Wikidata and Persistent Identifiers

Arthur Smith - American Physical Society

PIDapalooza 2016

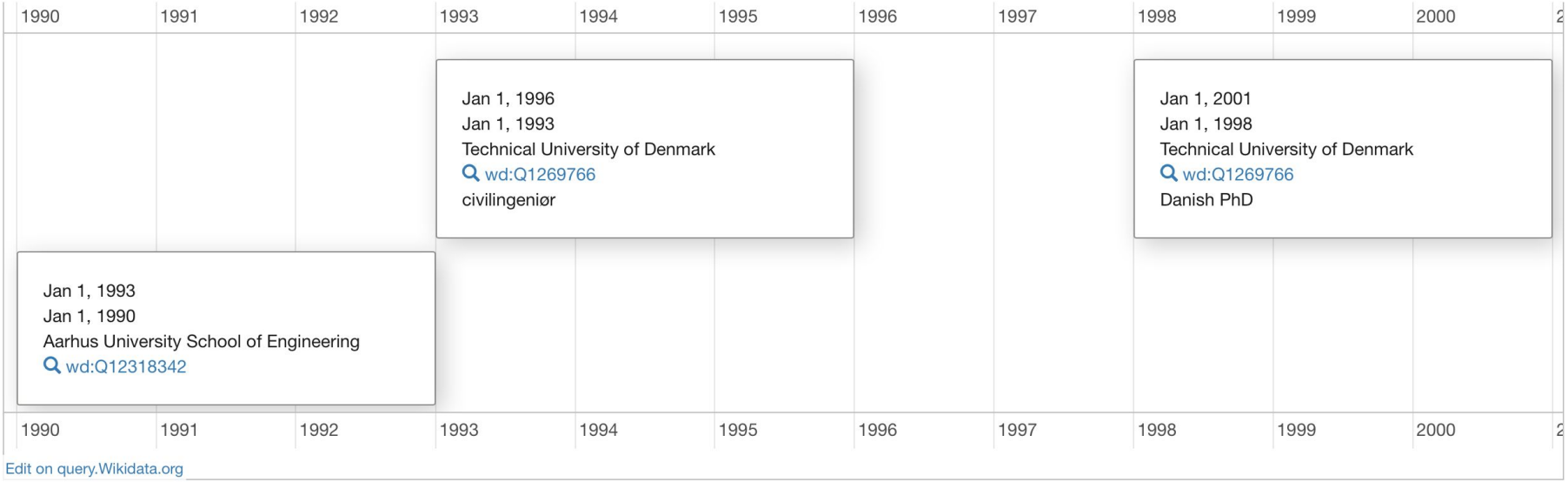
Peter Murray-Rust @petermurrayrust · Oct 9

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

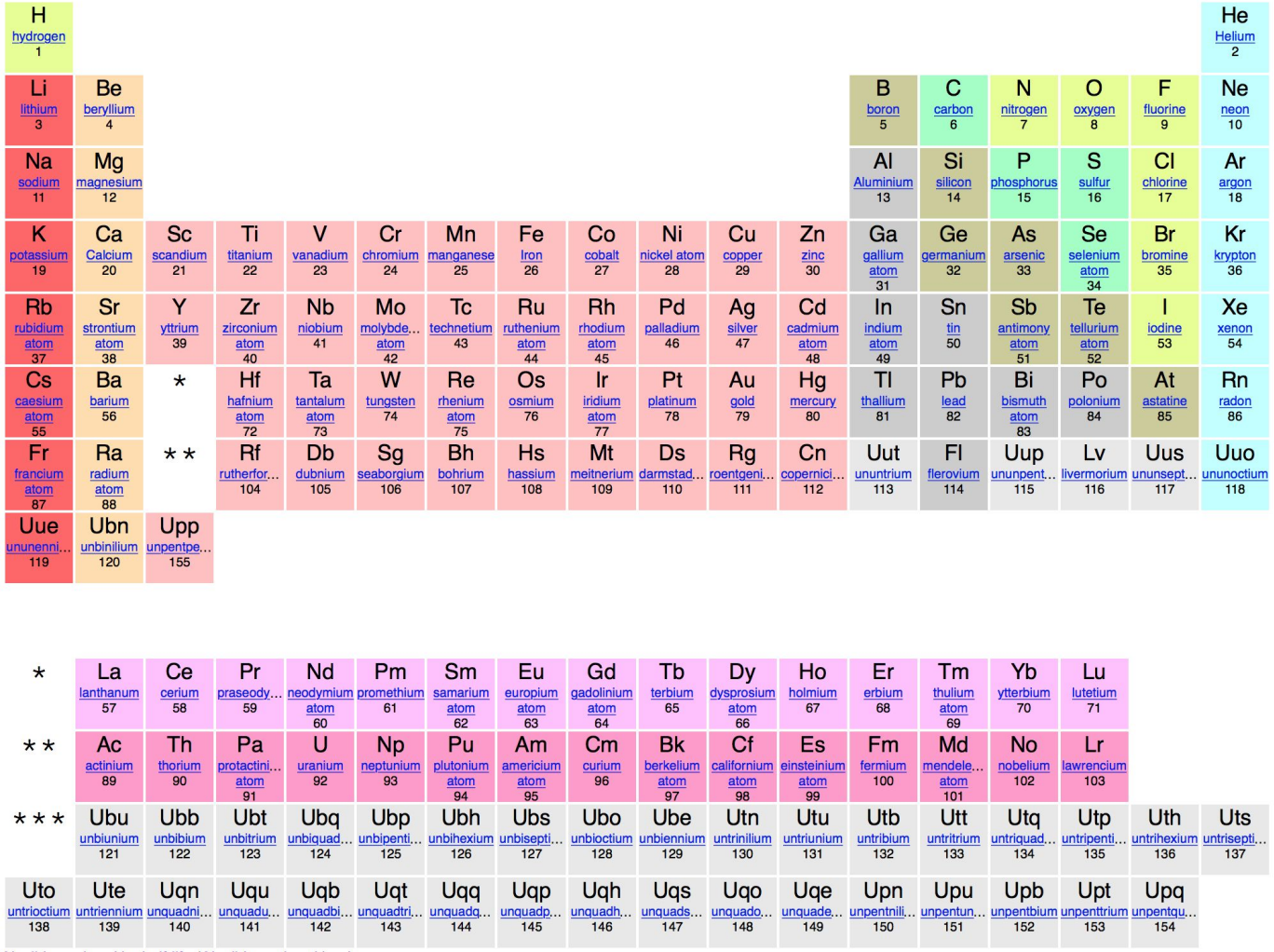


Edit on query.Wikidata.org

LIST OF PUBLICATIONS

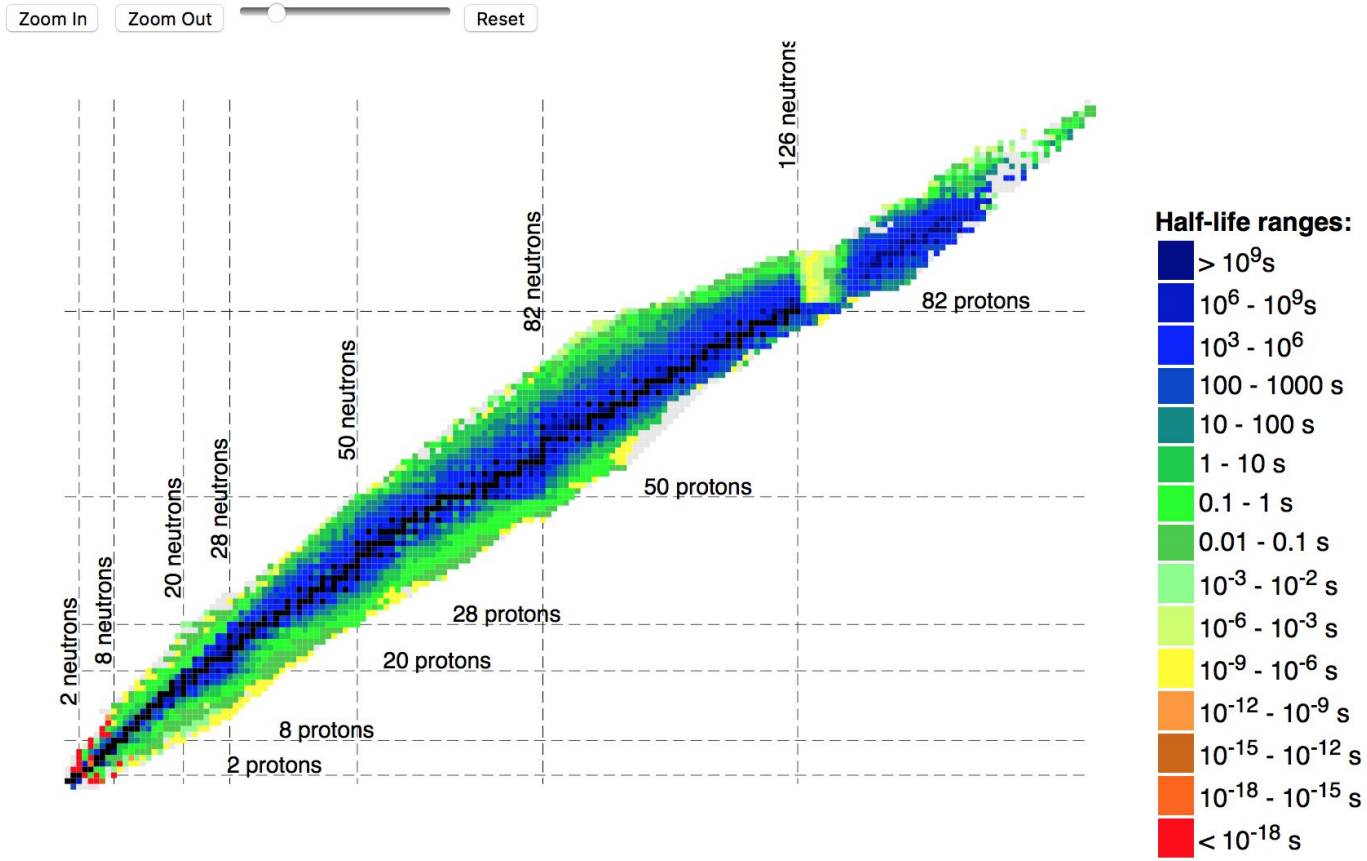
wd:Q26698262	data, models and neuroinformatics	2014			Brian MacWhinney, Ina Bornkessel-Schlesewsky, Michael A. Arbib
wd:Q26698303	Brede tools and federating online neuroinformatics databases	Jan 1, 2014	11	Neuroinformatics	Finn Årup Nielsen
wd:Q22329167	Wikipedia in the eyes of its beholders: A systematic 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



Wikidata chart of the nuclides

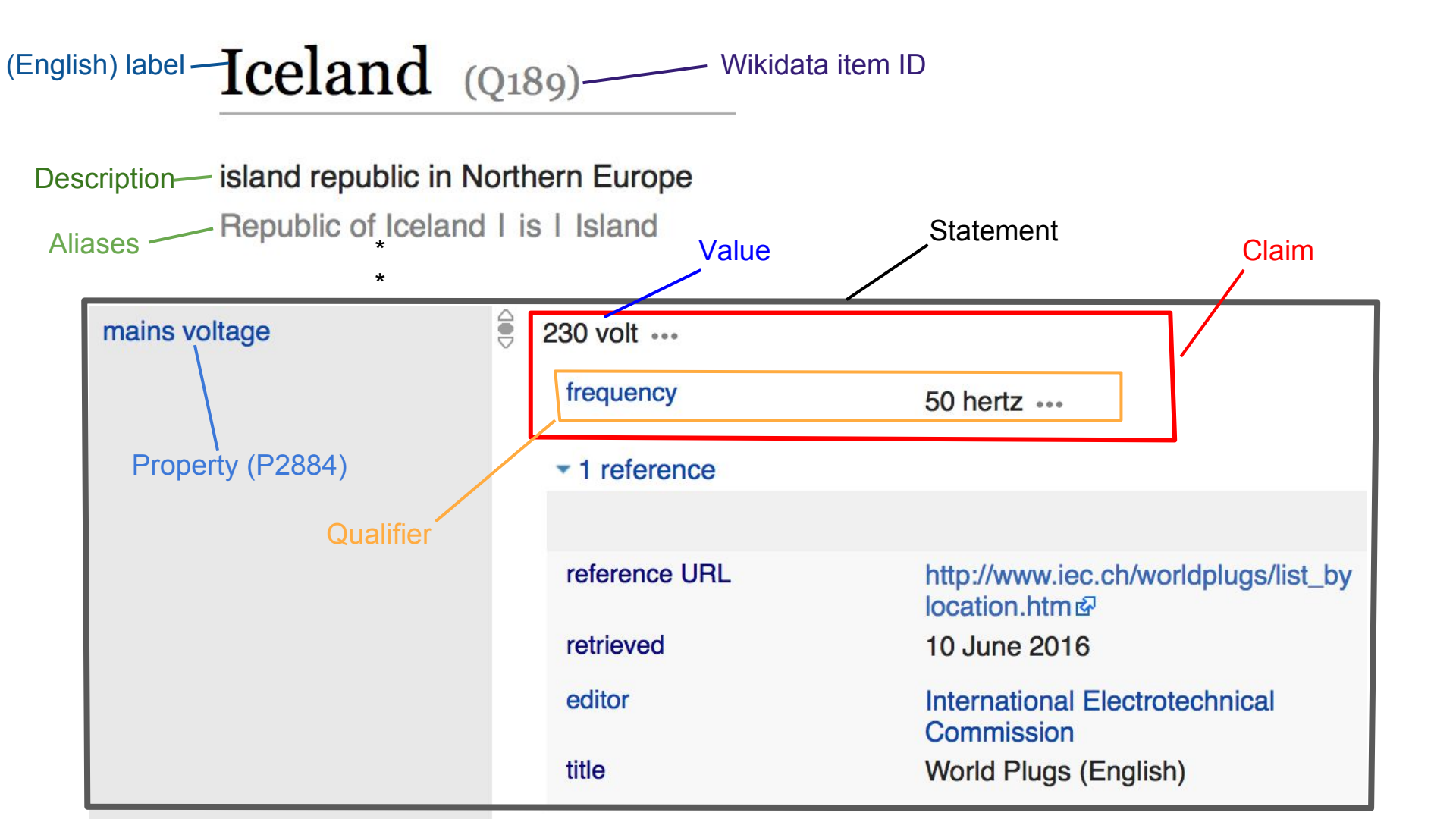
By half life (stable nuclides are black)



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 | is | Island
*
*
Statement
Value
Claim
mains voltage
230 volt
frequency
50 hertz
Property (P2884)
Qualifier
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

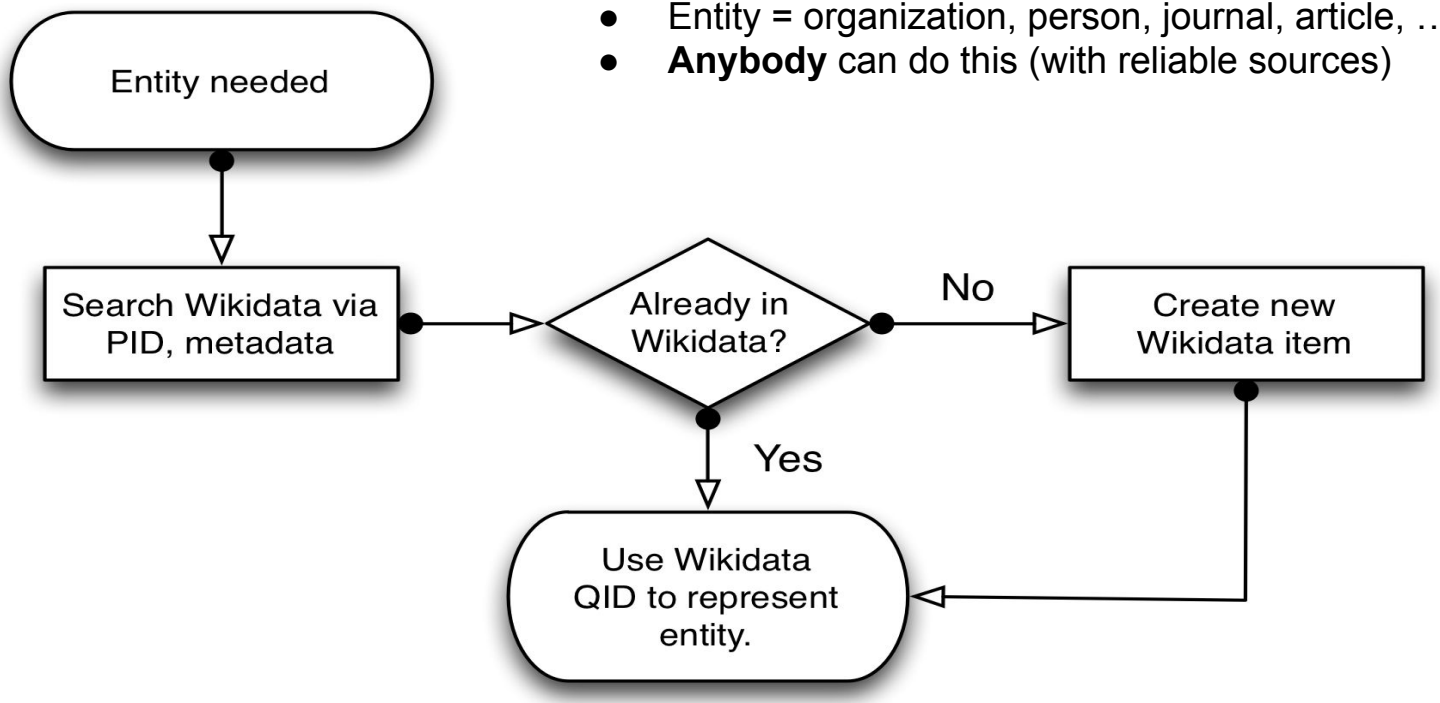
- Over 1300 properties defined with “external ID” datatype
- Includes most ID’s of interest to scholarly publishing:
  - ISBN-13 (P212)
  - ISNI (P213)
  - VIAF ID (P214)
  - ISSN (P236)
  - OCLC control number (P243)
  - LCAuth ID (P244)
  - DOI (P356)
  - MeSH ID (P486)
  - ORCID (P496)
  - arXiv ID (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/j.biopsych.2007.07.009	Frontolimbic Serotonin 2A Receptor Binding in Healthy Subjects Is Associated v

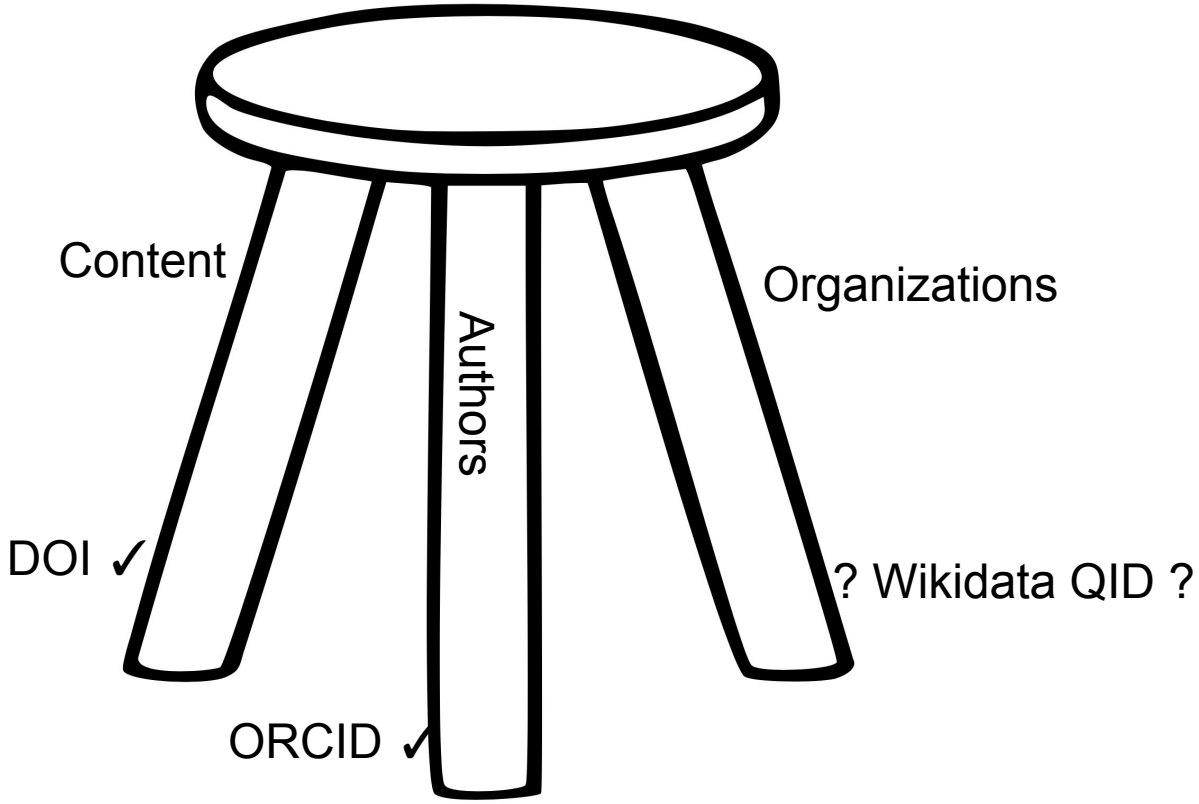
Wikidata items can be grouped by class

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]

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