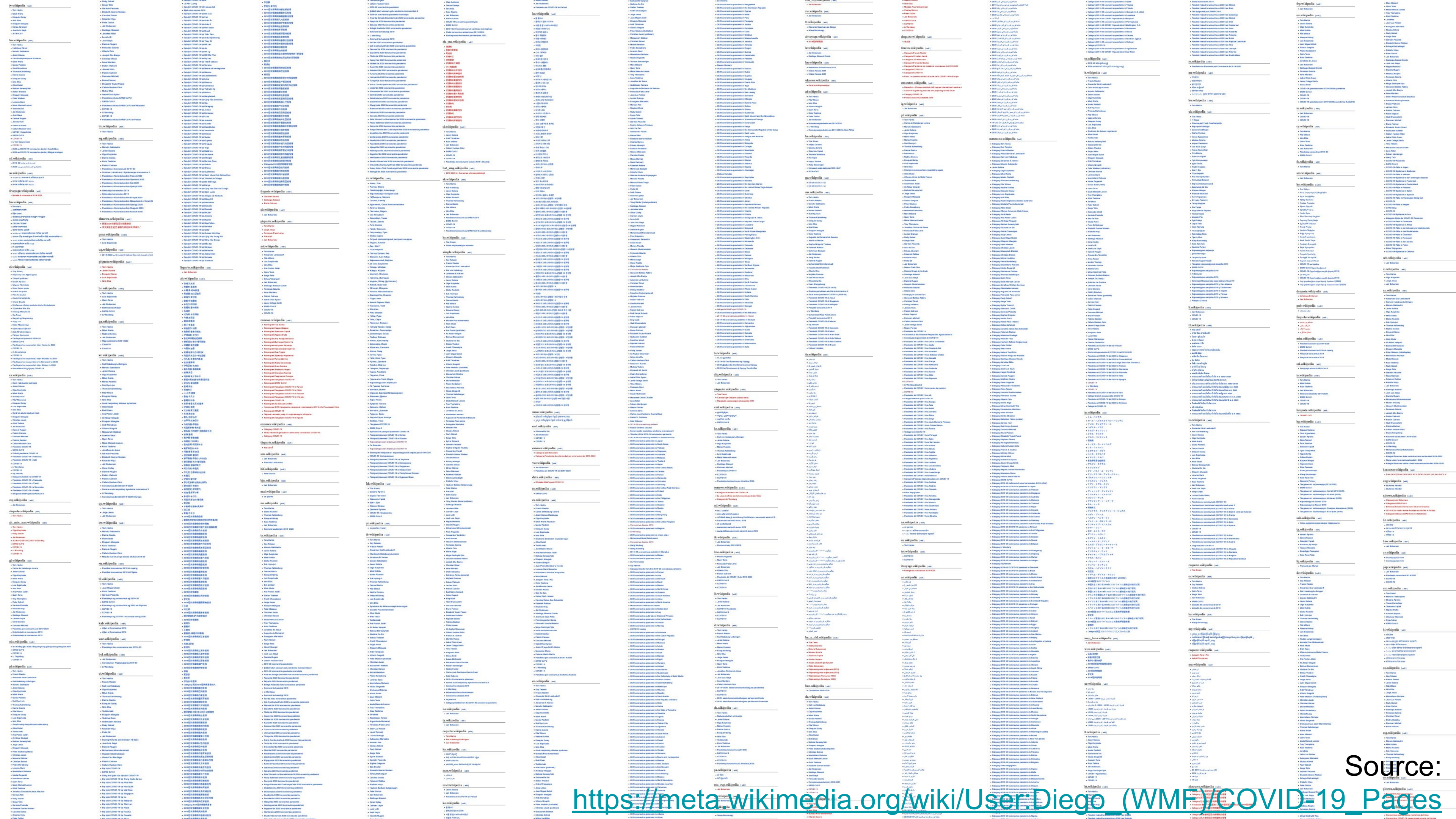
Topic Classification for Wikipedia

Me: Isaac (WMF) -- isaac@wikimedia.org
With plenty of support from: Martin Gerlach (WMF), Aaron Halfaker (WMF), Diego Sáez-Trumper (WMF), Growth Team (WMF), and many volunteers

Research Showcase -- 18 March 2020
https://www.mediawiki.org/wiki/Wikimedia_Research/Showcase#March_2020

Challenge One: Wikipedia is messy.

(aka Why are topics important to Wikipedia?)



Source: https://meta.wikimedia.org/wiki/User:Diego_(WMF)/COVID-19_Pages

So...

What if someone creates a **new article** about the pandemic in their country -- how can we automatically route that content to the right subject matter experts?

What if a **new reader** wants to find medical content?

What if a **new editor** wants to find relevant articles to which they can contribute?

So...

What if someone creates a **new article** about the pandemic in their country -- how can we automatically route that content to the right subject matter experts?

What if a **new reader** wants to find medical content?

What if a **new editor** wants to find relevant articles to which they can contribute?

Goal: label any Wikipedia article with high-level topics so that content can be routed to relevant readers and editors.

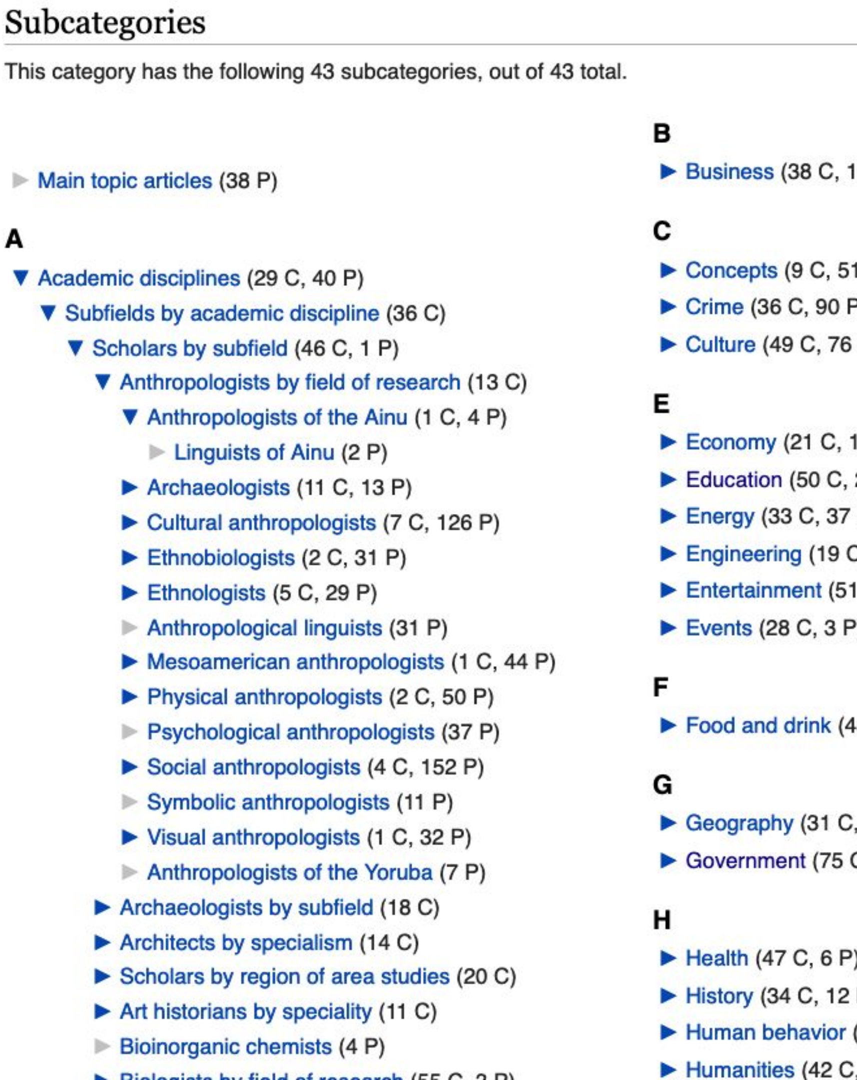
Potential Approaches

Category Network

https://en.wikipedia.org/wiki/Category:Main_topic_classifications (Image source)

- 43 high-level categories...but ~2M categories in English Wikipedia alone
- Powerful but very messy
- Often over-labeled -- e.g., alumni and Education

See Also: https://www.mediawiki.org/wiki/Wikimedia_Research/Showcase#March_2018



Subcategories

This category has the following 43 subcategories, out of 43 total.

▶ Main topic articles (38 P)

A

- ▼ Academic disciplines (29 C, 40 P)
  - ▼ Subfields by academic discipline (36 C)
    - ▼ Scholars by subfield (46 C, 1 P)
      - ▼ Anthropologists by field of research (13 C)
        - ▼ Anthropologists of the Ainu (1 C, 4 P)
          - ▶ Linguists of Ainu (2 P)
        - ▶ Archaeologists (11 C, 13 P)
        - ▶ Cultural anthropologists (7 C, 126 P)
        - ▶ Ethnobiologists (2 C, 31 P)
        - ▶ Ethnologists (5 C, 29 P)
        - ▶ Anthropological linguists (31 P)
        - ▶ Mesoamerican anthropologists (1 C, 44 P)
        - ▶ Physical anthropologists (2 C, 50 P)
        - ▶ Psychological anthropologists (37 P)
        - ▶ Social anthropologists (4 C, 152 P)
        - ▶ Symbolic anthropologists (11 P)
        - ▶ Visual anthropologists (1 C, 32 P)
        - ▶ Anthropologists of the Yoruba (7 P)
      - ▶ Archaeologists by subfield (18 C)
      - ▶ Architects by specialism (14 C)
      - ▶ Scholars by region of area studies (20 C)
      - ▶ Art historians by speciality (11 C)
      - ▶ Bioinorganic chemists (4 P)
      - ▶ Biologists by field of research (55 C, 2 P)

B

- ▶ Business (38 C, 1

C

- ▶ Concepts (9 C, 51
- ▶ Crime (36 C, 90 P
- ▶ Culture (49 C, 76

E

- ▶ Economy (21 C, 1
- ▶ Education (50 C,
- ▶ Energy (33 C, 37
- ▶ Engineering (19 C
- ▶ Entertainment (51
- ▶ Events (28 C, 3 P

F

- ▶ Food and drink (4

G

- ▶ Geography (31 C,
- ▶ Government (75 C

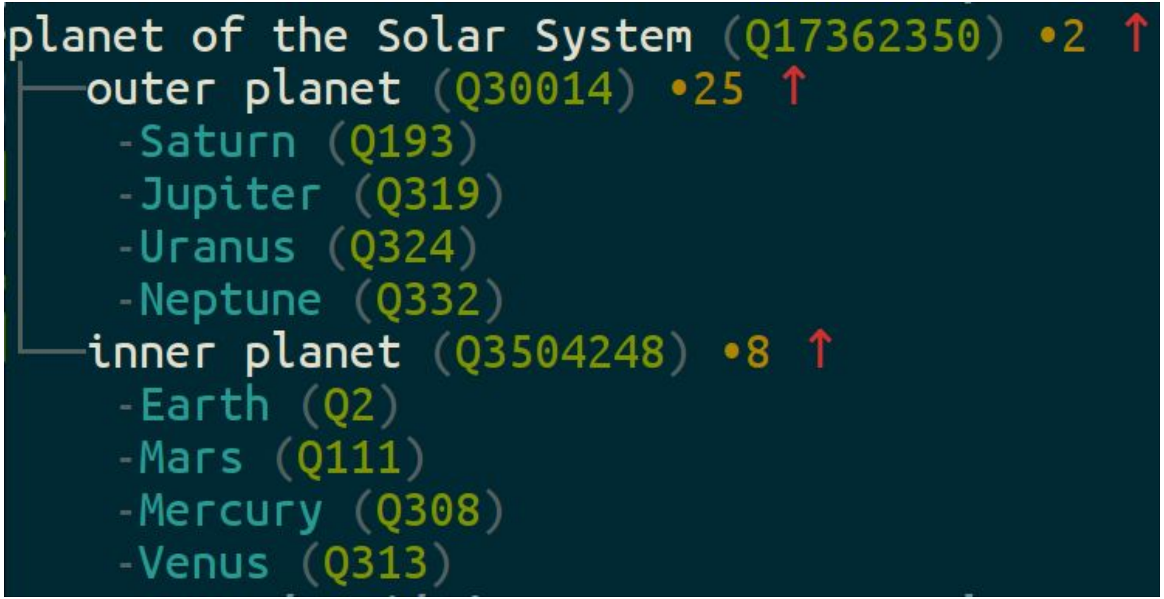
H

- ▶ Health (47 C, 6 P)
- ▶ History (34 C, 12
- ▶ Human behavior (
- ▶ Humanities (42 C

Wikidata Taxonomy

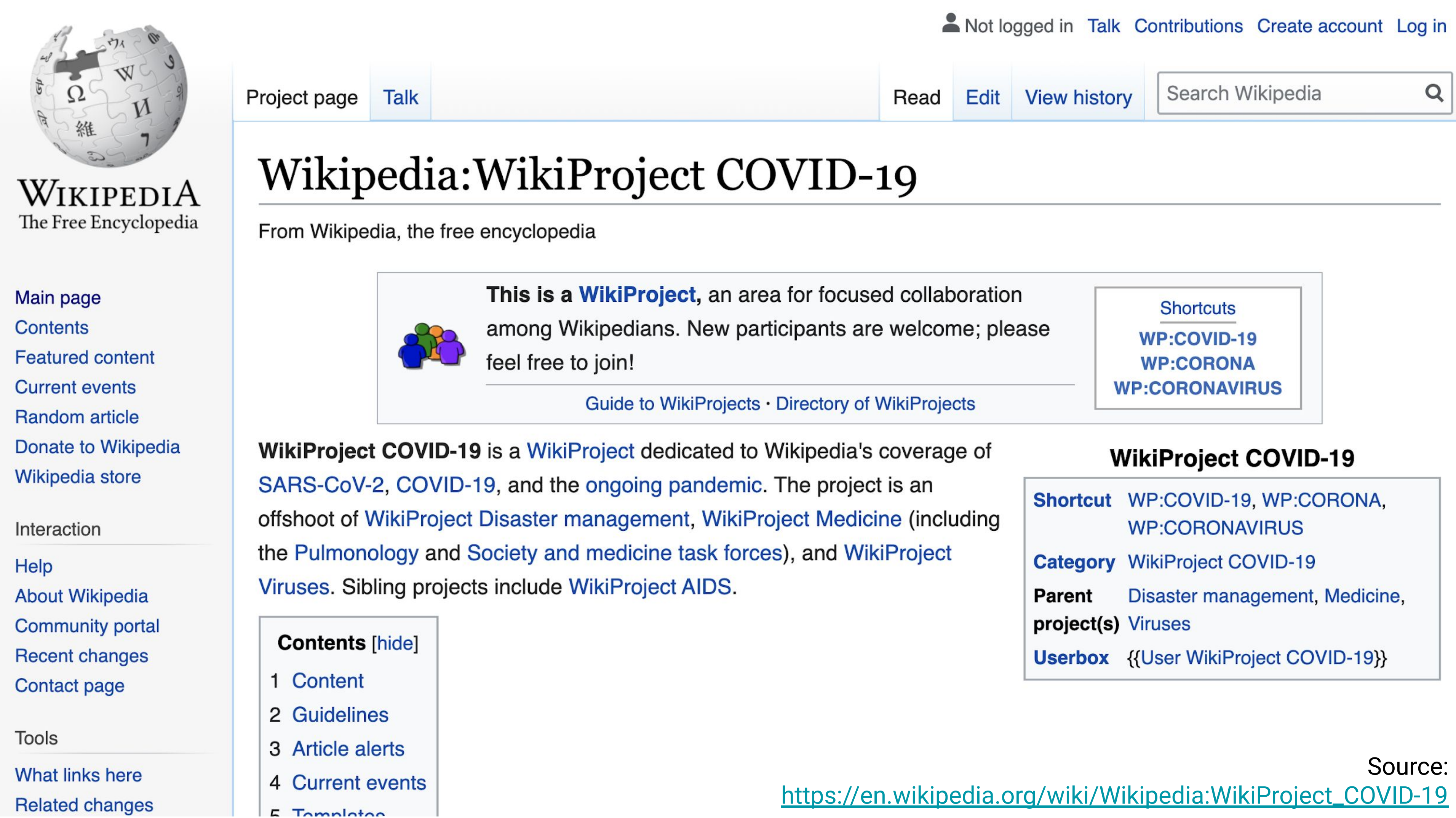
https://wikitech.wikimedia.org/wiki/Wikidata_Concepts_Monitor#WDCM_Taxonomy

- 14 high-level topics… but that is incomplete
- Designed to organize a knowledge base not be readily interpretable by humans



```
planet of the Solar System (Q17362350) •2 ↑
├──outer planet (Q30014) •25 ↑
│    -Saturn (Q193)
│    -Jupiter (Q319)
│    -Uranus (Q324)
│    -Neptune (Q332)
└──inner planet (Q3504248) •8 ↑
     -Earth (Q2)
     -Mars (Q111)
     -Mercury (Q308)
     -Venus (Q313)
```

Source: https://github.com/nichtich/wikidata-taxonomy/blob/master/img/wdtaxonomy-example.png



Wikipedia:WikiProject COVID-19

From Wikipedia, the free encyclopedia

This is a WikiProject, an area for focused collaboration among Wikipedians. New participants are welcome; please feel free to join!

Guide to WikiProjects · Directory of WikiProjects

Shortcuts
WP:COVID-19
WP:CORONA
WP:CORONAVIRUS

WikiProject COVID-19 is a WikiProject dedicated to Wikipedia's coverage of SARS-CoV-2, COVID-19, and the ongoing pandemic. The project is an offshoot of WikiProject Disaster management, WikiProject Medicine (including the Pulmonology and Society and medicine task forces), and WikiProject Viruses. Sibling projects include WikiProject AIDS.

WikiProject COVID-19

Shortcut	WP:COVID-19, WP:CORONA, WP:CORONAVIRUS
Category	WikiProject COVID-19
Parent project(s)	Disaster management, Medicine, Viruses
Userbox	{{User WikiProject COVID-19}}

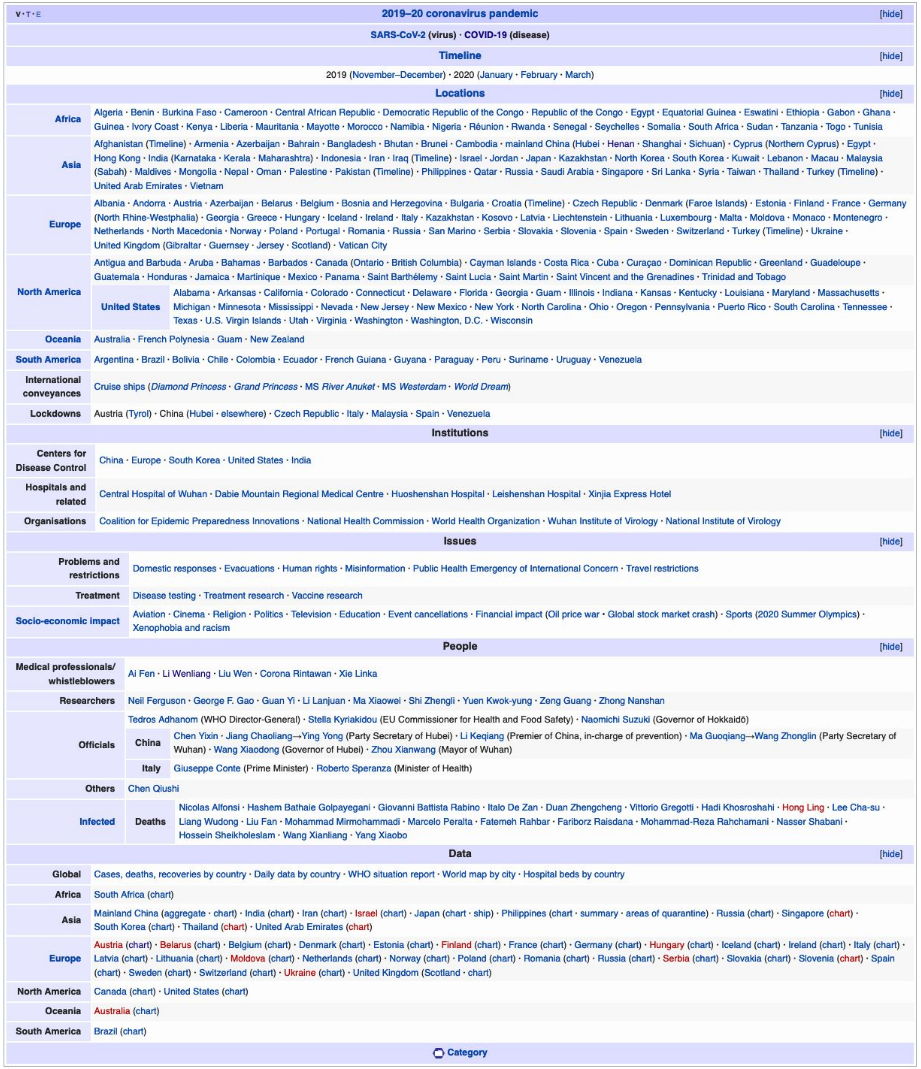
Contents [hide]

Source: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_COVID-19

Example: COVID-19 WikiProject (enwiki)

- 300 articles and growing
- Over 100 (sub)categories
- Over 60 participants (and hundreds of other editors)
- Hundreds of talk page discussions
- Over 1000 watchers
- 1.5 million pageviews / day
- And similar efforts in other languages, Commons, Wikidata, Wikivoyage etc.

Screenshot Source: https://en.wikipedia.org/wiki/Template:2019%E2%80%9320_coronavirus_pandemic



V · T · E **2019–20 coronavirus pandemic** [hide]

SARS-CoV-2 (virus) · COVID-19 (disease)

Timeline [hide]

2019 (November–December) · 2020 (January · February · March)

Locations [hide]

Africa		Algeria · Benin · Burkina Faso · Cameroon · Central African Republic · Democratic Republic of the Congo · Republic of the Congo · Egypt · Equatorial Guinea · Eswatini · Ethiopia · Gabon · Ghana · Guinea · Ivory Coast · Kenya · Liberia · Mauritania · Mayotte · Morocco · Namibia · Nigeria · Réunion · Rwanda · Senegal · Seychelles · Somalia · South Africa · Sudan · Tanzania · Togo · Tunisia
Asia		Afghanistan (Timeline) · Armenia · Azerbaijan · Bahrain · Bangladesh · Bhutan · Brunei · Cambodia · mainland China (Hubei · Henan · Shanghai · Sichuan) · Cyprus (Northern Cyprus) · Egypt · Hong Kong · India (Karnataka · Kerala · Maharashtra) · Indonesia · Iran · Iraq (Timeline) · Israel · Jordan · Japan · Kazakhstan · North Korea · South Korea · Kuwait · Lebanon · Macau · Malaysia (Sabah) · Maldives · Mongolia · Nepal · Oman · Palestine · Pakistan (Timeline) · Philippines · Qatar · Russia · Saudi Arabia · Singapore · Sri Lanka · Syria · Taiwan · Thailand · Turkey (Timeline) · United Arab Emirates · Vietnam
Europe		Albania · Andorra · Austria · Azerbaijan · Belarus · Belgium · Bosnia and Herzegovina · Bulgaria · Croatia (Timeline) · Czech Republic · Denmark (Faroe Islands) · Estonia · Finland · France · Germany (North Rhine-Westphalia) · Georgia · Greece · Hungary · Iceland · Ireland · Italy · Kazakhstan · Kosovo · Latvia · Liechtenstein · Lithuania · Luxembourg · Malta · Moldova · Monaco · Montenegro · Netherlands · North Macedonia · Norway · Poland · Portugal · Romania · Russia · San Marino · Serbia · Slovakia · Slovenia · Spain · Sweden · Switzerland · Turkey (Timeline) · Ukraine · United Kingdom (Gibraltar · Guernsey · Jersey · Scotland) · Vatican City
North America		Antigua and Barbuda · Aruba · Bahamas · Barbados · Canada (Ontario · British Columbia) · Cayman Islands · Costa Rica · Cuba · Curaçao · Dominican Republic · Greenland · Guadeloupe · Guatemala · Honduras · Jamaica · Martinique · Mexico · Panama · Saint Barthélemy · Saint Lucia · Saint Martin · Saint Vincent and the Grenadines · Trinidad and Tobago
	United States	Alabama · Arkansas · California · Colorado · Connecticut · Delaware · Florida · Georgia · Guam · Illinois · Indiana · Kansas · Kentucky · Louisiana · Maryland · Massachusetts · Michigan · Minnesota · Mississippi · Nevada · New Jersey · New Mexico · New York · North Carolina · Ohio · Oregon · Pennsylvania · Puerto Rico · South Carolina · Tennessee · Texas · U.S. Virgin Islands · Utah · Virginia · Washington · Washington, D.C. · Wisconsin
Oceania		Australia · French Polynesia · Guam · New Zealand
South America		Argentina · Brazil · Bolivia · Chile · Colombia · Ecuador · French Guiana · Guyana · Paraguay · Peru · Suriname · Uruguay · Venezuela
International conveyances		Cruise ships (*Diamond Princess* · *Grand Princess* · MS *River Anuket* · MS *Westerdam* · *World Dream*)
Lockdowns		Austria (Tyrol) · China (Hubei · elsewhere) · Czech Republic · Italy · Malaysia · Spain · Venezuela

Institutions [hide]

Centers for Disease Control	China · Europe · South Korea · United States · India
Hospitals and related	Central Hospital of Wuhan · Dabie Mountain Regional Medical Centre · Huoshenshan Hospital · Leishenshan Hospital · Xinjia Express Hotel
Organisations	Coalition for Epidemic Preparedness Innovations · National Health Commission · World Health Organization · Wuhan Institute of Virology · National Institute of Virology

Issues [hide]

Problems and restrictions	Domestic responses · Evacuations · Human rights · Misinformation · Public Health Emergency of International Concern · Travel restrictions
Treatment	Disease testing · Treatment research · Vaccine research
Socio-economic impact	Aviation · Cinema · Religion · Politics · Television · Education · Event cancellations · Financial impact (Oil price war · Global stock market crash) · Sports (2020 Summer Olympics) · Xenophobia and racism

People [hide]

Medical professionals/ whistleblowers		Ai Fen · Li Wenliang · Liu Wen · Corona Rintawan · Xie Linka
Researchers		Neil Ferguson · George F. Gao · Guan Yi · Li Lanjuan · Ma Xiaowei · Shi Zhengli · Yuen Kwok-yung · Zeng Guang · Zhong Nanshan
Officials		Tedros Adhanom (WHO Director-General) · Stella Kyriakidou (EU Commissioner for Health and Food Safety) · Naomichi Suzuki (Governor of Hokkaidō)
	China	Chen Yixin · Jiang Chaoliang→Ying Yong (Party Secretary of Hubei) · Li Keqiang (Premier of China, in-charge of prevention) · Ma Guoqiang→Wang Zhonglin (Party Secretary of Wuhan) · Wang Xiaodong (Governor of Hubei) · Zhou Xianwang (Mayor of Wuhan)
	Italy	Giuseppe Conte (Prime Minister) · Roberto Speranza (Minister of Health)
Others		Chen Qiushi
Infected	Deaths	Nicolas Alfonsi · Hashem Bathaie Golpayegani · Giovanni Battista Rabino · Italo De Zan · Duan Zhengcheng · Vittorio Gregotti · Hadi Khosroshahi · Hong Ling · Lee Cha-su · Liang Wudong · Liu Fan · Mohammad Mirmohammadi · Marcelo Peralta · Fatemeh Rahbar · Fariborz Raisdana · Mohammad-Reza Rahchamani · Nasser Shabani · Hossein Sheikholeslam · Wang Xianliang · Yang Xiaobo

Data [hide]

Global	Cases, deaths, recoveries by country · Daily data by country · WHO situation report · World map by city · Hospital beds by country
Africa	South Africa (chart)
Asia	Mainland China (aggregate · chart) · India (chart) · Iran (chart) · Israel (chart) · Japan (chart · ship) · Philippines (chart · summary · areas of quarantine) · Russia (chart) · Singapore (chart) · South Korea (chart) · Thailand (chart) · United Arab Emirates (chart)
Europe	Austria (chart) · Belarus (chart) · Belgium (chart) · Denmark (chart) · Estonia (chart) · Finland (chart) · France (chart) · Germany (chart) · Hungary (chart) · Iceland (chart) · Ireland (chart) · Italy (chart) · Latvia (chart) · Lithuania (chart) · Moldova (chart) · Netherlands (chart) · Norway (chart) · Poland (chart) · Romania (chart) · Russia (chart) · Serbia (chart) · Slovakia (chart) · Slovenia (chart) · Spain (chart) · Sweden (chart) · Switzerland (chart) · Ukraine (chart) · United Kingdom (Scotland · chart)
North America	Canada (chart) · United States (chart)
Oceania	Australia (chart)
South America	Brazil (chart)

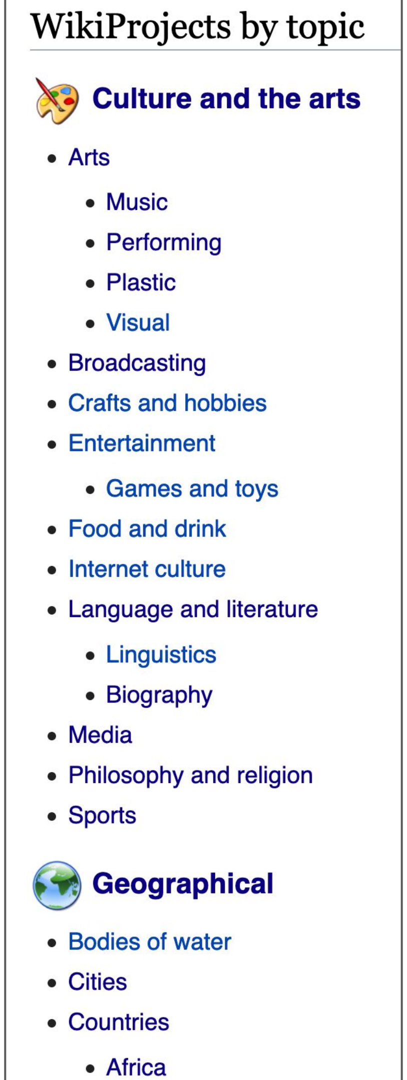
Category

WikiProject Directory

https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Directory

- ~60 topics
- Good coverage while still not too extensively over-labeled
- Manageable taxonomy generated by Wikipedians that we can easily adjust as needed for topic modeling -- e.g., WikiProject COVID-19 -> STEM.Medicine

See Also:
https://www.mediawiki.org/wiki/ORES#Topic_routing



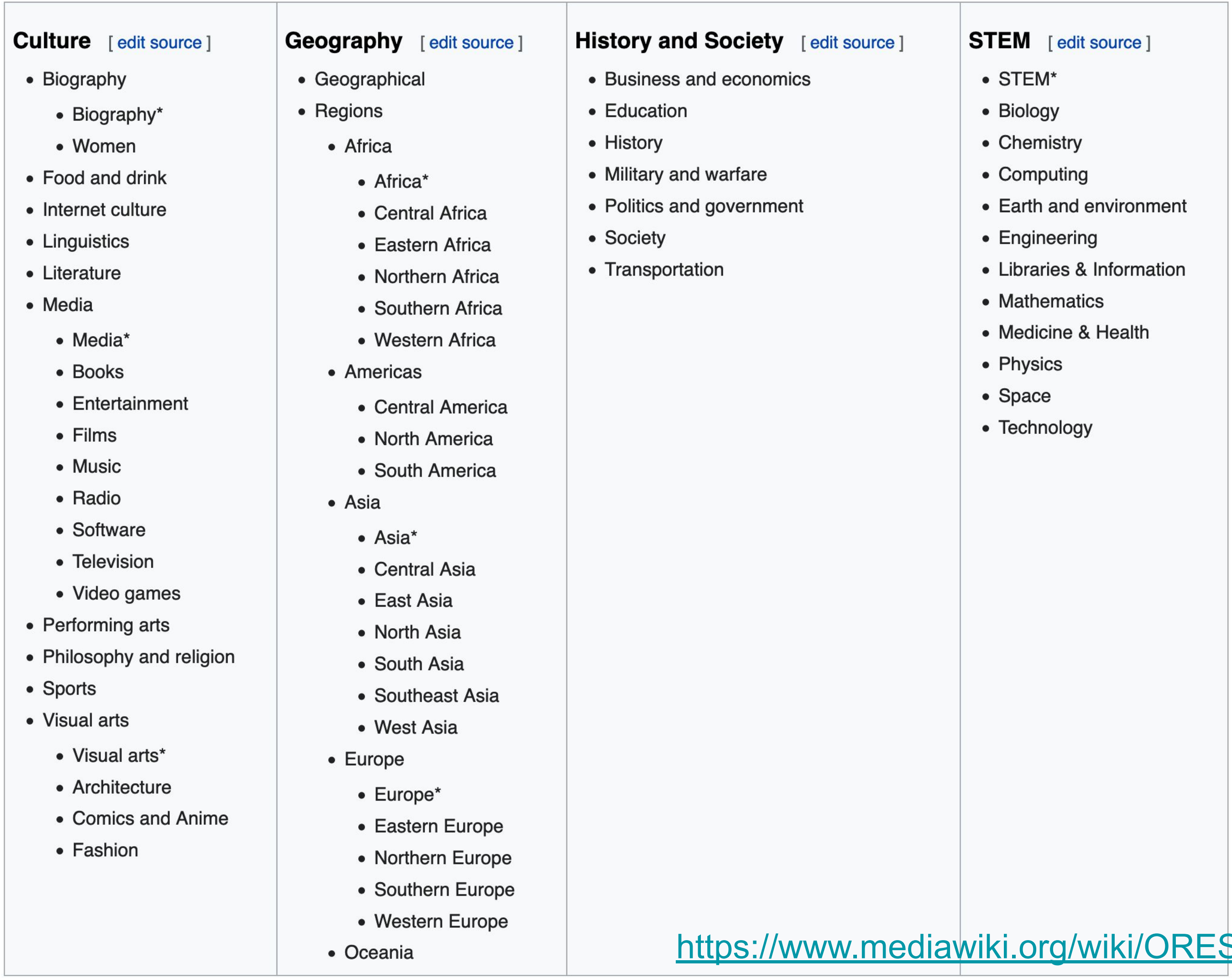
WikiProjects by topic

Culture and the arts

- Arts
  - Music
  - Performing
  - Plastic
  - Visual
- Broadcasting
- Crafts and hobbies
- Entertainment
  - Games and toys
- Food and drink
- Internet culture
- Language and literature
  - Linguistics
  - Biography
- Media
- Philosophy and religion
- Sports

Geographical

- Bodies of water
- Cities
- Countries
  - Africa



Culture [edit source]

- Biography
  - Biography*
  - Women
- Food and drink
- Internet culture
- Linguistics
- Literature
- Media
  - Media*
  - Books
  - Entertainment
  - Films
  - Music
  - Radio
  - Software
  - Television
  - Video games
- Performing arts
- Philosophy and religion
- Sports
- Visual arts
  - Visual arts*
  - Architecture
  - Comics and Anime
  - Fashion

Geography [edit source]

- Geographical
- Regions
  - Africa
    - Africa*
    - Central Africa
    - Eastern Africa
    - Northern Africa
    - Southern Africa
    - Western Africa
  - Americas
    - Central America
    - North America
    - South America
  - Asia
    - Asia*
    - Central Asia
    - East Asia
    - North Asia
    - South Asia
    - Southeast Asia
    - West Asia
  - Europe
    - Europe*
    - Eastern Europe
    - Northern Europe
    - Southern Europe
    - Western Europe
  - Oceania

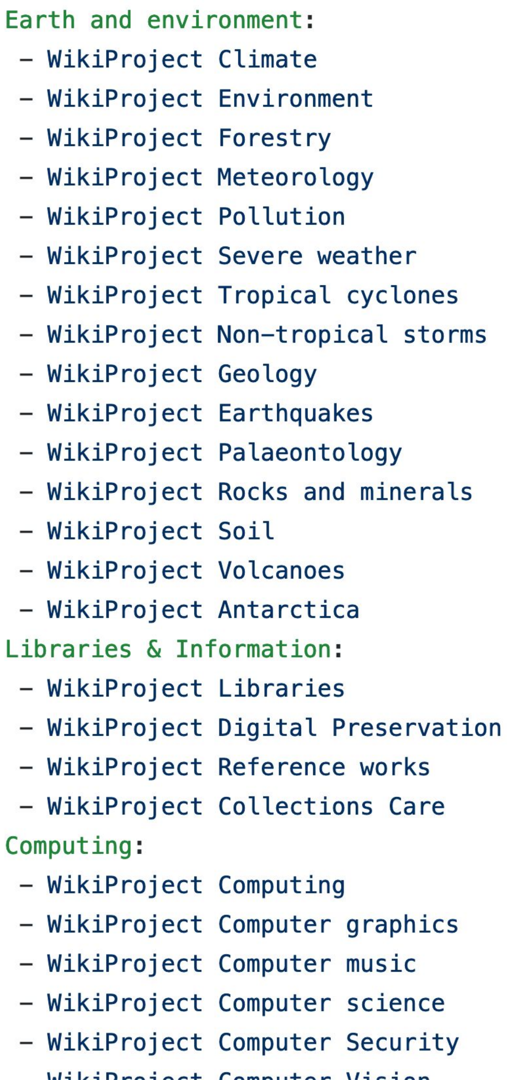
History and Society [edit source]

- Business and economics
- Education
- History
- Military and warfare
- Politics and government
- Society
- Transportation

STEM [edit source]

- STEM*
- Biology
- Chemistry
- Computing
- Earth and environment
- Engineering
- Libraries & Information
- Mathematics
- Medicine & Health
- Physics
- Space
- Technology

Source: https://www.mediawiki.org/wiki/ORES/Articletopic#Taxonomy

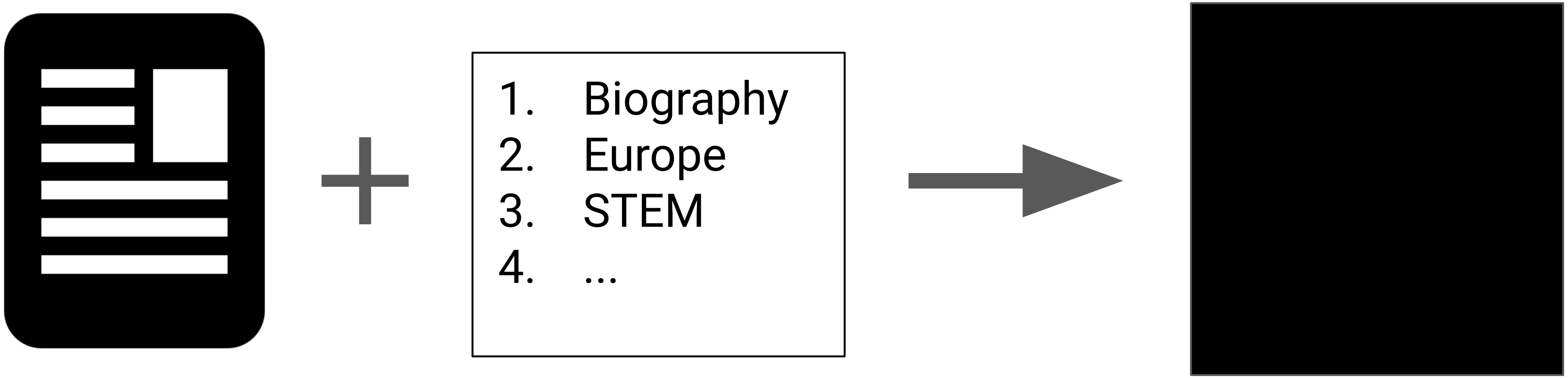


```
Earth and environment:
 - WikiProject Climate
 - WikiProject Environment
 - WikiProject Forestry
 - WikiProject Meteorology
 - WikiProject Pollution
 - WikiProject Severe weather
 - WikiProject Tropical cyclones
 - WikiProject Non-tropical storms
 - WikiProject Geology
 - WikiProject Earthquakes
 - WikiProject Palaeontology
 - WikiProject Rocks and minerals
 - WikiProject Soil
 - WikiProject Volcanoes
 - WikiProject Antarctica
Libraries & Information:
 - WikiProject Libraries
 - WikiProject Digital Preservation
 - WikiProject Reference works
 - WikiProject Collections Care
Computing:
 - WikiProject Computing
 - WikiProject Computer graphics
 - WikiProject Computer music
 - WikiProject Computer science
 - WikiProject Computer Security
 - WikiProject Computer Vision
```

Source:
https://github.com/halfak/wikitax/blob/master/taxonomies/wikiproject/halfak_20191202/taxonomy.yaml

Challenge Two: New Content

Training



1. Biography
2. Europe
3. STEM
4. ...

Wikipedia Article
(actually the average of word vectors for tokens in an article)

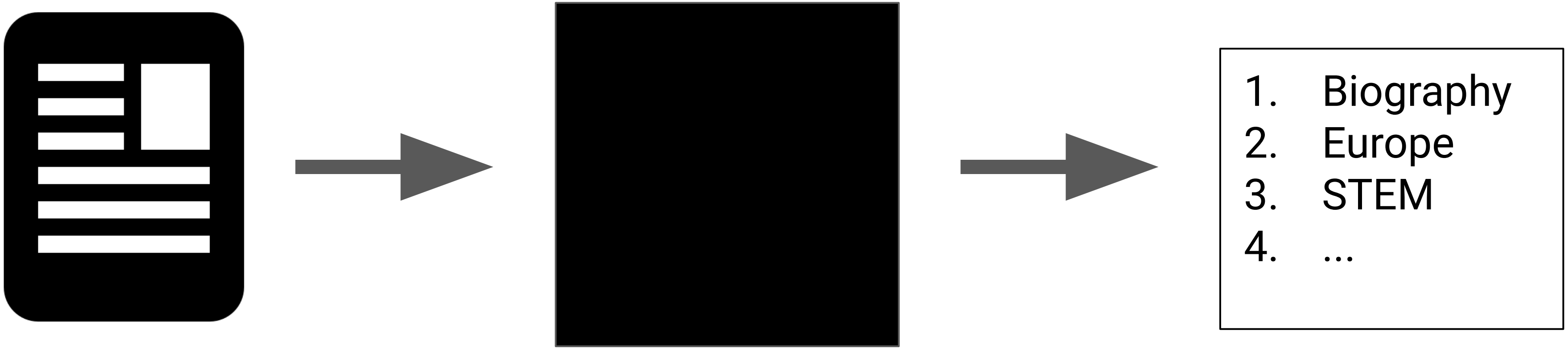
WikiProject-based Topics

Machine-Learning Model

Wikipedia Article Image Source: https://commons.wikimedia.org/wiki/File:OOjs_UI_icon_article-ltr.svg

Data: https://figshare.com/articles/Wikipedia_Articles_and_Associated_WikiProject_Templates/10248344

Prediction



1. Biography
2. Europe
3. STEM
4. ...

Wikipedia Article
(actually the average of word vectors for tokens in an article)

Machine-Learning Model

Topic Predictions

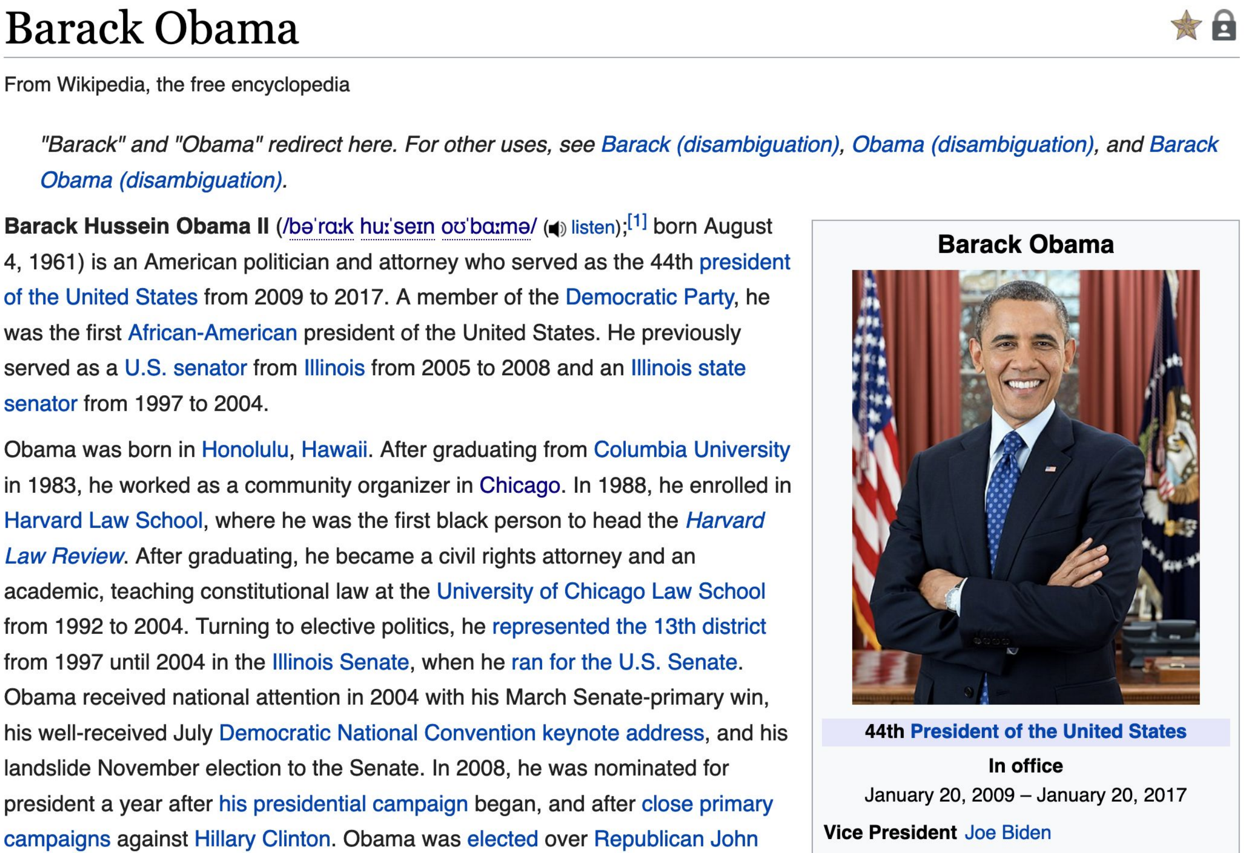
Wikipedia Article Image Source:
https://commons.wikimedia.org/wiki/File:OOjs_UI_icon_article-ltr.svg

Try it out:
https://ores.wmflabs.org/ui/

Available APIs

API	ORES articletopic
Input Data	Article text
Languages	Arabic, Czech, English, Korean, Vietnamese
Try it out:	https://ores.wmflabs.org/ui/

Barack Obama - ORES Model



Barack Obama

From Wikipedia, the free encyclopedia

"Barack" and "Obama" redirect here. For other uses, see Barack (disambiguation), Obama (disambiguation), and Barack Obama (disambiguation).

Barack Hussein Obama II (/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/ (🔊 listen);[1] born August 4, 1961) is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African-American president of the United States. He previously served as a U.S. senator from Illinois from 2005 to 2008 and an Illinois state senator from 1997 to 2004.

Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago. In 1988, he enrolled in Harvard Law School, where he was the first black person to head the *Harvard Law Review*. After graduating, he became a civil rights attorney and an academic, teaching constitutional law at the University of Chicago Law School from 1992 to 2004. Turning to elective politics, he represented the 13th district from 1997 until 2004 in the Illinois Senate, when he ran for the U.S. Senate. Obama received national attention in 2004 with his March Senate-primary win, his well-received July Democratic National Convention keynote address, and his landslide November election to the Senate. In 2008, he was nominated for president a year after his presidential campaign began, and after close primary campaigns against Hillary Clinton. Obama was elected over Republican John

Barack Obama

44th President of the United States

In office

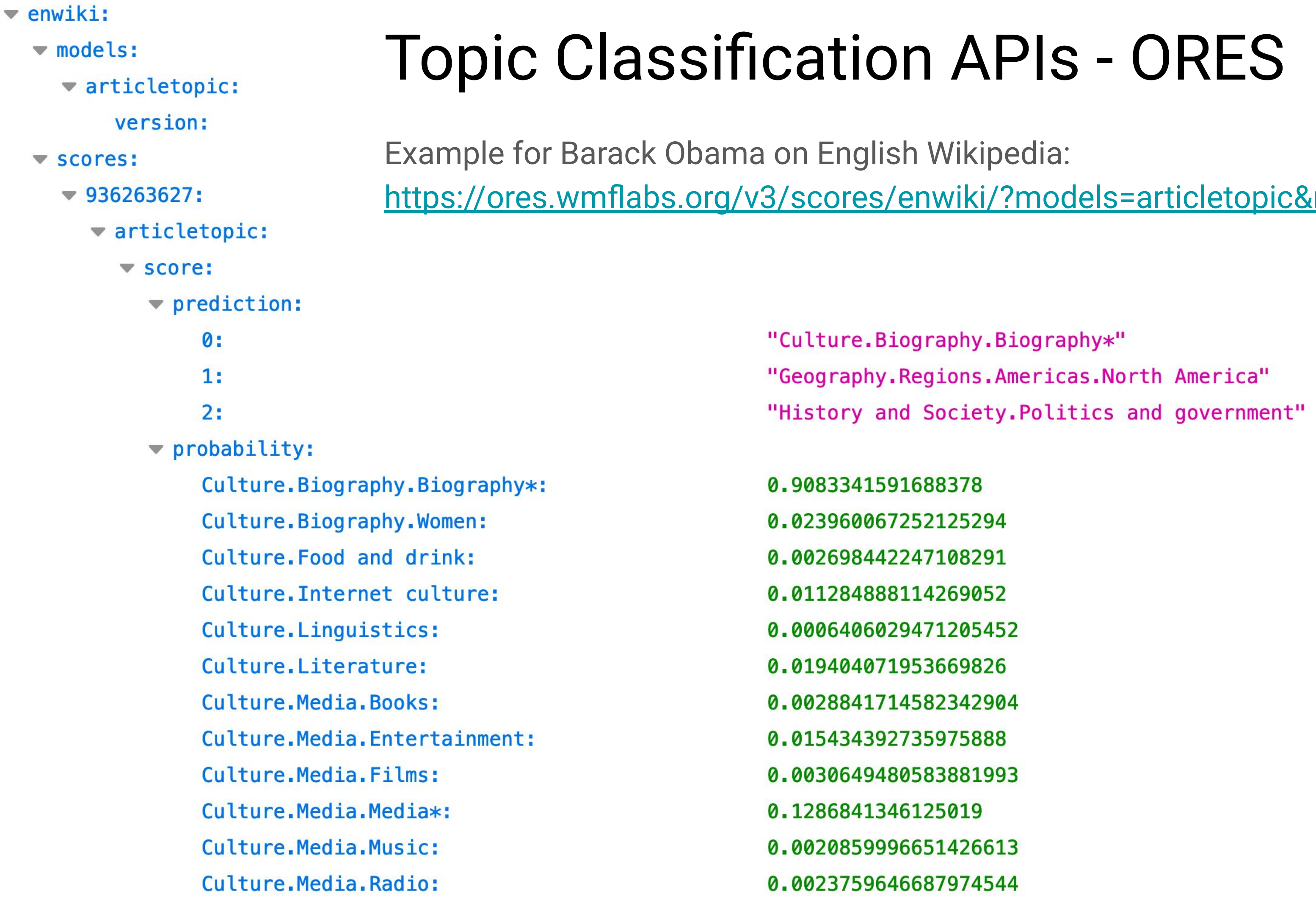
January 20, 2009 – January 20, 2017

Vice President Joe Biden

Source:
https://en.wikipedia.org/wiki/Barack_Obama

Topic Classification APIs - ORES

Example for Barack Obama on English Wikipedia: https://ores.wmflabs.org/v3/scores/enwiki/?models=articletopic&revids=936263627



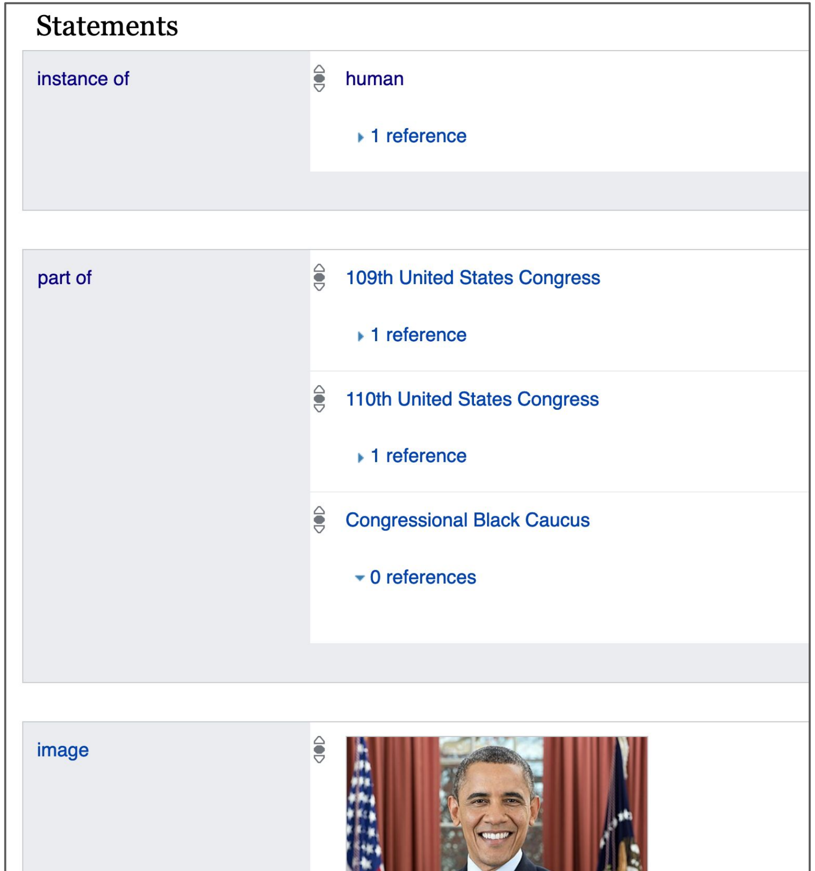
```
▼ enwiki:
  ▼ models:
    ▼ articletopic:
        version:
  ▼ scores:
    ▼ 936263627:
      ▼ articletopic:
        ▼ score:
          ▼ prediction:
              0:                                            "Culture.Biography.Biography*"
              1:                                            "Geography.Regions.Americas.North America"
              2:                                            "History and Society.Politics and government"
          ▼ probability:
              Culture.Biography.Biography*:                 0.9083341591688378
              Culture.Biography.Women:                      0.023960067252125294
              Culture.Food and drink:                       0.002698442247108291
              Culture.Internet culture:                     0.011284888114269052
              Culture.Linguistics:                          0.0006406029471205452
              Culture.Literature:                           0.019404071953669826
              Culture.Media.Books:                          0.0028841714582342904
              Culture.Media.Entertainment:                  0.015434392735975888
              Culture.Media.Films:                          0.0030649480583881993
              Culture.Media.Media*:                         0.1286841346125019
              Culture.Media.Music:                          0.0020859996651426613
              Culture.Media.Radio:                          0.0023759646687974544
```

Challenge Three: Multilingual

Available APIs

API	ORES articletopic	Experimental Wikidata
Input Data	Article text	Wikidata statements and Identifiers
Languages	Arabic, Czech, English, Korean, Vietnamese	All (if there's a Wikidata item)
Try it out:	https://ores.wmflabs.org/ui/	https://tools.wmflabs.org/wiki-topic/

Barack Obama - Wikidata Model



Statements

instance of

human

1 reference

part of

109th United States Congress

1 reference

110th United States Congress

1 reference

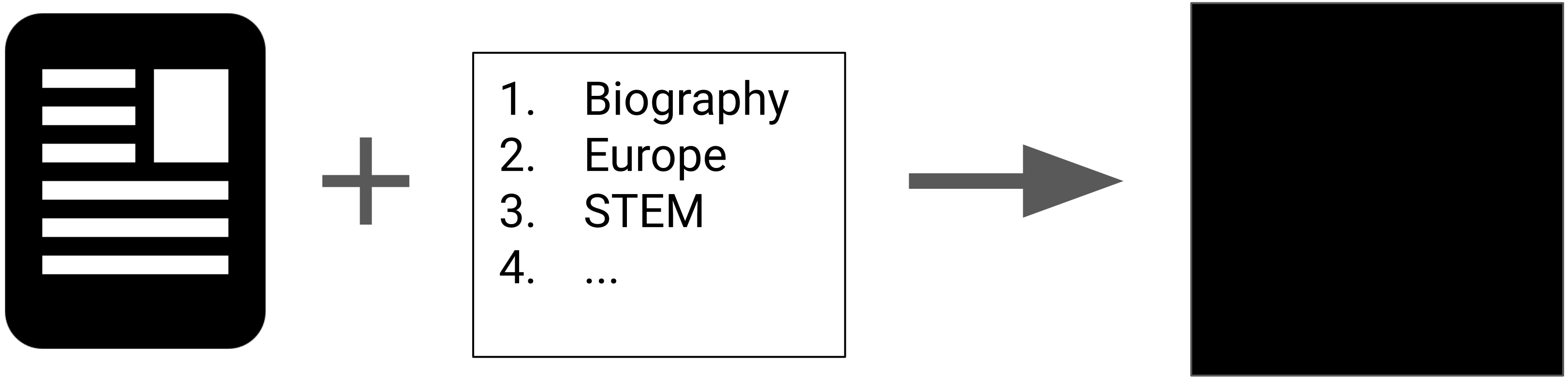
Congressional Black Caucus

0 references

image

Source:
https://www.wikidata.org/wiki/Q76

Training



1. Biography
2. Europe
3. STEM
4. ...

Wikidata Item
(actually the average of entity vectors for properties/values in an item)

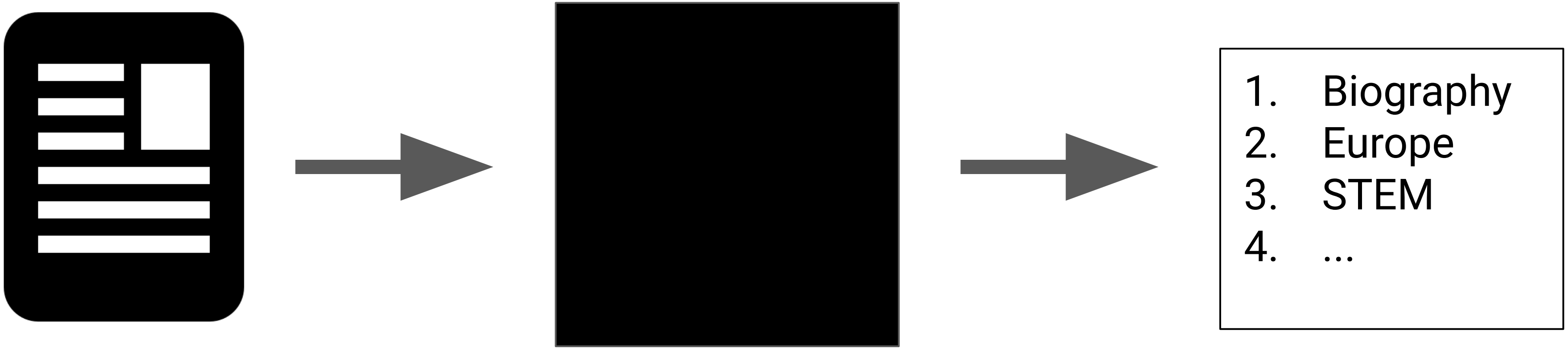
WikiProject-based Topics

Machine-Learning Model

Wikipedia Article Image Source:
https://commons.wikimedia.org/wiki/File:OOjs_UI_icon_article-ltr.svg

Data:
https://figshare.com/articles/Wikipedia_Articles_and_Associated_WikiProject_Templates/10248344

Prediction



1. Biography
2. Europe
3. STEM
4. ...

Wikidata Item
(actually the average of entity vectors for properties/values in an item)

Machine-Learning Model

Topic Predictions

Wikipedia Article Image Source: https://commons.wikimedia.org/wiki/File:OOjs_UI_icon_article-ltr.svg

Try it out: https://tools.wmflabs.org/wiki-topic/

Further Challenges, Improvements, etc.

Further Challenges, Improvements, etc.

- How do we scale to many languages while still capturing the richness of articles for a language edition?
- Is the English WikiProject Taxonomy applicable to other languages?
- What about more specific topics or topics outside of the taxonomy?

Thank you! Questions?

Me: Isaac (WMF) -- isaac@wikimedia.org

With plenty of support from: Martin Gerlach (WMF), Aaron Halfaker (WMF), Diego Sáez-Trumper (WMF), Growth Team (WMF), and many volunteers

- ORES Topic Model: https://www.mediawiki.org/wiki/ORES#Topic_routing
- Wikidata Topic Model: https://tools.wmflabs.org/wiki-topic/
- Wikipedia Categories: https://www.mediawiki.org/wiki/Wikimedia_Research/Showcase#March_2018
- WikiProject COVID-19: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_COVID-19