

Wikidata and Persistent Identifiers

Arthur Smith - American Physical Society
PIDapalooza 2016



Peter Murray-Rust @petermurrayrust · Oct 9

blog.wikimedia.org/2016/10/07/new...

[#wikidata](#) one of most important
developments in scientific information this
decade 2/n [#wikifactmine](#)
[@TheContentMine](#)

FINN ÅRUP NIELSEN

Scholarly profile page constructed from queries to Wikidata Query Service. Read more on [Twitter](#).

EDUCATION

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2
			Jan 1, 1996 Jan 1, 1993 Technical University of Denmark Q wd:Q1269766 civilingeniør					Jan 1, 2001 Jan 1, 1998 Technical University of Denmark Q wd:Q1269766 Danish PhD			2
		Jan 1, 1993 Jan 1, 1990 Aarhus University School of Engineering Q wd:Q12318342									
1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2

[Edit on query.Wikidata.org](#)

LIST OF PUBLICATIONS

wd:Q26698262	data, models and neuroinformatics	2014			Brian MacWhinney, Ina Bornkessel-Schlesewsky, Michael A. Arbib
Q wd:Q26698303	Brede tools and federating online neuroinformatics databases	Jan 1, 2014	11	Neuroinformatics	Finn Årup Nielsen
Q wd:Q22329167	Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership	Dec 1, 2012	23	Journal of the Association for Information Science and Technology	Arto Lanamäki, Mohamad Mehdi, Chitu Okoli, Mostafa Mesgari, Finn Årup Nielsen

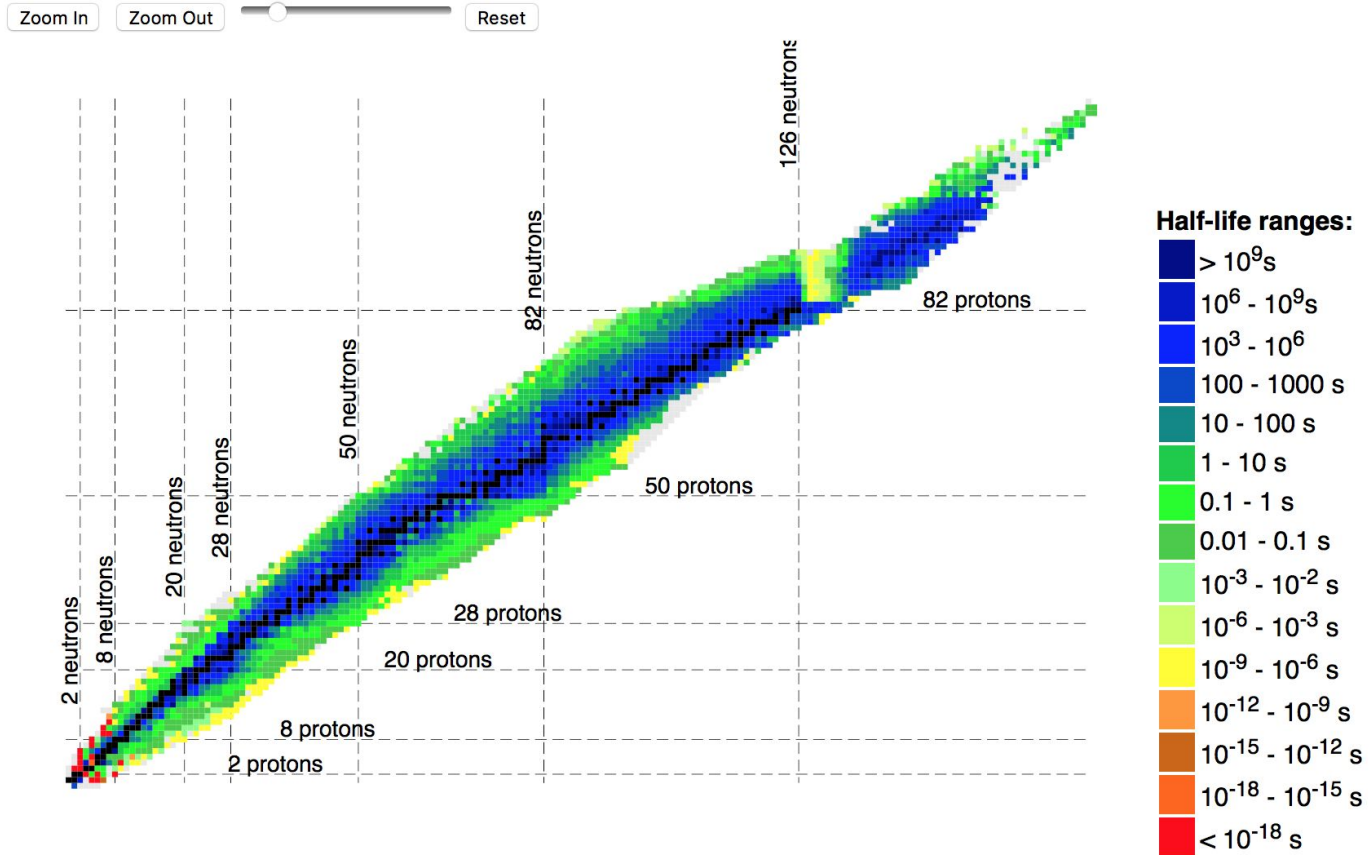
Wikidata periodic table · API · license

H hydrogen 1																He Helium 2	
Li lithium 3	Be beryllium 4											B boron 5	C carbon 6	N nitrogen 7	O oxygen 8	F fluorine 9	Ne neon 10
Na sodium 11	Mg magnesium 12											Al Aluminium 13	Si silicon 14	P phosphorus 15	S sulfur 16	Cl chlorine 17	Ar argon 18
K potassium 19	Ca Calcium 20	Sc scandium 21	Ti titanium 22	V vanadium 23	Cr chromium 24	Mn manganese 25	Fe Iron 26	Co cobalt 27	Ni nickel atom 28	Cu copper 29	Zn zinc 30	Ga gallium atom 31	Ge germanium 32	As arsenic 33	Se selenium atom 34	Br bromine 35	Kr krypton 36
Rb rubidium atom 37	Sr strontium atom 38	Y yttrium 39	Zr zirconium atom 40	Nb niobium atom 41	Mo molybde... atom 42	Tc technetium atom 43	Ru ruthenium atom 44	Rh rhodium atom 45	Pd palladium atom 46	Ag silver 47	Cd cadmium atom 48	In indium atom 49	Sn tin 50	Sb antimony atom 51	Te tellurium atom 52	I iodine 53	Xe xenon 54
Cs caesium atom 55	Ba barium 56	*	Hf hafnium atom 72	Ta tantalum atom 73	W tungsten 74	Re rhenium atom 75	Os osmium 76	Ir iridium atom 77	Pt platinum 78	Au gold 79	Hg mercury 80	Tl thallium 81	Pb lead 82	Bi bismuth atom 83	Po polonium 84	At astatine 85	Rn radon 86
Fr francium atom 87	Ra radium 88	* *	Rf rutherford... atom 104	Db dubnium 105	Sg seaborgium 106	Bh bohrium 107	Hs hassium 108	Mt meitnerium 109	Ds darmstad... atom 110	Rg roentgeni... atom 111	Cn copernici... atom 112	Uut ununtrium 113	Fl flerovium 114	Uup ununpent... atom 115	Lv livermorium 116	Uus ununsept... atom 117	Uuo ununoctium 118
Uue ununenni... atom 119	Ubn unbinilium 120	Upp unpentpe... atom 155															

*	La lanthanum 57	Ce cerium 58	Pr praseody... atom 59	Nd neodymium atom 60	Pm promethium 61	Sm samarium atom 62	Eu europium atom 63	Gd gadolinium atom 64	Tb terbium 65	Dy dysprosium atom 66	Ho holmium 67	Er erbium 68	Tm thulium atom 69	Yb ytterbium 70	Lu lutetium 71													
* *	Ac actinium 89	Th thorium 90	Pa protactini... atom 91	U uranium 92	Np neptunium atom 93	Pu plutonium atom 94	Am americium atom 95	Cm curium 96	Bk berkelium atom 97	Cf californium atom 98	Es einsteinium atom 99	Fm fermium 100	Md mendele... atom 101	No nobelium 102	Lr lawrencium 103													
* * *	Ubu unbinium 121	Ubb unbibium 122	Ubt unbitrium 123	Ubq unbiquad... atom 124	Ubp unbipenti... atom 125	Ubh unbihexium atom 126	Ubs unbisepti... atom 127	Ubo unbioctium atom 128	Ube unbiennium atom 129	Utn untrinium 130	Utu untrium 131	Utb untribium 132	Utt untriquad... atom 133	Utq untriquad... atom 134	Utp untripenti... atom 135	Uth untriheium 136	Uts untrisepti... atom 137											
	Uto untrioctium 138	Ute untriennium 139	Uqn unquadni... atom 140	Uqu unquadu... atom 141	Uqb unquadbi... atom 142	Uqt unquadtri... atom 143	Uqq unquadqu... atom 144	Uqp unquadp... atom 145	Uqh unquadh... atom 146	Uqs unquads... atom 147	Uqo unquado... atom 148	Uqe unquade... atom 149	Upn unpentni... atom 150	Upu unpentun... atom 151	Upb unpentbium 152	Upt unpentrium 153	Upq unpentqu... atom 154											

Wikidata chart of the nuclides

By half life (stable nuclides are black)



What is wikidata?

- Started October 2012
- Free (CC0), collaborative
- Multilingual (over 200 languages)
- Structured, sourced data
- Funded by Wikimedia Foundation
- Supports wikipedias & other WMF projects
- API's for fetch and update
- Query API via SPARQL
- Wide variety of services and tools
- Large active community (16,000 editors)
- 24 million “items” so far



www.wikidata.org

(English) label **Iceland** (Q189) Wikidata item ID

Description island republic in Northern Europe

Aliases Republic of Iceland I is I Island

*
*

Value

Statement

Claim

mains voltage

Property (P2884)

Qualifier

230 volt ...

frequency 50 hertz ...

1 reference

reference URL	http://www.iec.ch/worldplugs/list_by_location.htm
retrieved	10 June 2016
editor	International Electrotechnical Commission
title	World Plugs (English)

PIDs in Wikidata

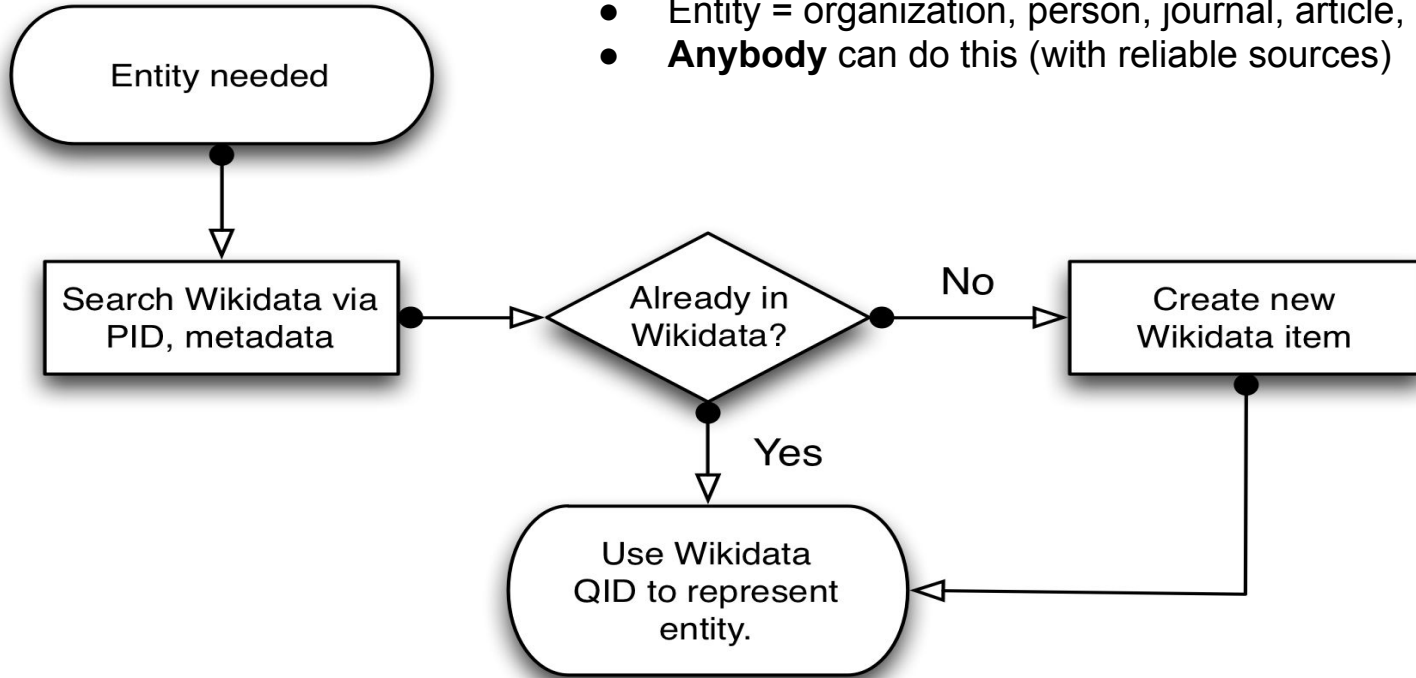
- Over 1300 properties defined with “external ID” datatype
- Includes most ID’s of interest to scholarly publishing:
 - ISBN-13 (P212)
 - ISNI (P213)
 - VIAF ID (P214)
 - ISSN (P236)
 - OCLC control number (P243)
 - LCAuth ID (P244)
 - DOI (P356)
 - MeSH ID (P486)
 - ORCID (P496)
 - arXiv ID (P818)
 - ADS bibcode (P819)
 - JSTOR article ID (P888)
 - Mathematical Reviews ID (P889)
 - Social Science Research Network ID (P893)
 - Zentralblatt MATH (P894)
 - ResearcherID (P1053)
 - Scopus Author ID (P1153)
 - Scopus EID (P1154)
 - Scopus Affiliation ID (P1155)
 - Scopus Source ID (P1156)
 - CODEN (P1159)
 - (many more!)
- New properties regularly added: community approval process usually takes a few weeks.

Peter Higgs (Q192112)

- VIAF ID: 159866219
- LCAuth ID: no2010189476
- Mathematics Genealogy Project ID: 35098
- Freebase ID: /m/01xvs9
- SUDOC authorities: 170001660
- Gran Enciclopèdia Catalana ID: 0523082
- ISNI: 0000 0001 0713 2514
- NKCR AUT ID: ntk2014824614
- NNDB people ID: 305/000169795
- Nobel prize ID: physics/laureates/2013/higgs

Wikidata can hold metadata for almost any PID

- Entity = organization, person, journal, article, ...
- **Anybody** can do this (with reliable sources)







Relationships between identifiers via Wikidata statements

- DOI <references same article as> Pubmed ID
 - Q27468554 -P356 (DOI)- 10.1128/mBio.01239-16
 - Q27468554 -P698 (Pubmed ID)- 27729507
- DOI <cites> DOI:
 - Q21065691 -P2860 (cites)- Q21092713 [and each have P356(DOI) claims]
- DOI <has author> ORCID
 - Q21012586 -P50 (author)- Q20980928 (Finn Årup Nielsen, has ORCID claim)
- ORCID <employed by> (organization)
 - Q20980928 -P108 (employer)- Q1269766 (T.U. Denmark)
- DOI <topic> (vocabulary ID)
 - Q27468554 -P921 (main subject)- Q202864 (Zika virus, MeSH ID D000071244)

Wikidata SPARQL queries

- Easy interface at <https://query.wikidata.org/>
- Example: find articles with DOI's where a given ORCID was an author:

```
SELECT ?personLabel ?article ?doi ?articleLabel WHERE {  
  ?person wdt:P496 '0000-0001-6128-3356' .  
  ?article wdt:P50 ?person ; wdt:P356 ?doi .  
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". } }
```

personLabel	article	doi	articleLabel
Finn Årup Nielsen	 wd:Q20973139	10.1093/cercor/bhh119	Right Temporoparietal Cortex Activation during Visuo-proprioceptive Conflict
Finn Årup Nielsen	 wd:Q20978712	10.1002/hbm.10012	Modeling of activation data in the BrainMap database: Detection of outliers
Finn Årup Nielsen	 wd:Q20980925	10.1126/science.291.5506.987	The Real Power of Artificial Markets
Finn Årup		10.1016/i.biopsych.2007.07.009	Frontolimbic Serotonin 2A Receptor Binding in Healthy Subjects Is Associated with

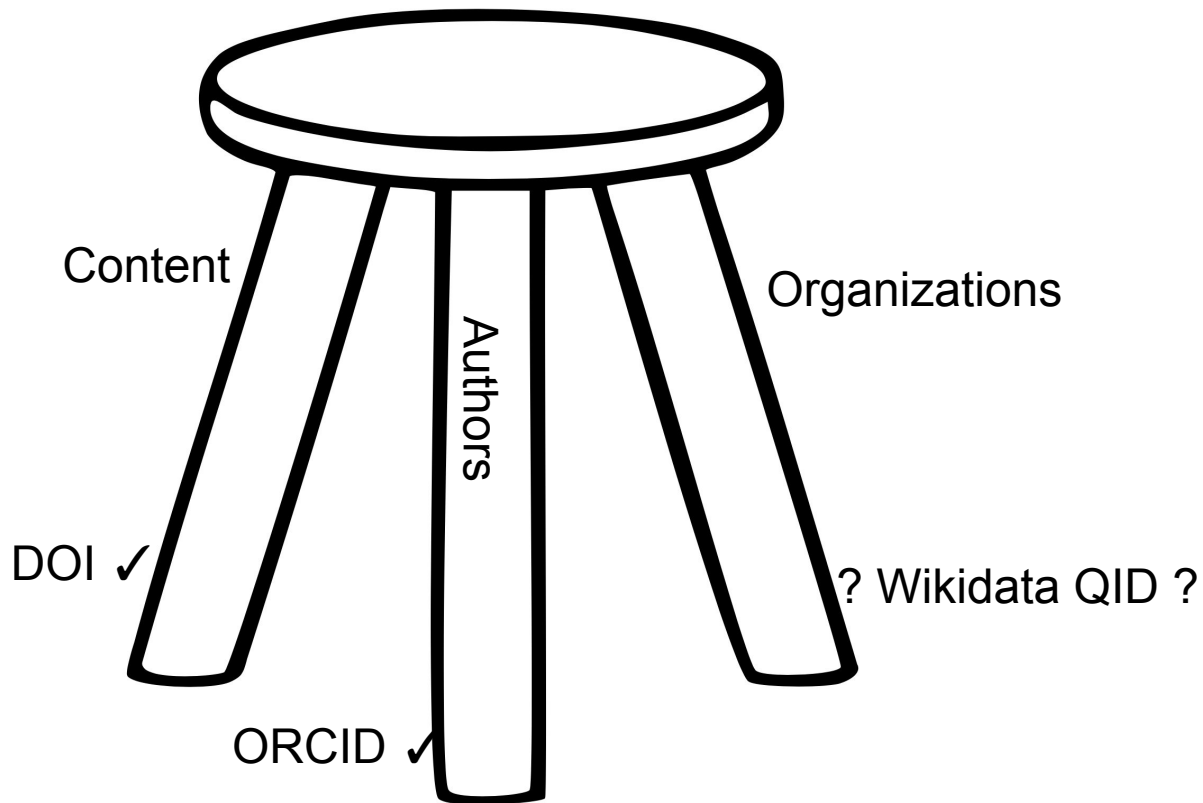
Wikidata items can be grouped by class

Property 31 = “instance of”: membership in a class

- Human (Q5) - 3.3 million
- Location (Q17334923 & subclasses) - 4.6 million
- Organization (Q43229 & subclasses) - 860,000
- Academic journal article (Q18918145 & subclasses) - 301,000
 - (all creative works Q17537576 - 1.5 million)
- Gene (subclass of Q7187) - 610,000

Note varying levels of completeness: only limited selection of people (note also privacy concerns for living people) and creative works; quite comprehensive on concepts, organizations, locations.

Scholarly PID ecosystem



(stolen from Geoff Bilder Crossref LIVE16 talk)

From “Organization Identifier Project: A Way Forward”,

Draft Framing Principles (paraphrased):

- Open, permanent, and unambiguous identification
- Not limited by type or geographic or national boundaries.
- Transparent and non-discriminatory terms of use.
- Identifier and metadata available under CC-0
- Sustainable business model ... long-term openness, persistence, and reliability
- Open Source Software
- Transparent and representative governance
- Living will: continue to operate under these principles if acquired.

Most of these are satisfied by the WMF and Wikidata.

My proposal for the Org Id Project:

- PID = Wikidata QID's
- Main software interface already in place: Wikidata itself
- Establish a representative group/project to:
 - support and better standardize organization metadata in wikidata
 - systematically add/update organization information from reliable sources
 - encourage institutions to curate their own wikidata entries
 - regularly extract curated subsets of wikidata's organization data
 - provide open-source tools for working with organization data in wikidata
- Every organization involved in scholarly work should be sufficiently notable...

Wikidata is a Community!

Not a place to dump data...

Community processes within wikidata

- Learn the norms of the community, ask questions, get experience
- Follow written policies: “Assume good faith.” “Use common sense.” etc.
- Any major efforts should contact relevant “WikiProject” members for advice (eg. WikiProject Physics, WikiProject Source Metadata); also use Project Chat, or other appropriate venues such as “talk” pages.
- Significant automated (“bot”) activities should be presented for review - do 100 sample edits and ensure they meet community standards
- “Vandalism” is watched for carefully: vandals are blocked; edits reverted.
- No “spamming”!
- Wikidata is still young, norms evolving: your voice can make a difference!

Wikidata Notability

See <https://www.wikidata.org/wiki/Wikidata:Notability>

1. At least one valid sitelink to another Wikimedia Foundation site (eg. wikipedia)
2. OR - “an instance of a **clearly identifiable conceptual or material entity**. [...] **can be described using serious and publicly available references.** “
3. OR - “fulfills **some structural need**” - i.e. makes wikidata more useful

reliable source => likely acceptable by criterion 2.

What about errors & conflicts?

- Wrong claims without a source can be replaced with corrections
- If two sources are in conflict, both claims can be presented, referencing the two distinct sources:
 - “The world is complicated. There is no single truth. Record what various sources say instead.”
- Outdated information can be “deprecated”, current info “preferred”
- Duplicate items can be merged: obsolete QID redirects to new one.
- Community consensus resolves disputes. Administrators adjudicate edit wars.
- Data imports should include a method for sustainability: how will data quality be maintained in future?
- Vandals seem to be caught quickly.
- Wikidata information may not be perfect truth, but where it's been curated it's pretty close.

Sample organizations using Wikidata

- Finnish Broadcaster Yle - tagged content with Wikidata items since April 2016
- Flemish art museums imported metadata on artworks, make dataset available as Linked Open Data
- Gene Wiki - imported genes and protein data to wikidata, use in infoboxes
- Google - Knowledge Graph: moved from Freebase to Wikidata
- Schema.org - linked wikidata id's with schema.org "types"
- TED organization - imported metadata for thousands of TED talks
- UNESCO - world heritage sites
- VIAF (OCLC and British Library) - close to 1 million VIAF id's and metadata imported, curation & reuse
- Wikimedia Foundation - linking language wikipedias and other WMF initiatives via wikidata entities

Wikidata main site: <https://www.wikidata.org/>

Queries: <https://query.wikidata.org/>

Gamification: <https://tools.wmflabs.org/wikidata-game/>

Join us!

Questions?

[Email apsmith@aps.org or twitter [@arthursmith](https://twitter.com/arthursmith)]

(Thanks to Léa Lacroix and Lydia Pintscher of Wikimedia DE and other Wikidata speakers for inspiration on slides)