

# Linking Named Entities in Diderot's *Encyclopédie* to Wikidata

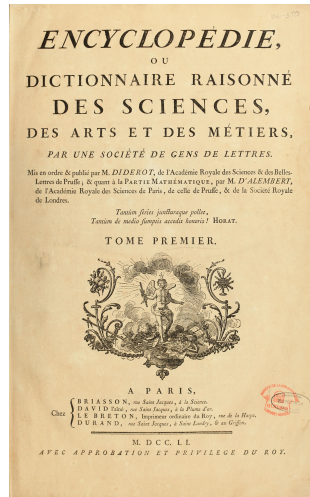
Pierre Nugues

Pierre.Nugues@cs.lth.se

LREC-COLING 2024  
Torino, May 22-24, 2024

# Diderot's *Encyclopédie*

- Encyclopedia published between 1751 and 1772
- Consists of 17 volumes of text and 11 volumes of plates
- By its size and influence, a milestone in the intellectual history of Europe.



Source of image: Wikimedia

# Age of the Enlightenment

The *Encyclopédie* embodies the ideas from the Enlightenment:

- A belief in rationalism summarized by *Sapere aude* 'Dare to know';
- Gathers the knowledge of its time;
- Asserted in the *explicit liber* of the *Encyclopédie*  
*ce Dictionnaire, destiné particulièrement à être le dépôt des connaissances humaines.*

*"this Dictionary, intended particularly to be the repository of human knowledge."*

Arguably, Wikipedia is today's most prominent reference work. It is the epitome of knowledge collection practices brought forward by the internet

- No competing multilingual encyclopedia or similar project has its reach, scope, and volume.
- Millions of contributors
- Popular reference for students, journalists, and academics,
- Endeavor stated in its Prime Objective:  
*Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. That's what we're doing*

# Contributions of this Work

A resource for entity linking:

- Annotate entries with their corresponding Wikidata identifiers
- Focused on location entries (15,274 entries) and the human beings they contain
- Completed the annotation of more than 9,400 entries
- Dataset available at [https://github.com/pnugues/encyclopédie\\_1751](https://github.com/pnugues/encyclopédie_1751) in the JSON format

# Potential Applications

- Connects the *Encyclopédie* to the Wikidata graph, Wikipedia, and to contemporary knowledge
- Enables the extraction of supplementary information: Coordinates for the places, status for people, etc.
- Entity linking resource for historical texts, for instance to train models
- Building block to understand knowledge transmission processes

# Material Structure and Organization

- About 74,000 entries, many tagged with a field and signed by an author
- Elaborate knowledge organization in fields and subfields;
- OCREd and proofread in three steps by University of Chicago, Wikisource, and Science Academy of France (ENCCRE)
- We used the ENCCRE version  
(<http://enccre.academie-sciences.fr/encyclopedia/>)

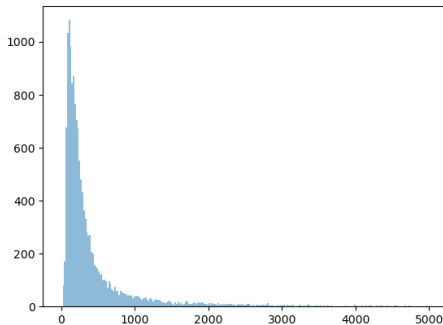


- We extracted all the entries with the geography field: 15,274 entries
- Corresponds to 20.6% of the entries

# Elementary Data Analysis

## Geography entries

- 1 Nearly 1,900,000 words in total
- 2 Mean of 123.5 words per entry and median of 42
- 3 Mean of 700 characters per entry and median of 232
- 4 The three longest entries are: *Fontaine* 'Spring', *Paris*, and *Géographie physique* 'Physical geography'



Frequency histogram of the distribution of entries by length in number of characters



# Entry Structure

Many entries describe only one location such as:

*WINDELINGEN, ou WINDLING, (Géog. mod.) petite ville d'Allemagne dans la Suabe, au duché de Wirtemberg, sur le Necker, près de l'embouchure de la Lauter. (D. J.)*

*“WINDELINGEN, or WINDLING, (Geog. mod.) small town in Germany in Swabia, in the Duchy of Wirtemberg, on the Necker, near the mouth of the Lauter.”*

Although the name has changed, easy to locate it from its geographic description or a simple query with search engine

This entry corresponds to *Wendlingen am Neckar* in Wikipedia and Q81860 identifier in Wikidata.

# Linking to Wikidata

*Encyclopédie* WINDELINGEN, ou WINDLING, (Géog. mod.) petite ville d'Allemagne dans la Suabe, au duché de Wirtemberg, sur le Neckar, près de l'embouchure de la Lauter. (D. J.)

## Wikidata item: Q81860

## Versions

## Structured information

**WIKIDATA**  
Main page  
Community portal  
Project chat  
Create a new item  
Recent changes  
Random item  
Query Service  
Nearby  
Help  
Donate  
Less prominent data  
Create a new Lexeme  
Recent changes  
Revision Lexeme  
Tools  
What links here  
Related changes  
Special pages  
Permanent link  
Page information  
Cite this page  
Get shortened URL  
Download QR code

**Wendlingen am Neckar** (Q81860)  
municipality in Germany  
Wendlingen  
+ 15 more languages  
edit

Language	Label	Description	Also known as
English	Wendlingen am Neckar	municipality in Germany	Wendlings
French	Wendlingen am Neckar	commune allemande	Wendlings
Swedish	Wendlingen am Neckar	stad i Tyskland	
American English	No label defined	No description defined	

All entered languages

**Statements**

Instance of


urban municipality in Germany

+ 0 references

+ add reference

+ add value

logo image


  
Signed Wendlingen am Neckar.jpg  
258 × 104, 34 KB

+ 0 references

+ add reference

+ add value

image

  
Tiefpunkt Stadtbild.jpg  
2,893 × 1,810, 544 KB

+ 0 references

+ add reference

+ add value

country

Germany

+ 0 references

+ add reference

+ add value

**Wikipedia** (36 entries) [edit](#)

als

Wendlingen am Neckar

ar

وندلينغن آن دير نكا

azb

وندلينغن آن دير نكا

ceb

Wendlingen am Neckar (munisipyo)

ce

Неккар-тиера-Вендлинген

de

Wendlingen am Neckar

en

Wendlingen

eo

Wendlingen am Neckar

es

Wendlingen am Neckar

fa

وندلينغن آن دير نكا

fi

Wendlingen am Neckar

fr

Wendlingen am Neckar

hu

Wendlingen am Neckar

hy

Նավադինգեն ամ Լեյքար

it

Wendlingen am Neckar

ja

グェンドリンゲン・アム・ネッカー

ku

Wendlingen

la

Wendlingen am Neckar

lld

Wendlingen am Neckar

lmo

Wendlingen am Neckar

ms

Wendlingen

nl

Wendlingen am Neckar

pl

Wendlingen am Neckar

pt

Wendlingen (Neckar)

ro

Wendlingen am Neckar

ru

Вендлинген-ам-Неккар

sh

Wendlingen am Neckar

sr

Вендлинген ам Некар

sv

Wendlingen am Neckar

tr

Wendlingen am Neckar


located in time zone

UTC+01:00  
valid in period  
standard time  
~ 0 references

located in or next to body of water

Neckar  
~ 0 references

coordinate location

  
48°40'29"N, 9°22'54"E

We encoded the corresponding entry as a JSON dictionary with the keys:

- `texte` that contains the text of the entry
- `qid`, a list wikidata identifiers
- `v17-1386-0`, the ENCCRE identifier that links this entry to the OCREd dataset: volume and entry number.

For the previous example, we have

```
{"texte": "WINDELINGEN, ou WINDLING, (Géog. mod.) petite ville  
d'Allemagne dans la Suabe, au duché de Wirtemberg, sur le  
Necker, près de l'embouchure de la Lauter. (D. J.)",  
"qid": ["Q81860"],  
"entreeid": "v17-1386-0"}
```

# Changing Names

Many entries have headwords that do not match those of Wikidata such as:

*BOINITZ, (Géog.) ville de la haute Hongrie, au comté de Zoll, rémarquable par ses bains & son safran. Long. 36. 40. lat. 48. 42.*

*“BOINITZ, (Geog.) town in upper Hungary, in the county of Zoll, remarkable for its baths & its saffron. Long. 36. 40. lat. 48. 42.”*

We have to look at other geographic sources to locate it.

This entry corresponds to *Bojnice*, today in Slovakia and its QID is [‘Q788753’]

# Sources of Geographic Information

The *Encyclopédie* authors reused earlier sources to write their entries, for example:

- *Grand dictionnaire géographique et critique*, A. A. Bruzen de La Martinière, 1726-1739
- *Dictionnaire géographique-portatif*, J.-B. Ladvocat (used the pseudonym Vosgien), 1749

In these sources, the headwords are often the same as in the *Encyclopédie*, but their entry contains more precise descriptions. It makes it easier to identify the places.

In a few cases, some contemporary authors compiled lists of updated names as in the paper *La Hongrie dans l'Encyclopédie* by Imre Vörös.

# Entry Structure: Multiple Subentries

Some entries have multiple subentries.

They are sometimes numbered and sometimes enumerated in a list:

*Chaumont, (Géog.) petite ville de France au Vexin. Il y a encore plusieurs petites villes de ce nom, une en Touraine, une autre en Savoie, & une troisieme au pays de Luxembourg.*

*“Chaumont, (Geog.) small town in France in Vexin. There are some other small towns of this name, one in Touraine, another in Savoie, and a third in the country of Luxembourg.”*

We encoded the links as a list of wikidata identifiers, here four:

```
'qid': ['Q737436', 'Q635143', 'Q819275', 'Q21551205']
```

# Unknown Places

Many references to Ancient Greece geographers such in the *Physcus* entry that mentions seven locations:

*“PHYSCUS, (Anc. Geog.) there are several places of this name; namely, 1°. A city in Asia Minor, [...] 2°. A city of the Ozoles of Locris, Plutarch speaks of it in his Greek questions; 3°. a city in Caria, according to Stephanus of Byzantium; 4°. a city in Macedonia, according to the same author; ...”*

We could not identify items 2, 3, 5 and 7.

We annotated *Physcus* with this list of Wikidata identifiers:

[`'Q209908'`, `'Q0'`, `'Q0'`, `'Q60792888'`, `'Q0'`, `'Q7826058'`, `'Q0'`],

where `'Q0'` does not exist and marks the unresolved entities.

# Entry Structure: Human Beings

Biographies show as subentries of geographical entries as

*“GRENOBLE, Gratianopolis, (Geogr.) ancient city of France [...]*

*Among the jurisconsults whose homeland is Grenoble are Pape (Guy), who died in 1487; his collection of decisions on the finest questions of law, has not yet been forgotten.*

*Mr. de Bouchenu de Valbonnais, (Jean Pierre Moret) first president of the parliament of Grenoble, born in this city on June 23, 1651, deserves the title of the most learned historiographer, [...]”*

We annotated the *Grenoble* entry with three Wikidata identifiers:

- ❶ Q1289, the city of Grenoble;
- ❷ Q41617345, Gui Pape (c. 1402-1487), French jurist-consult; and
- ❸ Q3169582, Jean-Pierre Moret de Bourchenu (1651-1730), French historian.



# Annotated Resource

The complete resource consists of a list of 15,274 dictionaries.

We annotated more than 9,400 entries with their QIDs:

- 841 entries contain the description of at least one human being
- We linked them to 2664 Wikidata identifiers including 1716 human beings

We annotated all the remaining entries with a `qid_region` key.

We used the main region in the entry definition such as *Italy* in:

*ASTRUNO, montagne d'Italie, au royaume de Naples, près de  
Puzzol ; [...]*

*“ASTRUNO, mountain of Italy, in the kingdom of Naples, near  
Puzzol; [...]”*

We restricted the values to 32 regions

We released the dataset on GitHub in JSON

([https://github.com/pnugues/encyclopedia\\_1751](https://github.com/pnugues/encyclopedia_1751)).

# Extracting Wikidata Information

Using the identifiers, we can extract Wikidata information with the SPARQL query language

- Queries consist of triples like:

```
wd:Q41617345 wdt:P31 ?type .
```

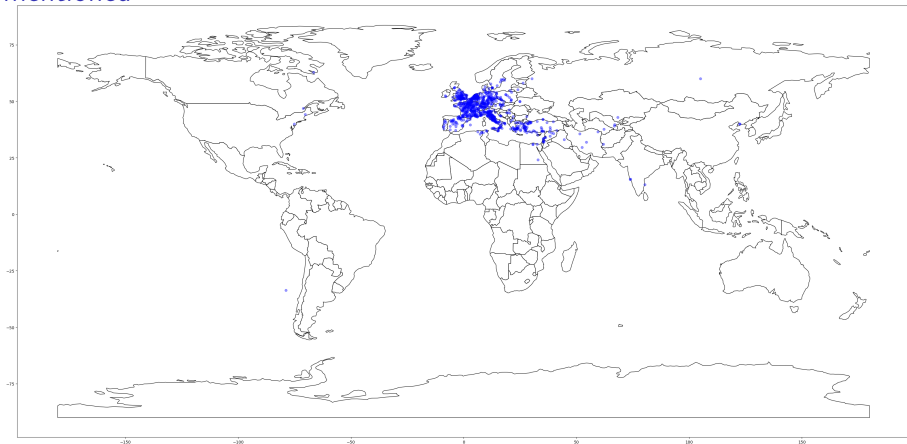
where `wd:Q41617345` is Gui Pape's identifier, `wdt:P31` is a property meaning *instance of*, and `?type` is the type we want to extract.

- The server returns the Q5 identifier denoting a human.

Using the same method, we extracted the geographical coordinates of the locations as well as the dates of birth and death of the human beings and their occupations

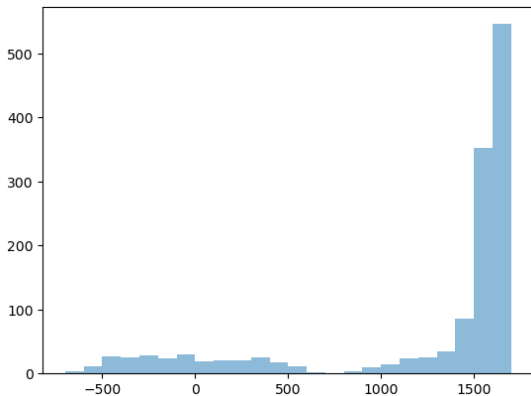
# Analyzing the Entities

Locations of the *Encyclopédie* headwords where a human being is mentioned



# Analyzing the Entities: Human Beings

Dates of deaths of the people mentioned in the *Encyclopédie* between -700 and 1700



# Analyzing the Entities: Occupations

Occupations of the human entities extracted from Wikidata. Note that an entity may have more than one occupation

<b>Qid</b>	<b>Description</b>	<b>Count</b>
Q36180	Writer	545
Q1234713	Theologian	285
Q49757	Poet	281
Q4964182	Philosopher	249
Q201788	Historian	245
Q1622272	University teacher	224
Q82955	Politician	199
Q250867	Catholic priest	144
Q333634	Translator	125
Q170790	Mathematician	123

# Future Work and Conclusion

As NLP resource, we hope this dataset will help:

- Train and assess entity solvers for historic text.
- Facilitate further connections with other data sources
- Serve research on knowledge transmission

As future work, we plan to annotate the rest of the geographic entities  
We believe this work could be adapted to other encyclopedias of the  
same time, in French or in other languages, like, in German, the  
*Universal-Lexicon* from 1731 to 1754.