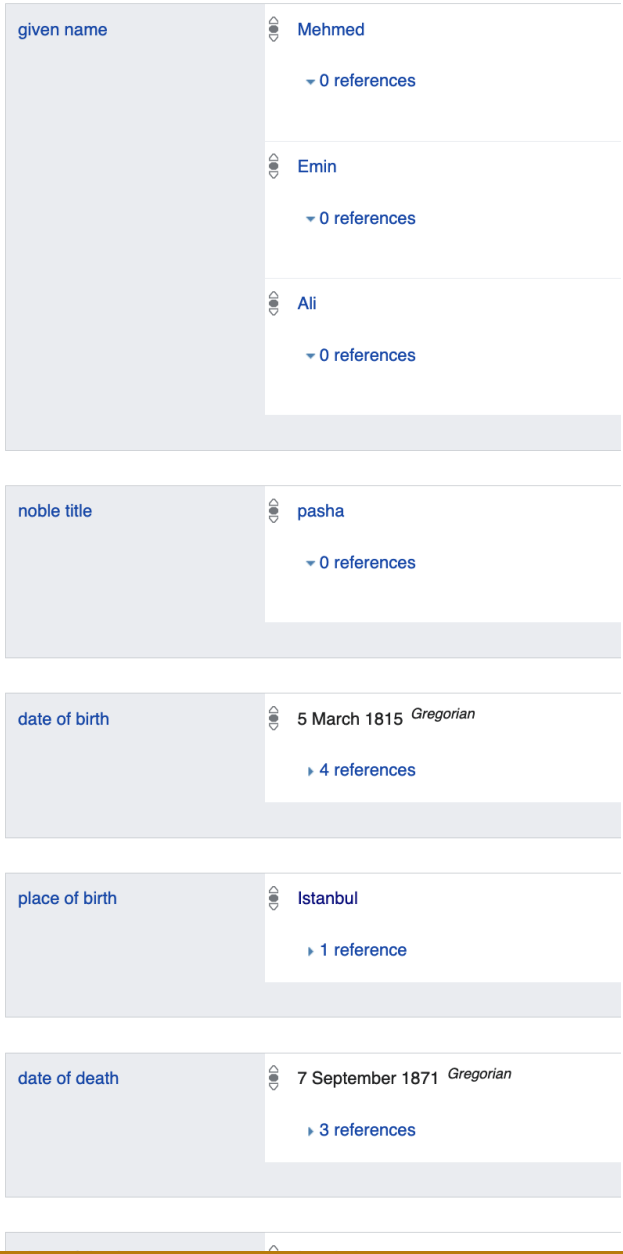
http://nenufar.huma-num.fr/

Example of Link

Some entries are easy to link like the person names:

> *AALI-PACHA, homme d'Etat turc, né à Constantinople. Il a attaché son nom à la politique de réformes du Tanzimat (1815-1871).*
> *'AALI-PASHA, Turkish statesman, born in Constantinople. He attached his name to the policy of reforms of the Tanzimat (1815-1871).'*

corresponds to Q439237 as the occupation and the dates of birth and death match those in wikidata.



Example of Link

Some other entries, like countries, can change in nature and form.

> *GRANDE-BRETAGNE et IRLANDE (Royaume-Uni de)...*
> *'GREAT BRITAIN and IRELAND (United Kingdom of)...'*

We chose the `Q174193` identifier, which describes a state that existed from 1801 to 1927 when the *Larousse* was published.

United Kingdom of Great Britain and Ireland (Q174193)

historical state (1801–1922), name in use until 1927

United Kingdom | UK | GB | UKGBI | Great Britain and Ireland | the United Kingdom | Britain

In more languages

Configure

Language	Label	Description	Also known as
English	United Kingdom of Great Britain and Ireland	historical state (1801–1922), name in use until 1927	United Kingdom UK GB UKGBI Great Britain and Ireland the United Kingdom Britain
French	Royaume-Uni de Grande-Bretagne et d'Irlande	État historique (1801–1922), nom en usage jusqu'en 1927	
Italian	Regno Unito di Gran Bretagna e Irlanda	stato europeo esistito dal 1801 al 1922	Regno di Gran Bretagna Regno Unito di Gran Bret
Swedish	Förenade kungariket Storbritannien och Irland	statsbildning mellan nuvarande Storbritannien och Irland 1801–1922	Förenade kungariket Storbritannien och Irland Förenade konungariket S

All entered languages

Statements

instance of	sovereign state		edit
	start time	1 January 1801 *Gregorian*	
	end time	12 April 1927 *Gregorian*	
	follows	Great Britain	

Tricky Cases

Works of art sometimes have changed name or attribution.
For example, this painting in the *Petit Larousse illustré:*

> *Automne (l'), tableau de Jordaens (Bruxelles) ;*

When ambiguous, we relied on the more detailed *Nouveau Larousse illustré* (1897–1904):

> *Automne (REPRÉSENTATIONS DIVERSES DE L'). On connaît l'Automne, tableau de Jordaens (musée de Bruxelles), allégorie des occupations et des dons de l'automne. (Cette déesse, drapée dans un manteau rouge, a les mains pleines de raisins. Parmi les personnages qui l'entourent, on remarque une nymphe nue, vue de dos, aux formes opulentes) ;*

Jacob Jordaens - Allegory of Fertility - Google Art Project, Q22787456

Entries with Two or More Links

Some entries contain lists of people or works of art, as for example:

ABAD Ier [bad'], premier roi maure de Séville, et chef de la dynastie des Abadites ; il régna de 1023 à 1042. — Son fils Abad II régna de 1042 à 1069, et son petit-fils, Abad III, de 1069 à 1095.

'ABBAD I, first Moorish king of Seville, and head of the Abbadid dynasty; he reigned from 1023 to 1042. — His son Abbad II reigned from 1042 to 1069, and his grand-son, Abbad III, from 1069 to 1095.'

- Most of the time, such lists contain a dash character '–' to separate the entities.
- Using this character, we extracted all the lists and we connected all the names mentioned in them.
- For the previous example, we linked the three entities to:

  `["Q305795", "Q30556", "Q299578"]`.

Dataset Format and Size

- We stored the annotation as a list of dictionaries in a JSON file.
- Each dictionary represents an entry and has two keys: the raw text of the entry and a list of wikidata identifiers, most often only one.
- For the AALI-PASHA entry, this corresponds to:

```
"texte": "AALI-PACHA, homme d'Etat turc, né à
         Constantinople.  Il a attaché son nom à la
         politique de réformes du Tanzimat
         (1815-1871).",
  "qid": ["Q439237"]
```

- At the end, the complete annotation of the 20,245 entries resulted in 22,357 links, where 18,905 entries have one link and 1340 have two or more.
- The annotated dataset is available from GitHub: https://github.com/pnugues/petit_larousse_1905/.

The Semantic Categories

- Using the identifiers, we extracted the semantic category of the entries from wikidata.
- We used the SPARQL query language and triples like this one:

  `wd:Q439237 wdt:P31 ?type .`

  where `wd:Q439237` is Aali-Pasha's identifier, `wdt:P31` is a property meaning *instance of*, and `?type` is the type we want to extract.
- The SPARQL server returns the `Q5` identifier denoting a human.
- When an entry contained more than one name, we considered only the first identifier in the list.

The Ten most Frequent Categories

Q-Number	Frequency	Description
Q5	7653	human
Q484170	3022	commune of France
Q1549591	849	big city
Q515	789	city
Q7725634	572	literary work
Q4022	447	river
Q3305213	308	painting
Q747074	255	*comune* of Italy
Q1637706	231	city with a population of more than 1,000,000
Q23442	221	island

The two first types are human with nearly 38% of the entries and French commune with nearly 15%.

The Human Beings

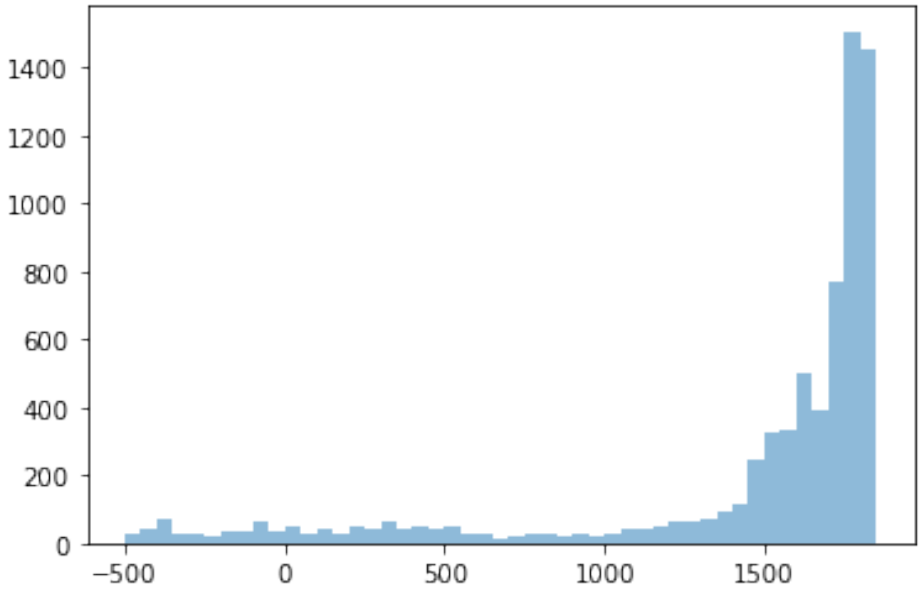
- Human being is the most frequent category and the corresponding wikidata items often provide the dates and places of birth and death of the person, sometimes the occupations, spouses, etc.
- Note that less known entities are not always completely documented.
- We extracted the dates of birth and death from wikidata.
- We extracted the dates with the SPARQL triples:

  ```
  wd:Q439237 wdt:P569 ?db .
  wd:Q439237 wdt:P570 ?dd .
  ```

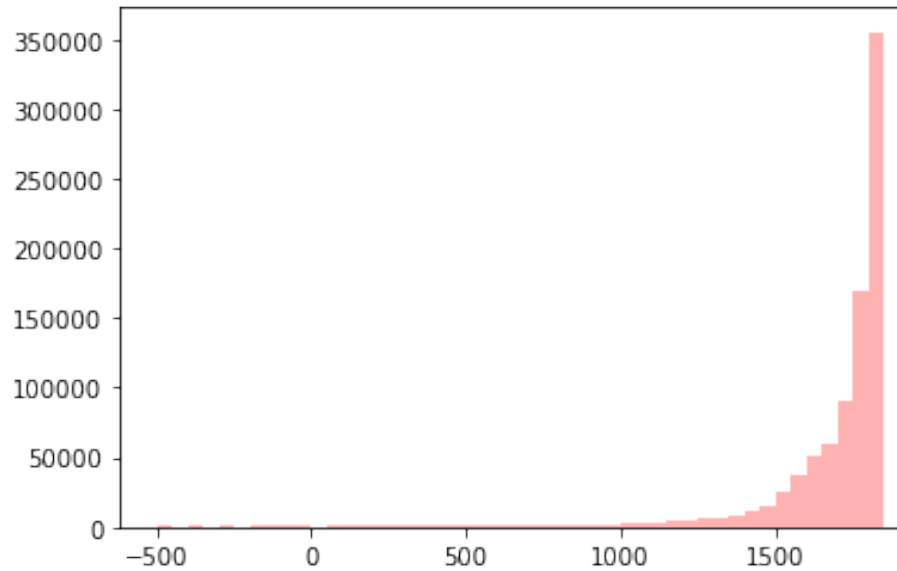
  where `wdt:P569` is the property for the date of birth and `wdt:P570` for the date of death.
- The wikidata queries for the human beings returned 7430 pairs.

Distribution Across the Years: Birth Dates

- Excluding the dates before -500 and after 1851, we computed the distribution of the birth dates of the humans.



Petit Larousse illustré

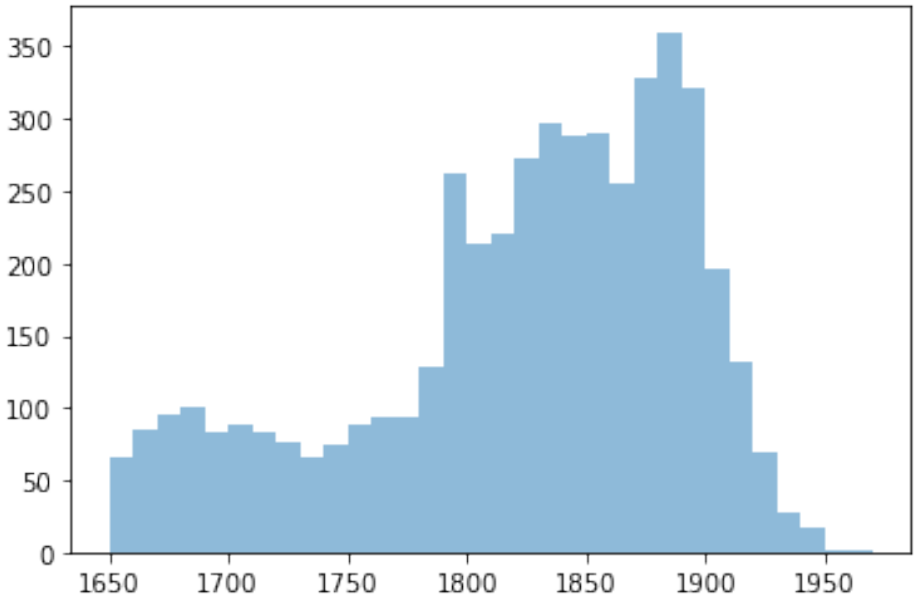


Wikidata

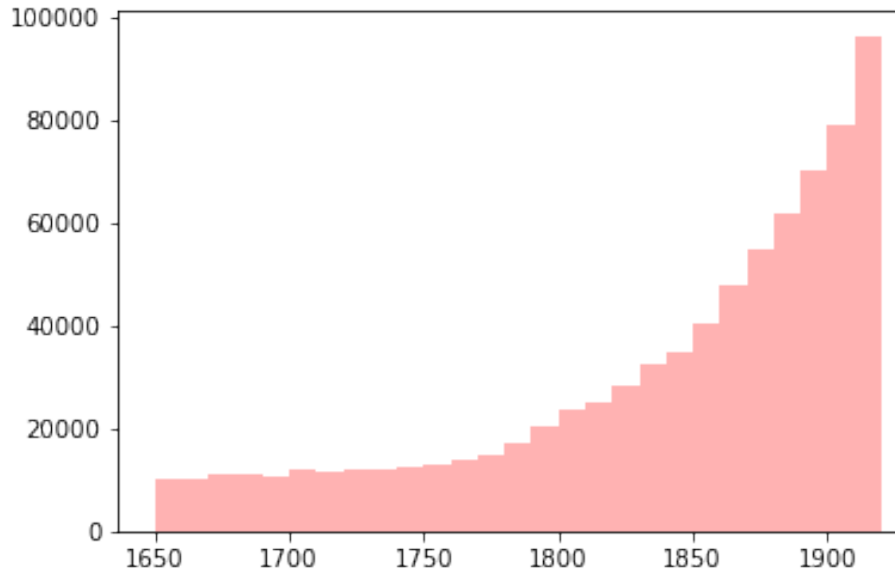
- The two largest bins are between 1750 and 1800 and 1800 and 1850. These two most recent bins are twice as large as the one just before them between 1700 and 1750.

Distribution Across the Years: Death Dates

- The death dates with a distribution starting at year 1650.



Petit Larousse illustré



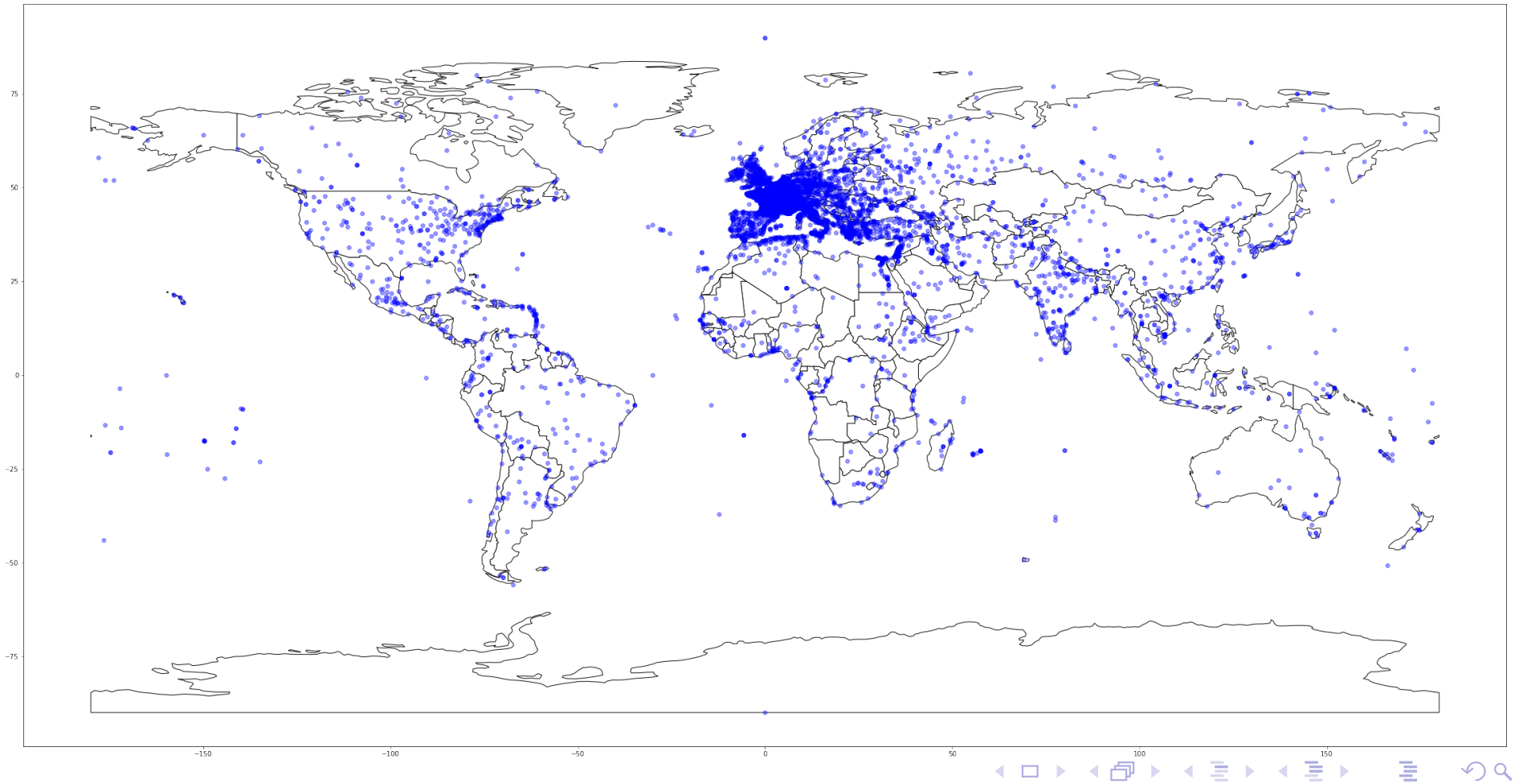
Wikidata

- Confirms the accent on contemporary or near-contemporary people. At the date of publication (1905), 334 people were still living
- Year 1800 is a turning point, reflecting a specific attention of the *Larousse* on the new era started by the French Revolution of 1789.

Geographical Entities

- The geographical entities in the *Larousse* convey a world view.
- We extracted their coordinates with the P625 property:

```
?e wdt:P625 ?geo .
```



Conclusion

- We annotated all the entries of the *history and geography* part of the *Petit Larousse illustré* from 1905 with their wikidata identifiers.
- We have outlined how to use the identifiers to extract, complement, and process knowledge on the entities.
- This has enabled us to exhibit more precisely the scope of information and world view conveyed by this dictionary.
- We have released the complete annotated corpus
- We hope that this language resource will be useful to train entity linking applications or facilitate new projects in digital humanities.