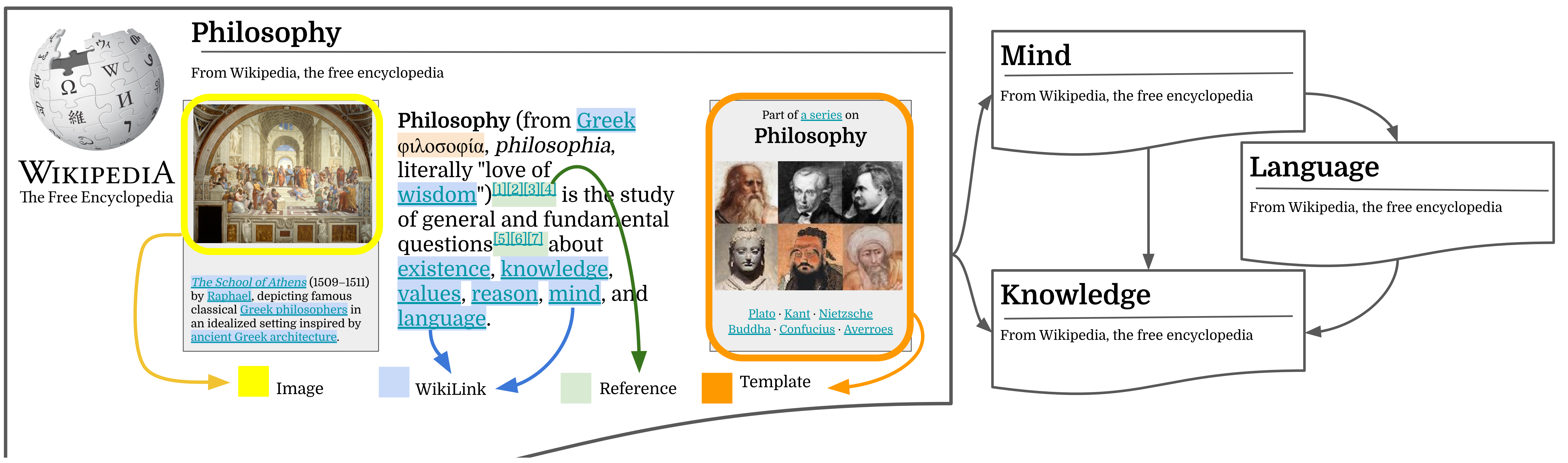
WikiLinkGraphs: A complete, longitudinal and multi-language dataset of the Wikipedia link networks

Cristian Consonni[1], David Laniado[2], Alberto Montresor[1]

[1] DISI, University of Trento, Italy [2] Eurecat, Centre Tecnològic de Catalunya, Spain

Motivation


WIKIPEDIA
The Free Encyclopedia
Philosophy
From Wikipedia, the free encyclopedia
The School of Athens (1509–1511) by Raphael, depicting famous classical Greek philosophers in an idealized setting inspired by ancient Greek architecture.
Philosophy (from Greek φιλοσοφία, philosophia, literally "love of wisdom")[1][2][3][4] is the study of general and fundamental questions[5][6][7] about existence, knowledge, values, reason, mind, and language.
Part of a series on
Philosophy
Plato · Kant · Nietzsche
Buddha · Confucius · Averroes
Image
WikiLink
Reference
Template
Mind
From Wikipedia, the free encyclopedia
Language
From Wikipedia, the free encyclopedia
Knowledge
From Wikipedia, the free encyclopedia

Dataset
on Zenodo
zenodo.org/record/2539424

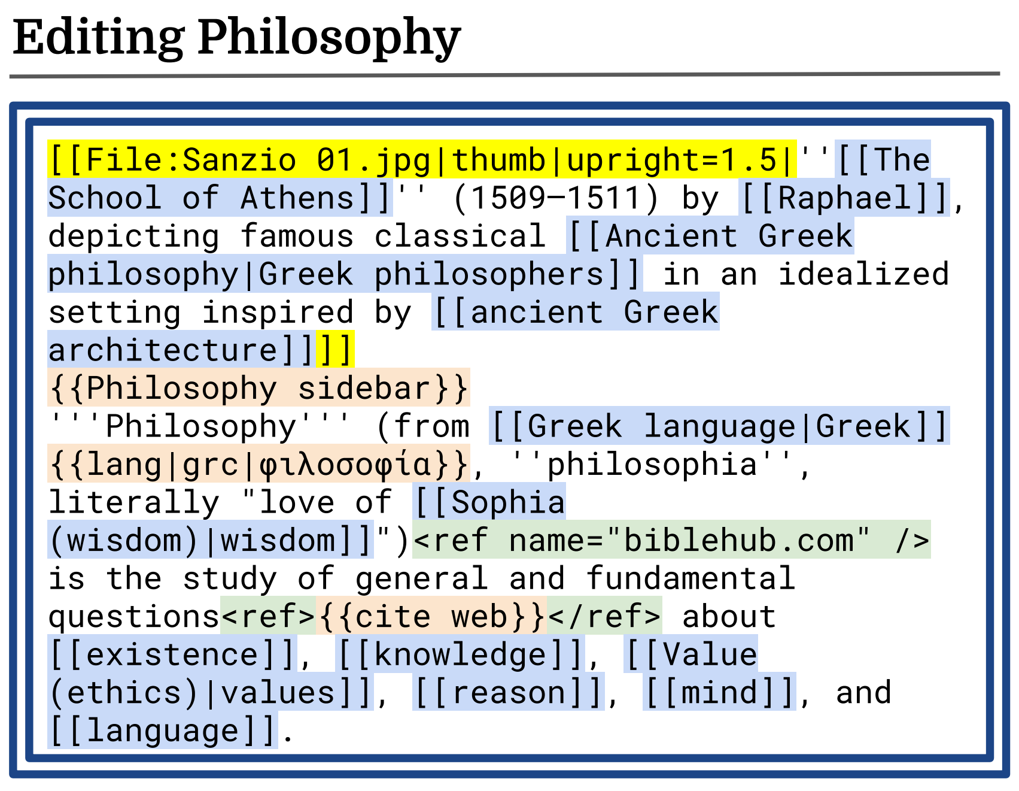
Additional datasets
on UniTN
cricca.disi.unitn.it/datasets

Code
on GitHub
github.com/WikiLinkGraphs

Wikipedia articles contain multiple links connecting a subject to other pages of the encyclopedia. In Wikipedia parlance, these links are called internal links or **wikilinks**. While previous work has mostly focused on the complete hyperlink graph which includes also links automatically generated by templates, we parsed each revision of each article to track links appearing in the main text. In this way we obtained a **cleaner network**, discarding more than half of the links and representing **all and only the links intentionally added by editors.**

Extraction process

Mining the Wikitext



Editing Philosophy

```
[[File:Sanzio 01.jpg|thumb|upright=1.5|''[[The
School of Athens]]'' (1509–1511) by [[Raphael]],
depicting famous classical [[Ancient Greek
philosophy|Greek philosophers]] in an idealized
setting inspired by [[ancient Greek
architecture]]]]
{{Philosophy sidebar}}
'''Philosophy''' (from [[Greek language|Greek]]
{{lang|grc|φιλοσοφία}}, ''philosophia'',
literally "love of [[Sophia
(wisdom)|wisdom]]")<ref name="biblehub.com" />
is the study of general and fundamental
questions<ref>{{cite web}}</ref> about
[[existence]], [[knowledge]], [[Value
(ethics)|values]], [[reason]], [[mind]], and
[[language]].
```

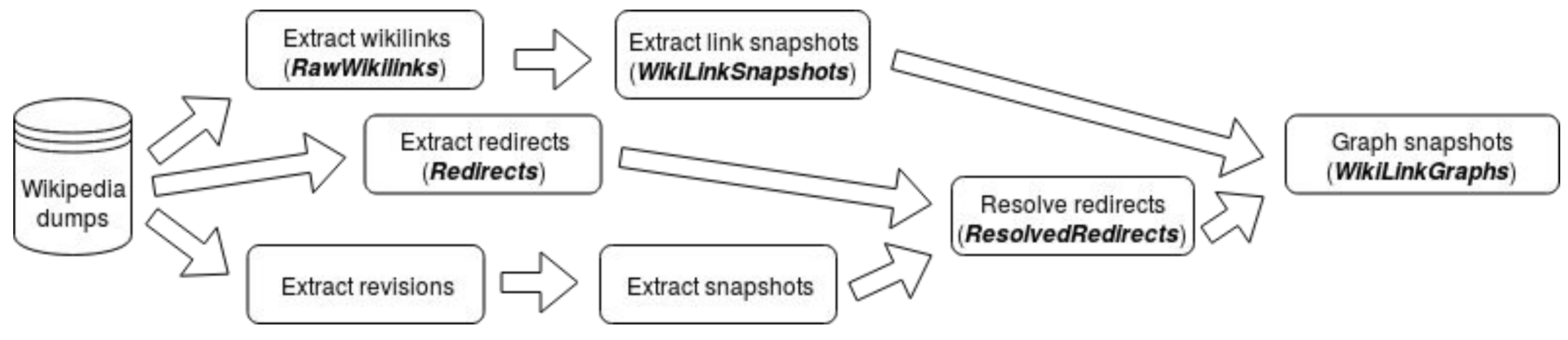
Example of Wikitext from the page “Philosophy” on English Wikipedia.

Wikilink Regex

```
\[\[
(?P<link>
[^\n\|\]\[\<\>\{\}]{0,256}
)
(?:
  \|
  (?P<anchor>
   [^\[]*?
  )
)?
\]\]
```

Regular expression used for the extraction of Wikilinks from Wikitext.

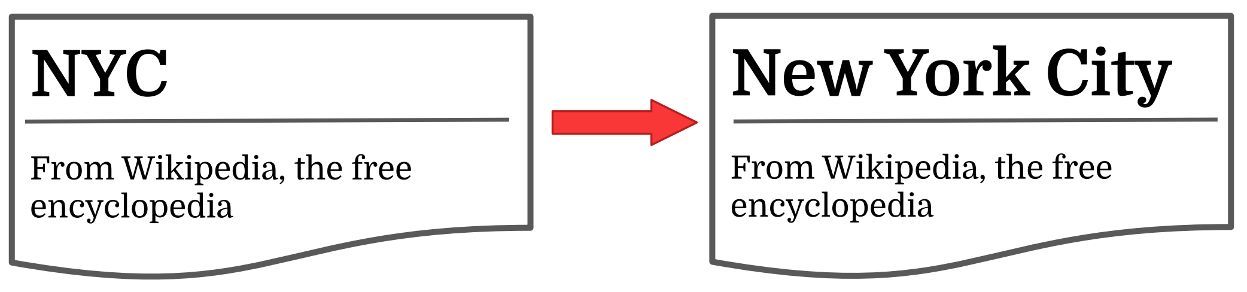
From Wikipedia’s XML dumps to WikiLinkGraphs


Wikipedia dumps
Extract wikilinks (RawWikilinks)
Extract link snapshots (WikiLinkSnapshots)
Extract redirects (Redirects)
Resolve redirects (ResolvedRedirects)
Graph snapshots (WikiLinkGraphs)
Extract revisions
Extract snapshots

The process to produce the WikiLinkGraphs dataset from the Wikipedia dumps. In bold and italics the name of the intermediate datasets produced.

Consider the Redirect

A **redirect** is a page that automatically sends users to another page.


NYC
From Wikipedia, the free encyclopedia
New York City
From Wikipedia, the free encyclopedia

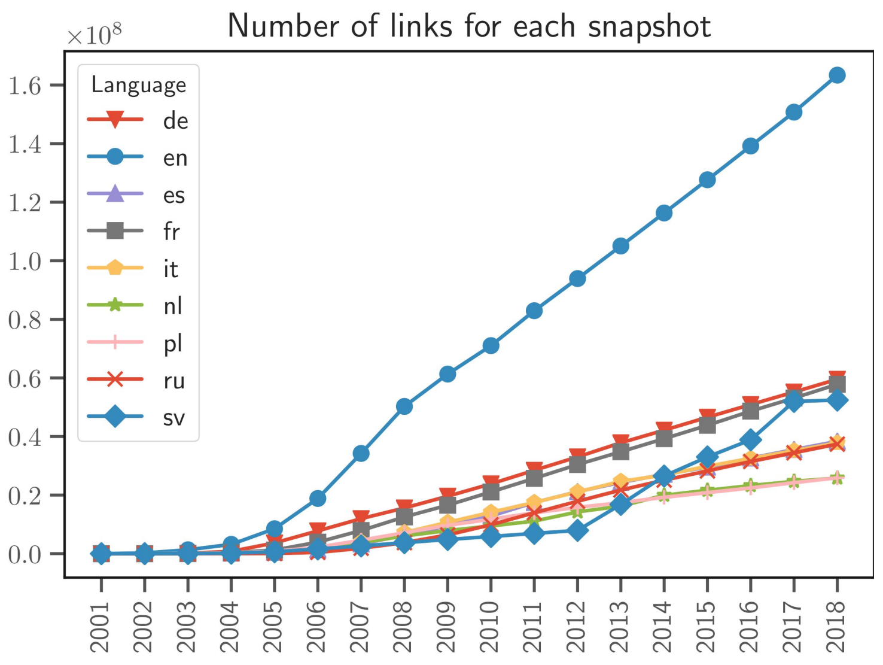
In the The WikiLinkGraphs dataset redirects are resolved to their targets.

lang	words
de	#WEITERLEITUNG
en	#REDIRECT
es	#REDIRECCIÓN, #REDIRECCION
fr	#REDIRECTION
it	#RINVIA, #RINVIO, #RIMANDO
nl	#DOORVERWIJZING
pl	#PATRZ, #PRZEKIERUJ, #TAM
ru‡	#PERENAPRAVLENIE, #PERENAPR
sv	#OMDIRIGERING

Words creating a redirect in MediaWiki for different languages. #REDIRECT is valid on all languages. (‡) For Russian Wikipedia, words are transliterated.

Dataset statistics

- **9 languages: de, en, es, fr, it, nl, pl, ru, sv**
- **17 years: from 2001-03 to 2018-03**
- **Yearly granularity**



Growth over time of the number of links in each snapshot in the WLG dataset.

ℓ	GB	N	E
de	5.7	3,588,883	59,535,864
en	17.0	13,685,337	163,380,007
es	3.0	3,034,113	38,348,163
fr	4.8	3,443,206	57,823,305
it	3.1	2,117,022	37,814,105
nl	2.0	2,626,527	25,834,057
pl	2.3	1,684,606	25,901,789
ru	3.2	3,360,531	37,394,229
sv	2.0	6,131,736	52,426,633

No. of nodes (N). edges (E) and size (GB), of the latest snapshot by language edition (ℓ).

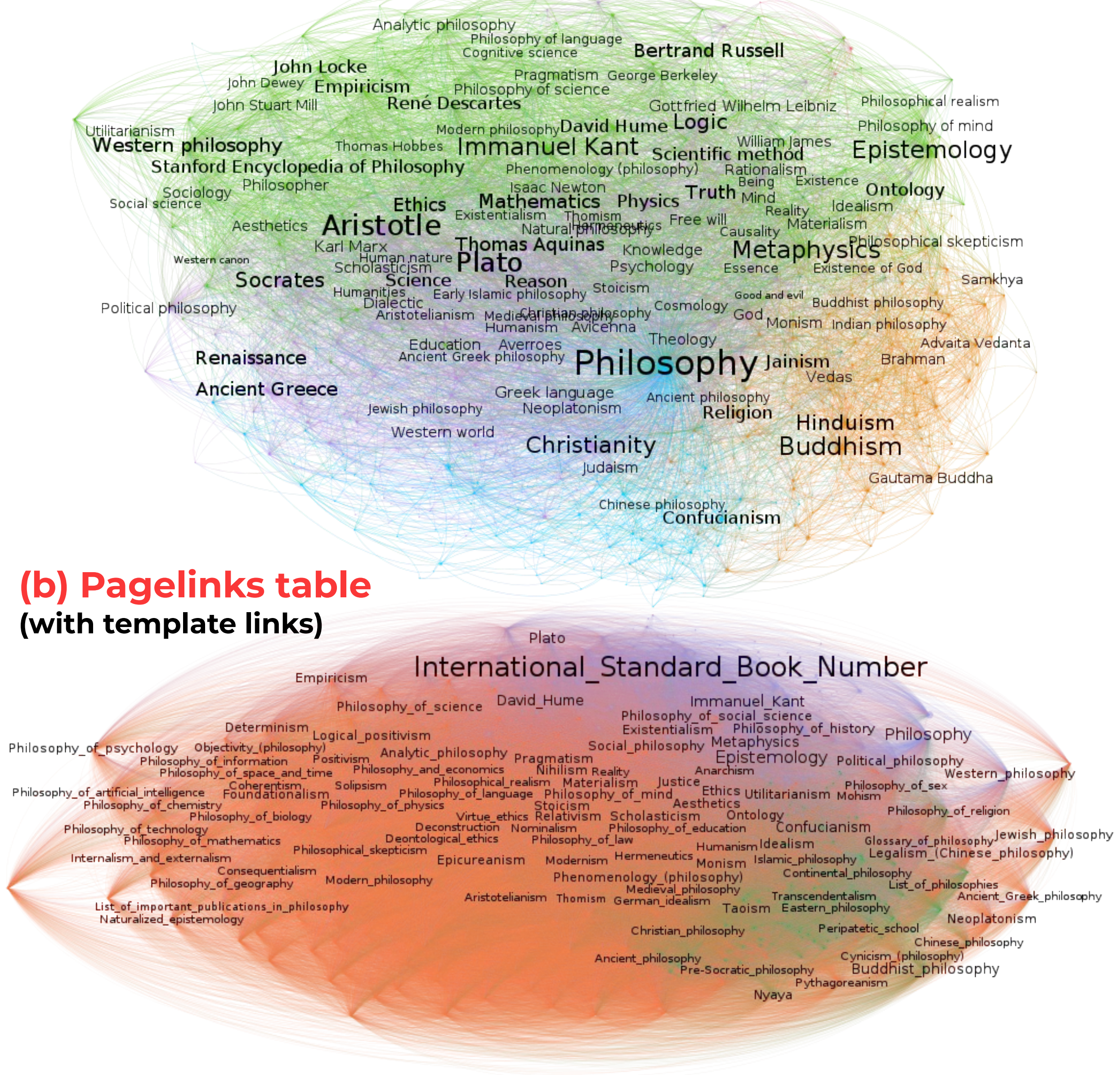
Applications: PageRank (α=0.85)

#	de		en		es		fr		it	
	article	score ($\times 10^{-3}$)	article	score ($\times 10^{-3}$)	article	score ($\times 10^{-3}$)	article	score ($\times 10^{-3}$)	article	score ($\times 10^{-3}$)
1	Vereinigte Staaten	1.646	United States	1.414	Estados Unidos	2.301	France	2.370	Stati Uniti d’America	3.076
2	Deutschland	1.391	World War II	0.654	España	2.095	États-Unis	2.217	Italia	1.688
3	Frankreich	1.020	United Kingdom	0.618	Francia	1.281	Paris	1.228	Comuni della Francia	1.303
4	Zweiter Weltkrieg	0.969	Germany	0.557	Idioma inglés	1.073	Allemagne	0.977	Francia	1.292
5	Berlin	0.699	The New York Times	0.527	Argentina	0.955	Italie	0.812	Germania	1.257
6	Österreich	0.697	Association football	0.525	Alemania	0.909	Royaume-Uni	0.773	Lingua inglese	1.228
7	Schweiz	0.691	List of sovereign states	0.523	Latín	0.867	Anglais	0.764	Roma	0.961
8	Englische Sprache	0.620	Race and ethnicity	0.500	Animalia	0.866	Francais	0.748	Centrocampista	0.861

Top-8 articles with the highest Pagerank score computed over the most recent snaphost of the WikiLinkGraphs dataset (2018-03-01) for German (de), English (en), Spanish (es), French (fr), and Italian (it) Wikipedia.

#	nl		pl		ru		sv	
	article	score ($\times 10^{-3}$)	article‡	score ($\times 10^{-3}$)	article‡	score ($\times 10^{-3}$)	article	score ($\times 10^{-3}$)
1	Kevers	3.787	Stany Zjednoczone	2.763	Soedinjonnye Shtaty Ameriki	3.290	Familj (biologi)	5.489
2	Vlinders	3.668	Polska	2.686	Sojuz Sovetskih Socialisticheskih Respublik	2.889	Släkte	5.184
3	Dierenrijk	3.294	Francja	2.360	Rossija	2.233	Nederbörd	4.696
4	Vliesvleugeligen	3.084	Jezyk angielski	2.110	Francija	1.190	Grad Celsius	4.144
5	Insecten	2.164	Łacina	1.914	Moskva	1.135	Djur	4.114
6	Geslacht (biologie)	2.101	Niemcy	1.698	Germanija	1.080	Catalogue of Life	3.952
7	Soort	1.954	Włochy	1.229	Sankt-Peterburg	0.881	Årsmedeltemperatur	3.878
8	Frankrijk	1.932	Wielka Brytania	1.124	Ukraina	0.873	Årsnederbörd	3.366

Top-8 articles with the highest Pagerank score computed over the most recent snaphost of the WikiLinkGraphs dataset (2018-03-01) for Dutch (nl), Polish (pl), Russian (ru), and Swedish (sv) Wikipedia. (‡) For Polish and Russian Wikipedia, article titles are transliterated



(a) WikiLinkGraphs
(without template links)

(b) Pagelinks table
(with template links)

Ego networks for the "Philosophy" page from English Wikipedia: (a) from the latest WikiLinkGraph snapshot (2018-03-01); (b) from the Pagelinks table in the Wikimedia dumps (2018-02-20), the latter contains links from templates. Bigger node size represents higher PageRank score within each network. Labels are show for nodes with degree higher than 50 for (a) and higher than 500 for (b) . colors represent clusters computed with the Louvain method.

License

Acknowledgements

This work has been supported by the European Union’s Horizon 2020 research and innovation programme under the EU Engineroom project, with Grant Agreement n° 780643.