

# Data Quality Days

8-15 September 2021

## Bibliometric-Enhanced Information Retrieval: A new alternative for the validation and enrichment of Wikidata Statements



Data Engineering and Semantics  
هندسة البيانات و دلالاتها



جامعة صفاقس  
University of Sfax

Houcemeddine Turki

Medical Student, Faculty of Medicine of Sfax  
Research Assistant, Data Engineering and Semantics Research Unit  
Member, Wiki Indaba Steering Committee  
Vice-Chair, Wikimedia Tunisia User Group  
Board Member, Wikimedia and Libraries User Group



# About Us

Data Engineering and Semantics  
Research Unit and Wikimedia Tunisia

# University of Sfax

Located in Tunisia, North Africa (270 km from Tunis)

Major University in Tunisia

Among the best universities in Africa in Computer Science Research



# Team



- » Houcemeddine Turki
- Research Assistant
  - Medical Student



- » Mohamed Ali Hadj Taieb
- Senior Researcher
  - Assistant Professor



- » Mohamed Ben Aouicha
- Head of Research Unit
  - Associate Professor

## Wikimedia Tunisia

- » Wikimedia Regional User Group from Tunisia
- » Created in May 2014
- » Tries to enhance and diversify the Wikimedia Movement in Tunisia
- » Organizes campaigns and initiatives about Tunisia-related topics

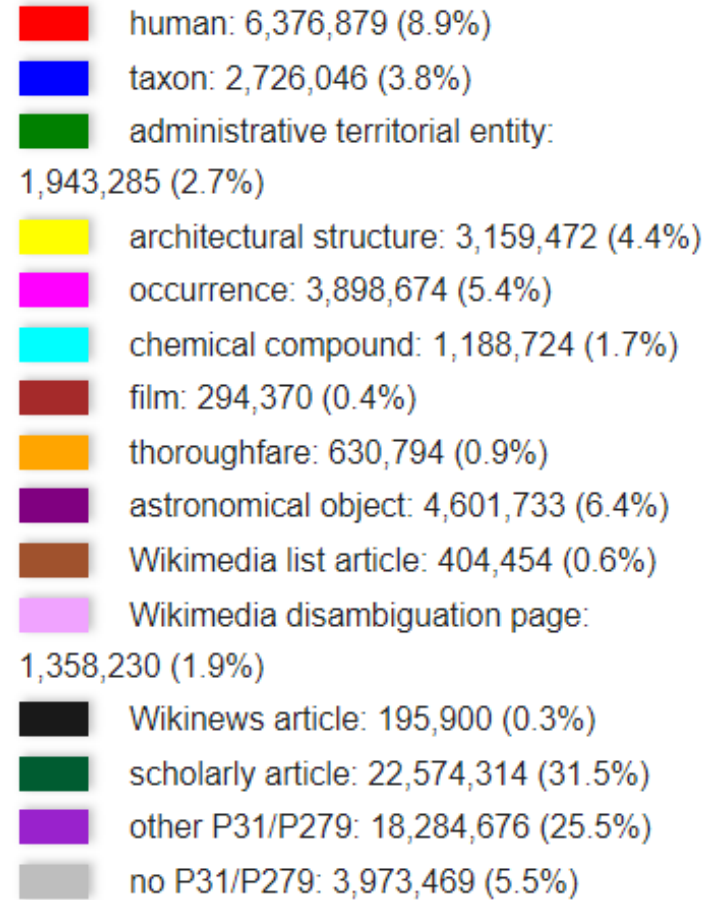
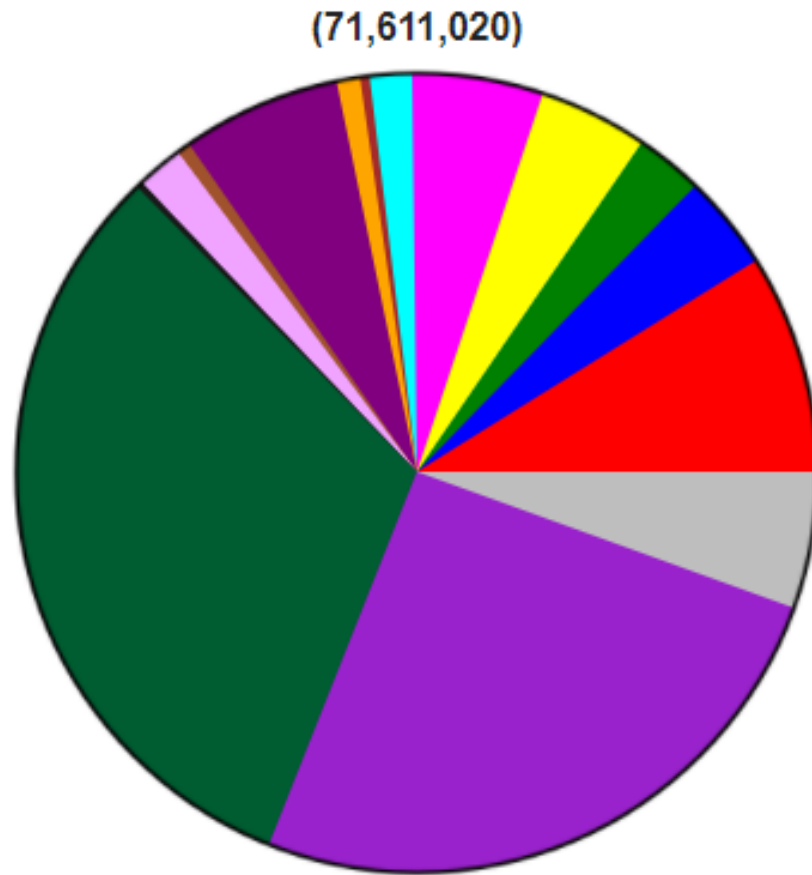




# Introduction

The State of Wikidata

# Wikidata items



Module:Statistical data/by project/classes, 2020-02-16

# Biomedical Knowledge in Wikidata as of March 2019

Biomedical entity (P31)	Number of items	Number of properties		Number of properties per item		Percentage of referenced data
		With references	Without references	With references	Without references	
Drugs	2713	75,259	35,302	27.7	13.0	68.1%
Drug classes	1043	16,855	10,537	16.2	10.1	61.5%
Human enzymes	89	1234	386	13.9	4.3	76.2%
Diseases	11,447	152,622	57,689	13.3	5.0	72.6%
<b>Human genes</b>	<b>58,691</b>	<b>671,282</b>	<b>12,949</b>	<b>11.4</b>	<b>0.2</b>	<b>98.1%</b>
<b>Human proteins</b>	<b>25,482</b>	<b>265,684</b>	<b>27,825</b>	<b>10.4</b>	<b>1.1</b>	<b>90.5%</b>
Human muscles	351	1690	2136	4.8	6.1	44.2%
Pains	171	725	858	4.2	5.0	45.8%
Syndromes	72	173	350	2.4	4.9	33.1%
Human arteries	418	964	2383	2.3	5.7	28.8%
Human joints	67	151	535	2.3	8.0	22.0%
Human bones	102	233	1119	2.3	11.0	17.2%
Human nerves	335	738	1738	2.2	5.2	29.8%
Human veins	220	478	1081	2.2	4.9	30.7%
Medical specialties	248	512	2069	2.1	8.3	19.8%
Therapies	487	931	2312	1.9	4.7	28.7%
Human ligaments	46	56	201	1.2	4.4	21.8%
Surgical procedures	244	261	1099	1.1	4.5	19.2%
<b>Overall</b>	<b>102,226</b>	<b>1,189,848</b>	<b>160,569</b>	<b>11.6</b>	<b>1.6</b>	<b>88.1%</b>





## Several inconsistencies in Wikidata

```
SELECT ?disease ?diseaseLabel ?drug ?drugLabel WHERE {  
  ?disease wdt:P2176 ?drug.  
  ?disease wdt:P31 wd:Q12140.  
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }  
}
```

disease	diseaseLabel	drug	drugLabel
<a href="#">Q422482</a>	gentamicin	<a href="#">Q217519</a>	bubonic plague
<a href="#">Q481757</a>	aceclofenac	<a href="#">Q474959</a>	myalgia
<a href="#">Q481757</a>	aceclofenac	<a href="#">Q683498</a>	arthralgia
<a href="#">Q7335107</a>	Rintatolimod	<a href="#">Q209733</a>	chronic fatigue syndrome

## Projects for the validation of Wikidata

- » Shape Expressions
- » SHACL
- » Logical Constraints in SPARQL
- » Reference Island
- » Property Constraints
- » ORES



# Bibliographic Metadata

Introduction to Citation Indexes

## Title

## Comparison of Rapid Antigen Tests for COVID-19

Seiya <sup>1</sup>, Michiko Koga <sup>2,3</sup>, Osamu Akasaka <sup>4</sup>, Ichiro Nakachi <sup>5</sup>, Hidefumi Koh <sup>6</sup>, Kenji Maeda <sup>7</sup>, Eisuke Adachi <sup>3</sup>, Makoto Saito <sup>2,3</sup>, Hiroyuki Nagai <sup>3</sup>, Kazuhiko Ikeuchi <sup>2,3</sup>, Takayuki Ogura <sup>8</sup>, Rie Baba <sup>5</sup>, Kensuke Fujita <sup>8</sup>, Takahiro Fukui <sup>6</sup>, Fumimaro Ito <sup>6</sup>, Shin-Ichiro Hattori <sup>7</sup>, Kei Yamamoto <sup>9</sup>, Takato Naito <sup>9</sup>, Yukihiro Ito <sup>1</sup>, Atsuhiko Yasuhara <sup>1</sup>, Michiko Ujje <sup>1</sup>, Shinya Yamada <sup>1</sup>, Mutsumi Ito <sup>1</sup>, Hiroaki Mitsuya <sup>7</sup>, Norio Omagari <sup>9</sup>, Hiroshi Yotsuyanagi <sup>2,3</sup>, Kiyoko Iwatsuki-Hori <sup>1</sup>, Masahito Ito <sup>1</sup>, Yoshihiro Kawaoka <sup>1,10,11</sup>

## Authors

Affiliations 

## Affiliations

- 1 Division of Virology, Department of Microbiology and Immunology, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan.
- 2 Division of Infectious Diseases, Advanced Clinical Research Center, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan.
- 3 Department of Infectious Diseases and Applied Immunology, IMSUT Hospital of Institute of Medical Science, the University of Tokyo, Tokyo 108-8639, Japan.
- 4 Emergency Medical Center, Fujisawa City Hospital, Kanagawa 251-0292, Japan.
- 5 Pulmonary Division, Department of Internal Medicine, Saitama Utsunomiya Hospital, Tochigi 321-0974, Japan.
- 6 Division of Pulmonary Medicine, Department of Internal Medicine, Tachikawa Hospital, Tokyo 190-8531, Japan.

## Affiliations

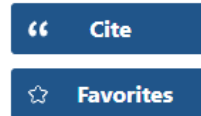
## References

1. Zhu N., Zhang D., Wang W., Li X., Yang B., Song J., Zhao X., Huang B., Li X., Yang B., et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med.* 2020;382:727–733. doi: 10.1056/NEJMoa2001017. - DOI - PMC - PubMed
2. Sethuraman N., Jeremiah S.S., Ryo A. Interpreting Diagnostic Tests for SARS-CoV-2. *JAMA.* 2020;323:2249–2251. doi: 10.1001/jama.2020.8259. - DOI - PubMed
3. Nagura-Ikeda M., Imai K., Tabata S., Miyoshi K., Murahara N., Mizuno T., Horiuchi M., Kato K., Imoto Y., Iwata M., et al. Clinical Evaluation of Self-Collected Saliva by Quantitative Reverse Transcription-PCR (RT-qPCR), Direct RT-qPCR, Reverse Transcription-Loop-Mediated Isothermal Amplification, and a Rapid Antigen Test to Diagnose COVID-19. *J. Clin. Microbiol.* 2020;58:e01438-20. doi: 10.1128/JCM.01438-20. - DOI - PMC - PubMed

## References



ACTIONS



SHARE



PAGE NAVIGATION

&lt; Title &amp; authors

Abstract

Conflict of interest statement

Figures

Free PMC article

## Abstract

Reverse transcription-quantitative PCR (RT-qPCR)-based tests are widely used to diagnose coronavirus disease 2019 (COVID-19). As a result that these tests cannot be done in local clinics where RT-qPCR testing capability is lacking, rapid antigen tests (RATs) for COVID-19 based on lateral flow immunoassays are used for rapid diagnosis. However, their sensitivity compared with each other and with RT-qPCR and infectious virus isolation has not been examined. Here, we compared the sensitivity among four RATs by using severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) isolates and several types of COVID-19 patient specimens and compared their sensitivity with that of RT-qPCR and infectious virus isolation. Although the RATs read the samples containing large amounts of virus as positive, even the most sensitive RAT read the samples containing small amounts of virus as negative. Moreover, all RATs tested failed to detect viral antigens in several specimens from which the virus was isolated. The current RATs will likely miss some COVID-19 patients who are shedding infectious SARS-CoV-2.

**Keywords:** COVID-19; SARS-CoV-2; diagnosis; rapid antigen test.

## Keywords

## Publication types

- > Comparative Study
- > Research Support, N.I.H., Extramural
- > Research Support, Non-U.S. Gov't

## Publication Type

## MeSH terms

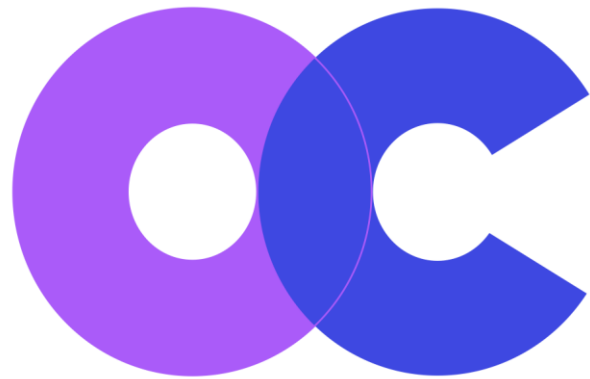
- > Antigens, Viral / analysis\*
- > COVID-19 / diagnosis\*
- > COVID-19 Serological Testing / methods\*
- > False Negative Reactions
- > Humans
- > Immunoassay
- > Point-of-Care Systems\*

## Controlled Keywords



# Bibliographic Databases

» OpenCitations



» DBLP



» PubMed



# Multiple ways for parsing bibliographic databases



## Find

Advanced Search  
Clinical Queries  
Single Citation Matcher



## Download

E-utilities API  
FTP  
Batch Citation Matcher



## Explore

MeSH Database  
Journals



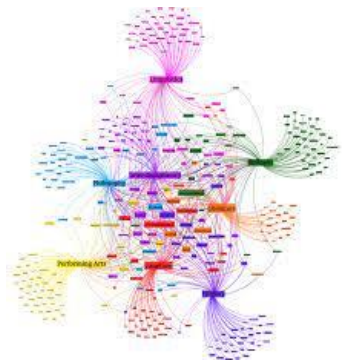


# Usefulness of Bibliographic Information

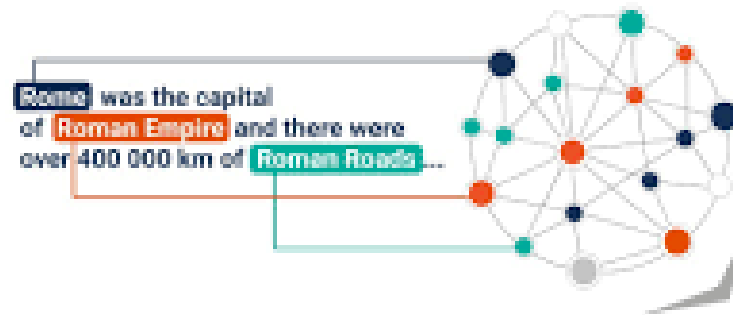
Usage in Knowledge Graph Enrichment  
and Validation

# Title and Abstract

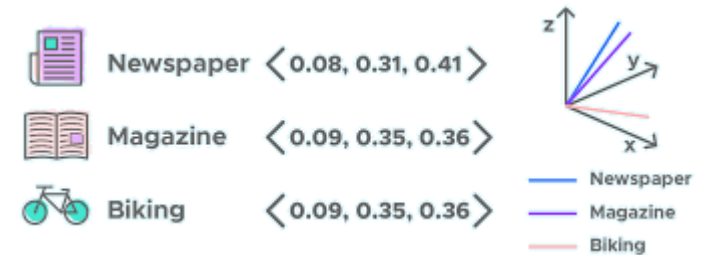
## » Topic Modelling



## » Semantic Annotation




## » Word and Graph Embeddings



Useful Resources to generate *main subject* (P921) statements  
Adding References to unsupported Wikidata Statements

main subject

Wikidata  edit

0 references

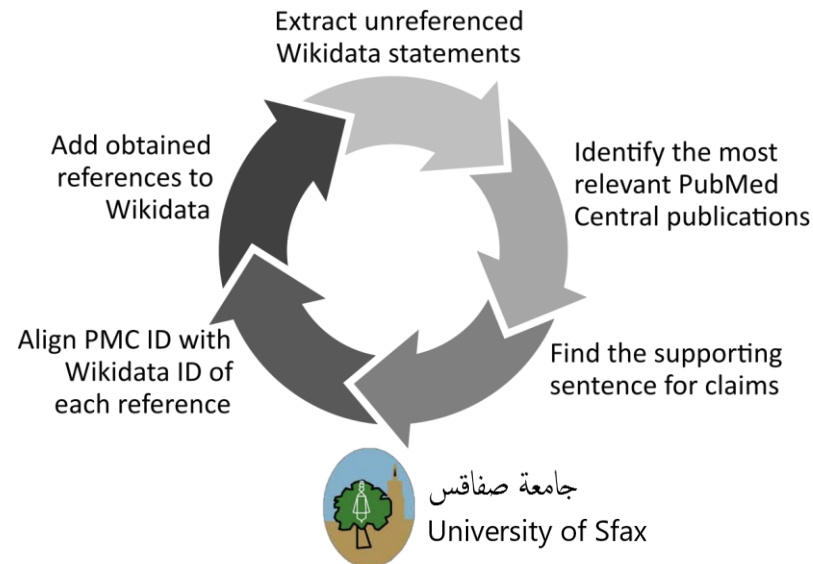
+ add reference



# Keywords and Controlled Keywords

- » Simply used for generating *main subject* (P921) statements.
- » Co-word analysis can be useful to generate statements in the form of triples reflecting facts about the scholarly publication.
- » Can be used alongside titles and abstracts to validate Wikidata statements.

{{{Ref<sub>B</sub>}}}



## MeSH terms

- > Antiviral Agents / adverse effects
- > Carbamates
- > Hepacivirus / genetics
- > Hepatitis C\* / complications
- > Hepatitis C\* / drug therapy
- > Hepatitis C, Chronic\* / complications
- > Hepatitis C, Chronic\* / drug therapy
- > Heterocyclic Compounds, 4 or More Rings
- > Humans
- > Liver Cirrhosis / complications
- > Liver Cirrhosis / drug therapy
- > Sofosbuvir / adverse effects
- > Sustained Virologic Response
- > Taiwan
- > Treatment Outcome



# Section and Source Title

## 3. COVID-19 pathology and immune response to SARS-CoV-2 Go to:

The lungs are exposed to thousands of liters of air daily, creating vast opportunities for airborne pathogens to enter the body [59]. Therefore, the immune system within the lungs has evolved to be highly sensitive and constantly active [60, 61, 62]. Mucus, a protective barrier in the lungs, coats the epithelial layers and entraps small particles and pathogens which are easily cleared from the body by coughing [63]. However, respiratory viruses such as coronavirus are able to permeate through this barrier. The virus infects the lung cells and triggers an immune response by recruiting cells that release inflammatory cytokines and prime T and B cells for immune response [64]. This process is intended for viral clearance; however, in some cases dysfunctional immune response occurs, causing severe damage to the lungs and eventually leading to systemic inflammation. Knowledge of the host immune response to SARS-CoV-2 is still not fully understood despite continuing research. However, clinical data obtained from SARS-CoV and MERS-CoV allows for some fundamental understanding and prediction of how the immune system will respond [65].

Section Titles can reflect the types of semantic relations that can be retrieved from the part, particularly when the publication type is review.

Source Titles can be added to Wikidata as *published in* (P1433) statements. They can also reflect the topic of papers.

published in

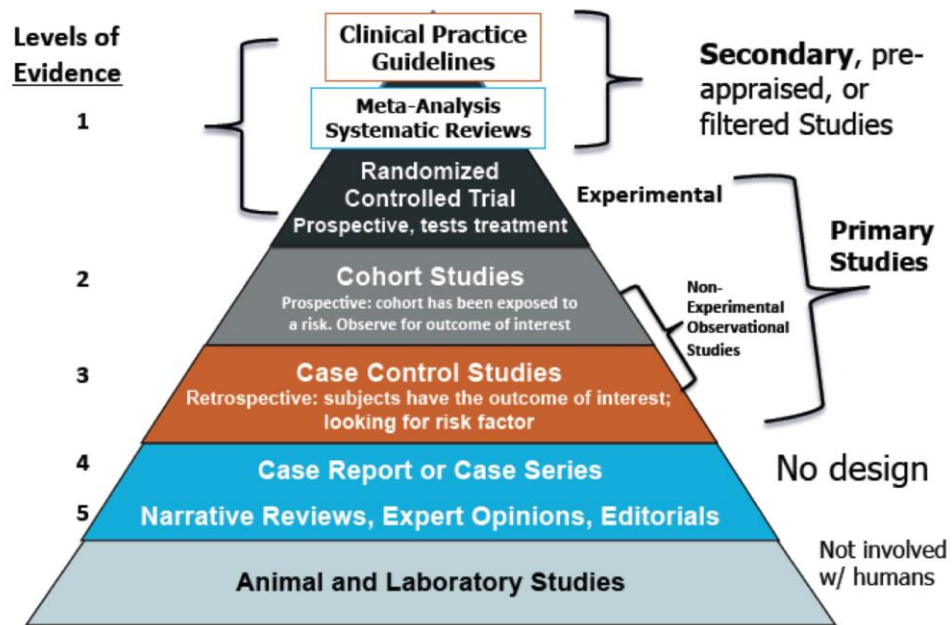


Journal of Biomedical Informatics

▶ 1 reference



# Publication Type, Publication Year and Status



# Conclusion

« **Bibliographic metadata** provide **useful** information about **scholarly publications** in a nutshell: **Findable** and **Reusable**. This can be consequently used to enrich and validate **Wikidata**»

## References:

- Turki, H., Shafee, T., Hadj Taieb, M. A., Ben Aouicha, M., Vrandečić, D., Das, D., & Hamdi, H. (2019). Wikidata: A large-scale collaborative ontological medical database. *Journal of biomedical informatics*, 99, 103292.
- Turki, H., Hadj Taieb, M. A., Ben Aouicha, M., Fraumann, G., Hauschke, C., & Heller, L. (2021). Enhancing Knowledge Graph Extraction and Validation From Scholarly Publications Using Bibliographic Metadata. *Frontiers in research metrics and analytics*, 6, 36.
- Turki, H. (2018). Citation analysis is also useful to assess the eligibility of biomedical research works for inclusion in living systematic reviews. *Journal of clinical epidemiology*, 97, 124-125.
- Turki, H., Hadj Taieb, M. A., & Ben Aouicha, M. (2018). MeSH qualifiers, publication types and relation occurrence frequency are also useful for a better sentence-level extraction of biomedical relations. *Journal of biomedical informatics*, 83, 217-218.



# Thank You



E-mail: [turkiabdelwaheb@hotmail.fr](mailto:turkiabdelwaheb@hotmail.fr)

Phone: +21629499418

Twitter: @Csisc1994

<https://dblp.org/pid/176/1531.html>

[https://www.researchgate.net/profile/Houcemeddine\\_Turki](https://www.researchgate.net/profile/Houcemeddine_Turki)

