Wikibase and Wikidata

Wiki* Family

Wikimedia Foundation: American 501(c)(3) nonprofit founded in 2001

Wikipedia: multilingual web-based free-content encyclopedia

MediaWiki: free and open-source wiki software

Wikidata.org: a collaboratively edited structured dataset used by Wikimedia sister projects & others

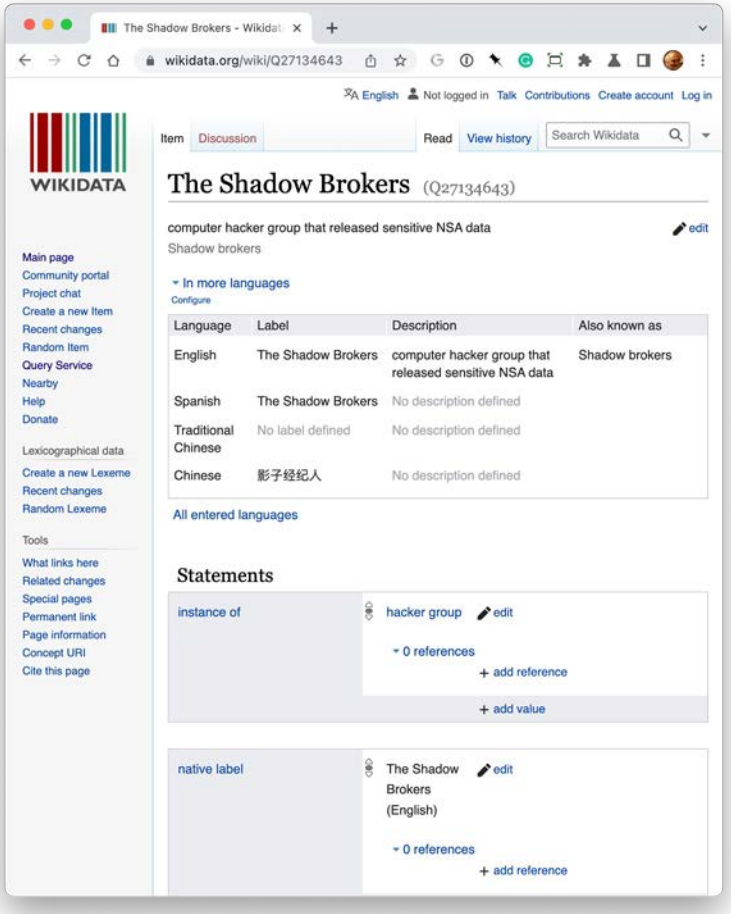
Wikibase: a MediaWiki extension to store and manage structured data

Wikimedia Commons: free media repository

And more…

Wikidata Knowledge Graph

- **Large knowledge graph** with ~1B facts about ~100M items
- **Fine-grained ontology**: ~2M types; ~10K properties
- **Multilingual**, names and strings tagged with language id
- Links entity's **Wikimedia pages** in many language and many other resources
- Entities have a canonical **name, aliases** and **short description** in multiple languages plus many key facts
- Queried using SPARQL query language
- See this account of its history and how it launched in 2012

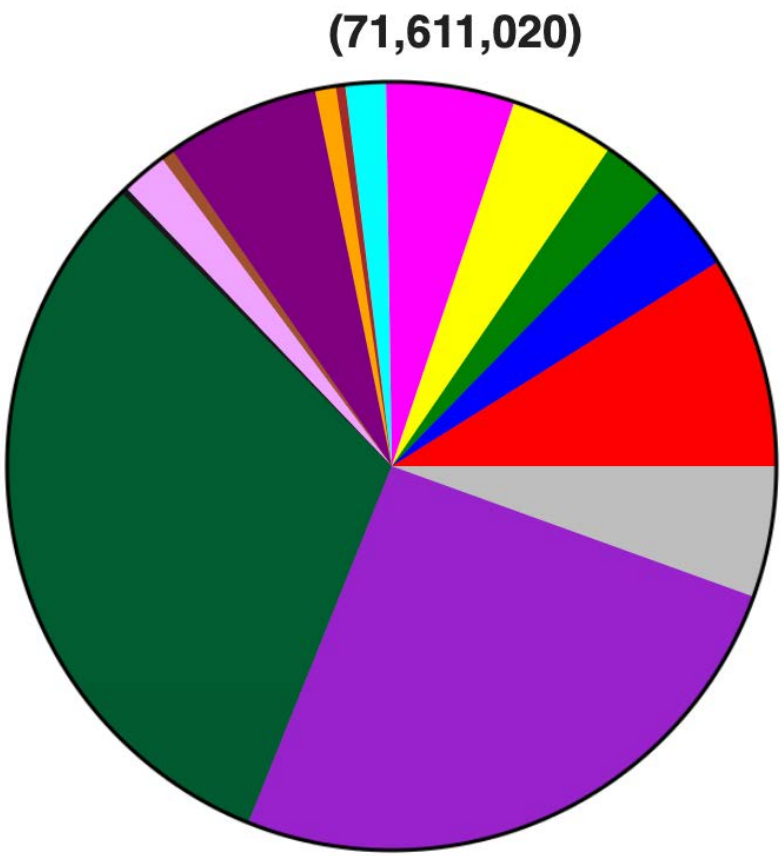


Q27134643

Wikibase is…

- Wikibase: structured data repository based on MediaWiki
- Complex/expressive data model has triples, provenance, qualifiers, and alternate values
- Export to standard formats including JSON, RDF/XML, N3, and Turtle
- Its RDF model is queryable via SPARQL
- Local installs via a Docker image
- Stored in a custom database rather than a standad triple store

Wikidata Statistics



Data from Feb. 2020, Wikidata is 50% larger now. But the proportions may be similar

- human: 6,376,879 (8.9%)
- taxon: 2,726,046 (3.8%)
- administrative territorial entity: 1,943,285 (2.7%)
- architectural structure: 3,159,472 (4.4%)
- occurrence: 3,898,674 (5.4%)
- chemical compound: 1,188,724 (1.7%)
- film: 294,370 (0.4%)
- thoroughfare: 630,794 (0.9%)
- astronomical object: 4,601,733 (6.4%)
- Wikimedia list article: 404,454 (0.6%)
- Wikimedia disambiguation page: 1,358,230 (1.9%)
- Wikinews article: 195,900 (0.3%)
- scholarly article: 22,574,314 (31.5%)
- other P31/P279: 18,284,676 (25.5%)
- no P31/P279: 3,973,469 (5.5%)

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

What may seem weird…

- Wikidata Items and properties do not have noraml, natural language based names
  - Motivation is to be truly multilingual
- **Items** begin with a **Q** and **properties** with a **P**; both continue with a sequence of digits
- Anyone can create a new Q item, but P properties are a controlled vocabulary
  - Adding a property starts with making a proposal
  - See this page of all properties organized by topic

Minding your Ps and Qs

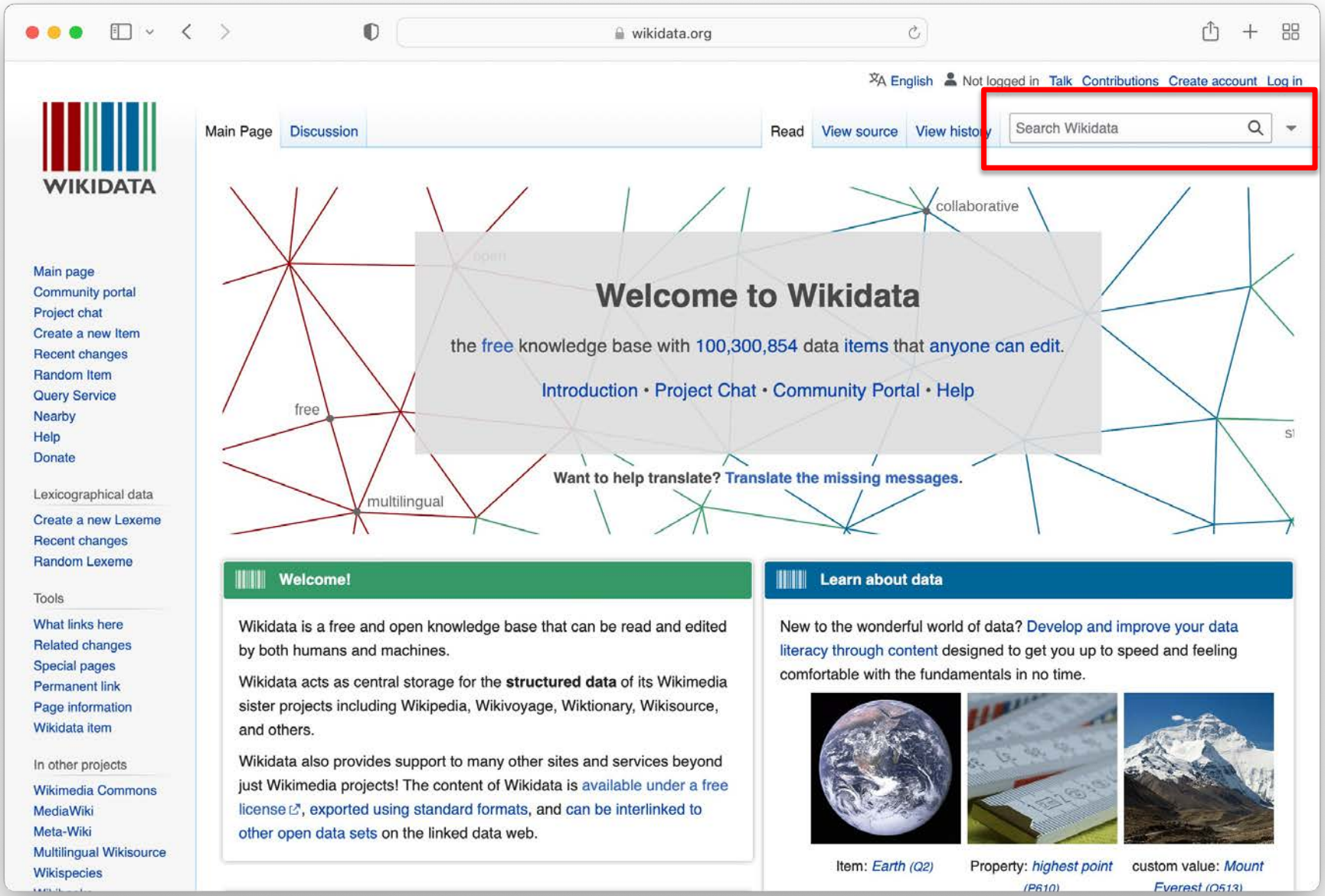
- Q1: Universe
- Q2: Earth
- Q3: Life
- Q4: Death
- Q5: human
- Q42: Douglas Adams
- Q735049: UMBC
- Q113553960: Hacking tool

- P31: instance of
- P279: subclass of
- P108: employer
- P306: operating system
- P683: Chem. formula ID

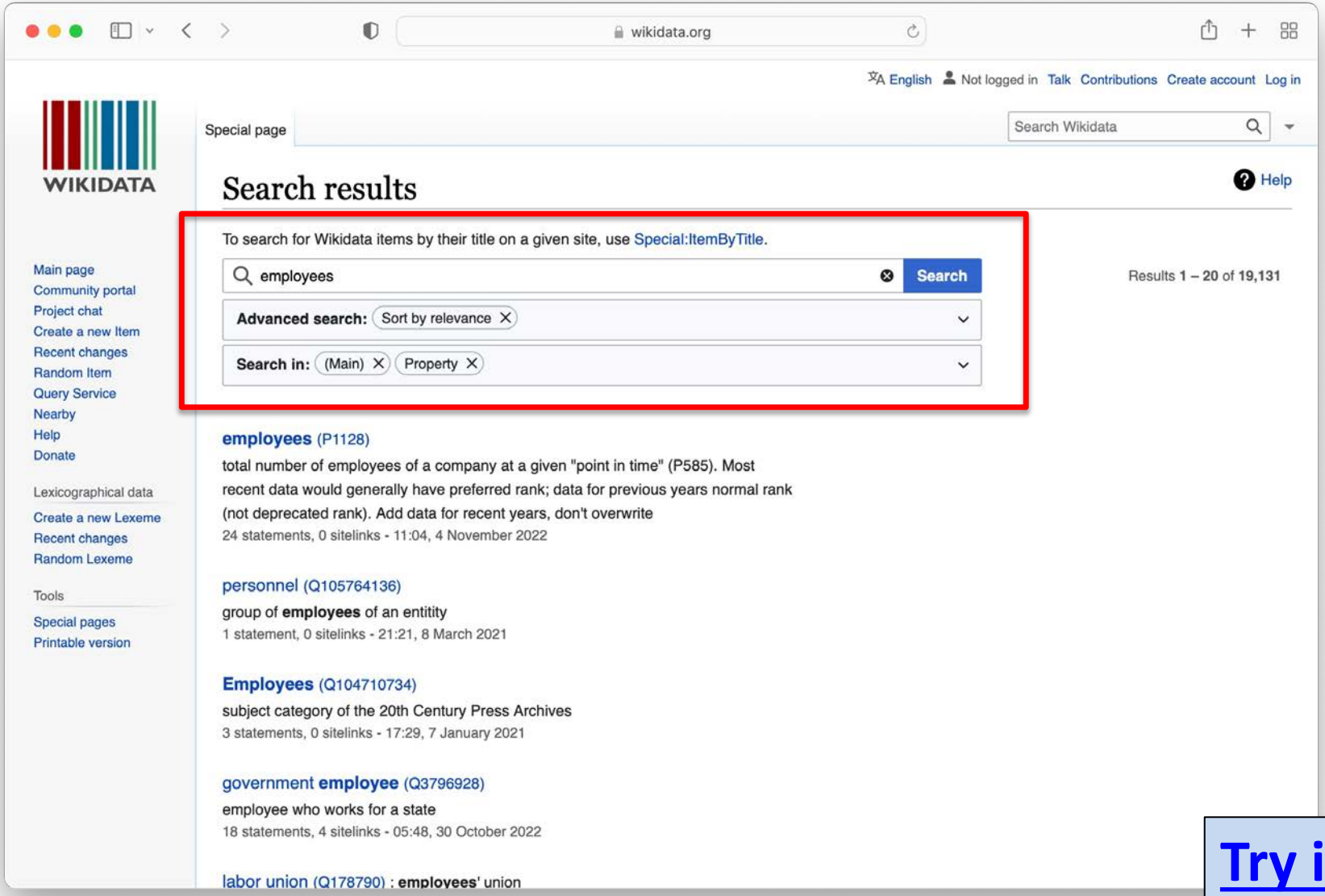
To make it easy for us,

- The human interface for viewing, exploring, and editing displays the RDFS label of an entity or property
- The SPARQL query interface shows

Finding things on Wikidata

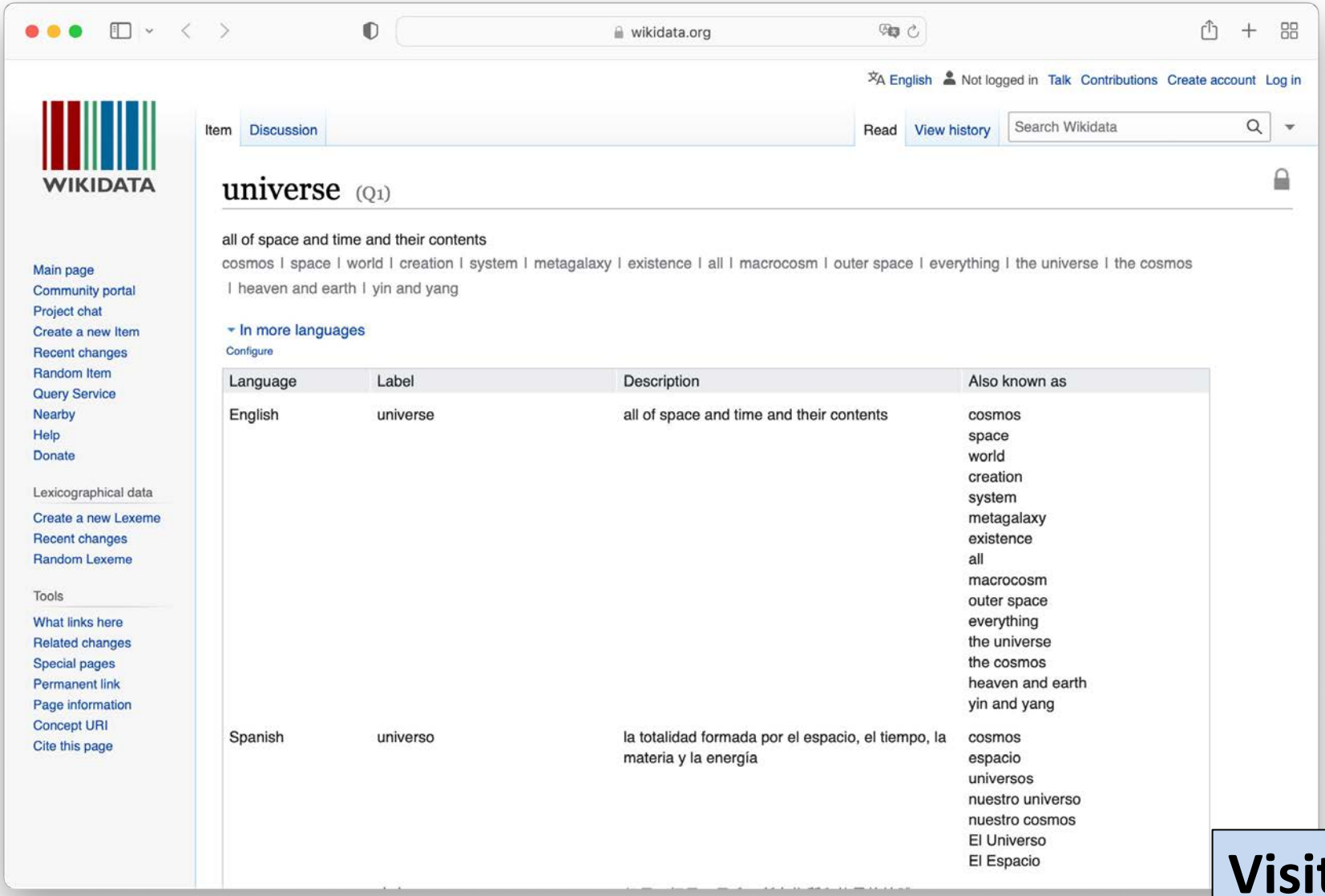


Wikidata's search has options



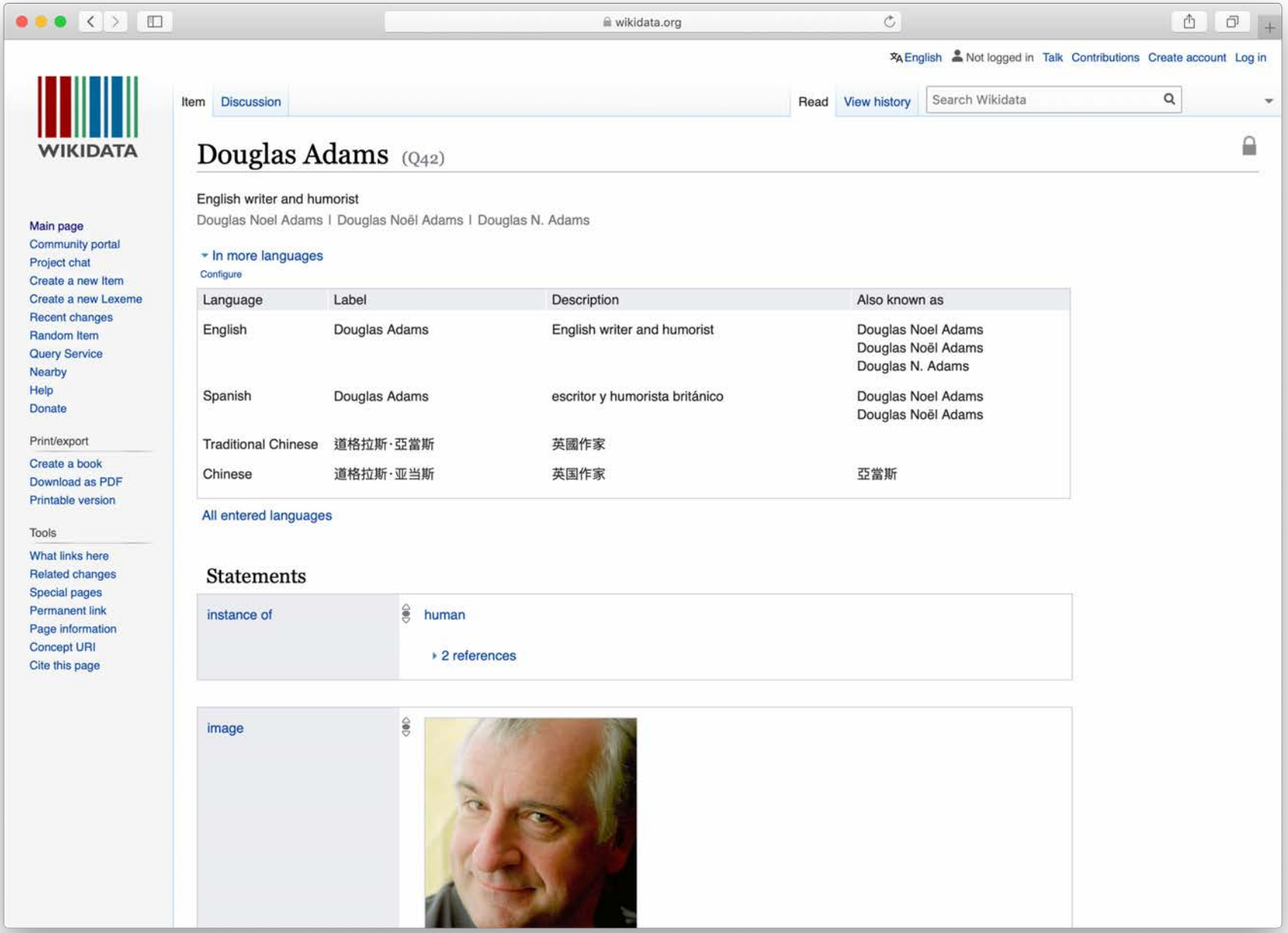
Try it

https://wikidata.org/wiki/Q1



Visit

https://www.wikidata.org/wiki/Q42



https://wikidata.org/wiki/Special:EntityData/Q42.json

The entity in JSON

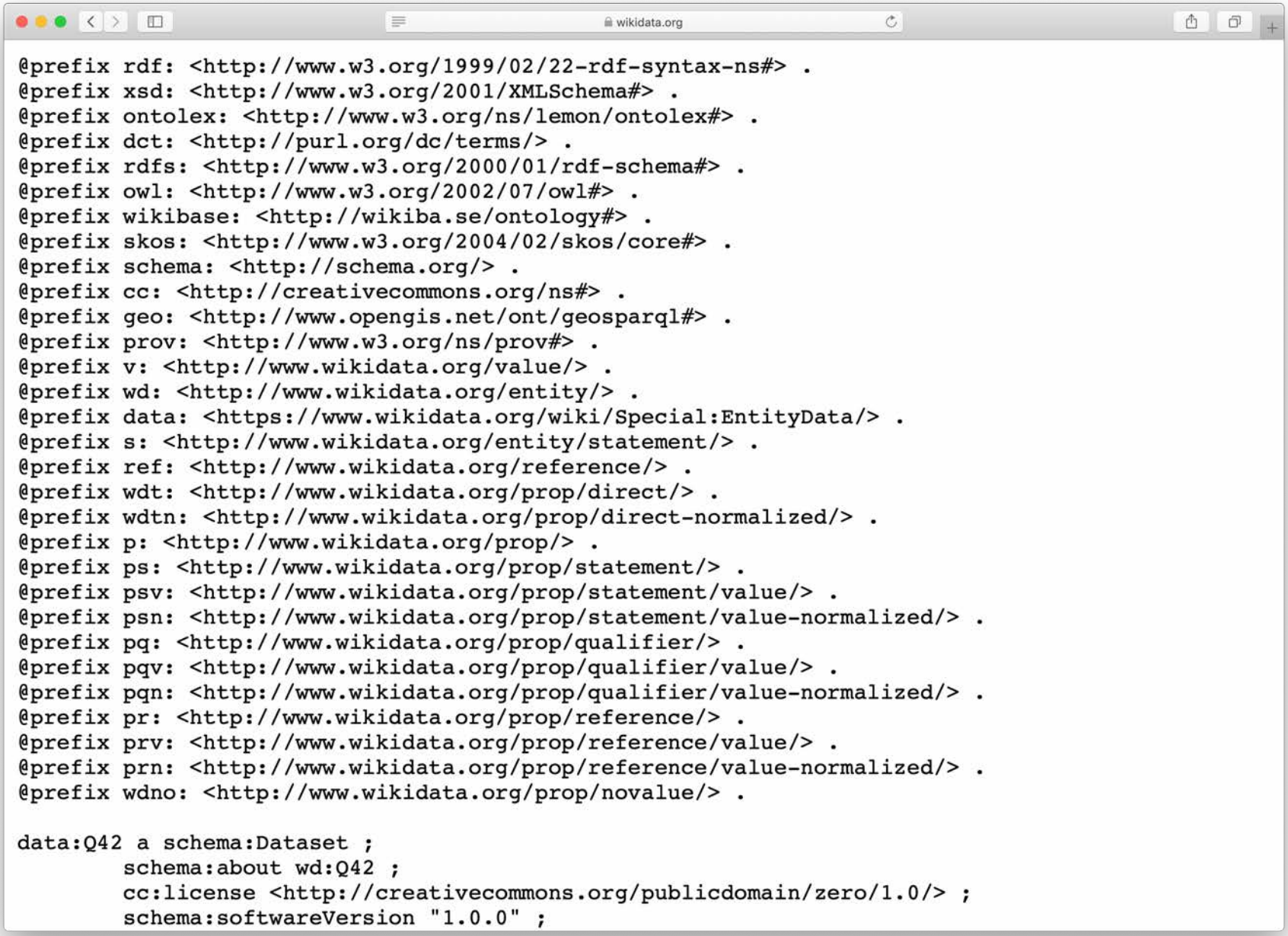
Q42.json — ~/Downloads

Q42.json


```
{
  "entities": {
    "Q42": {
      "pageid": 138,
      "ns": 0,
      "title": "Q42",
      "lastrevid": 1065557227,
      "modified": "2019-11-29T18:46:12Z",
      "type": "item",
      "id": "Q42",
      "labels": {…},
      "descriptions": {…},
      "aliases": {…},
      "claims": {…},
      "sitelinks": {…}
    }
  }
}
```


~/Downloads/Q42.json* 11:7 — 1 LF UTF-8 JSON GitHub Git (0) 1 update

https://wikidata.org/wiki/Special:EntityData/Q42.ttl



Wikidata Query Service (WDQS)

- Let's query Wikidata's WDQS to see when Douglas Adams was born
- We'll need to know that Douglas Adams is **wd:Q42** and date of birth is **wdt:P569**
- If we click on  on the left side of the WDQS we can see labels or enter names

Try it

Wikibase Data Model

- **Item** = subjects = entities
- **Property** = properties
- **Value** = entities or datatypes (string, number, …)
- **Snak** = basic assertion about item, i.e., Property-Value pair, ”small, but more than a byte”
  - Some are simple claims: *population Berlin 3,499,879*
  - Others (e.g., types) are structural: *type Berlin City*
  - Others include claim **qualifiers**
    Population Berlin 3,499,879; considering only territory of city, as estimated on 30 November 2011

Items have

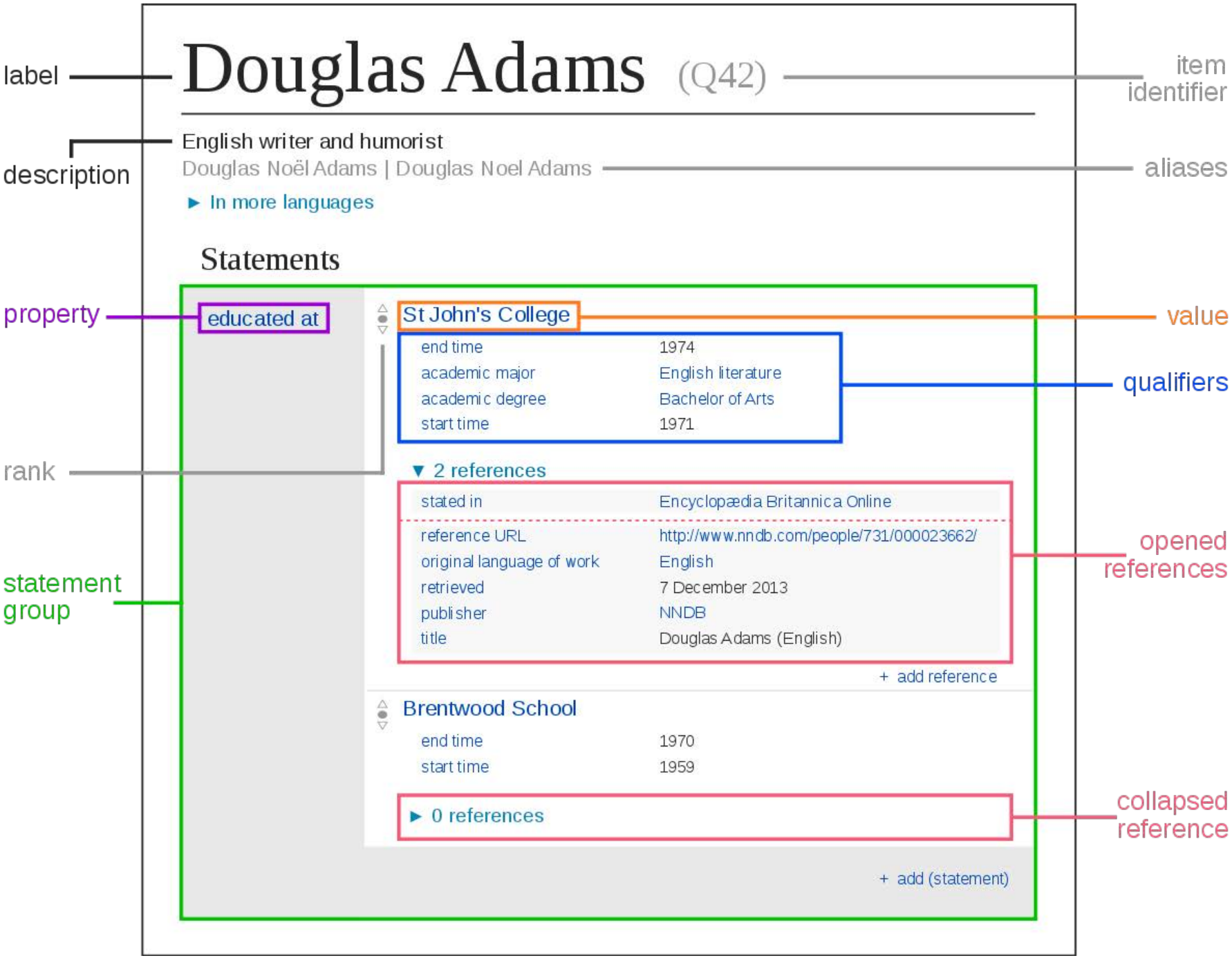
- **Item identifier** (number prefixed with *Q*)
- **Fingerprint**, consisting of:
  - Multilingual **label***
  - Multilingual **description***
  - Multilingual **aliases**
- **Statements**, each consisting of:
  - **Claim**, consisting of:
    - Property
    - Value
    - Qualifiers (additional property-value pairs)
  - **References** (each with one or more property-value pairs)
  - **Rank**
- **Site links**

Statements…

- A statement may have:
  - one property (in the example, P551 “residence")
  - one value (Q84 “London”)
  - optionally one or more qualifiers (e.g, property:P582, “end time” 11 May 2011)
  - optional reference(s) (e.g., property:P143 “imported from Wikipedia”)
- The property, value, and qualifiers together are also called the **claim**, which together with any source references forms a statement

Properties have …

- **Property identifier** (number prefixed with *P*)
- **Fingerprint**, consisting of:
  - Multilingual **label***
  - Multilingual **description***
  - Multilingual **aliases**
- **Statements**, each consisting of:
  - **Claim**, consisting of:
    - Property
    - Value
    - Qualifiers (additional property-value pairs)
  - **References** (each with one or more property-value pairs)
  - **Rank**
- **Datatype**


label
Douglas Adams
(Q42)
item identifier
description
English writer and humorist
Douglas Noël Adams | Douglas Noel Adams
aliases
► In more languages
Statements
property
educated at
St John's College
value
end time 1974
academic major English literature
academic degree Bachelor of Arts
start time 1971
qualifiers
rank
▼ 2 references
stated in Encyclopædia Britannica Online
reference URL http://www.nndb.com/people/731/000023662/
original language of work English
retrieved 7 December 2013
publisher NNDB
title Douglas Adams (English)
opened references
statement group
+ add reference
Brentwood School
end time 1970
start time 1959
► 0 references
collapsed reference
+ add (statement)

Statements…

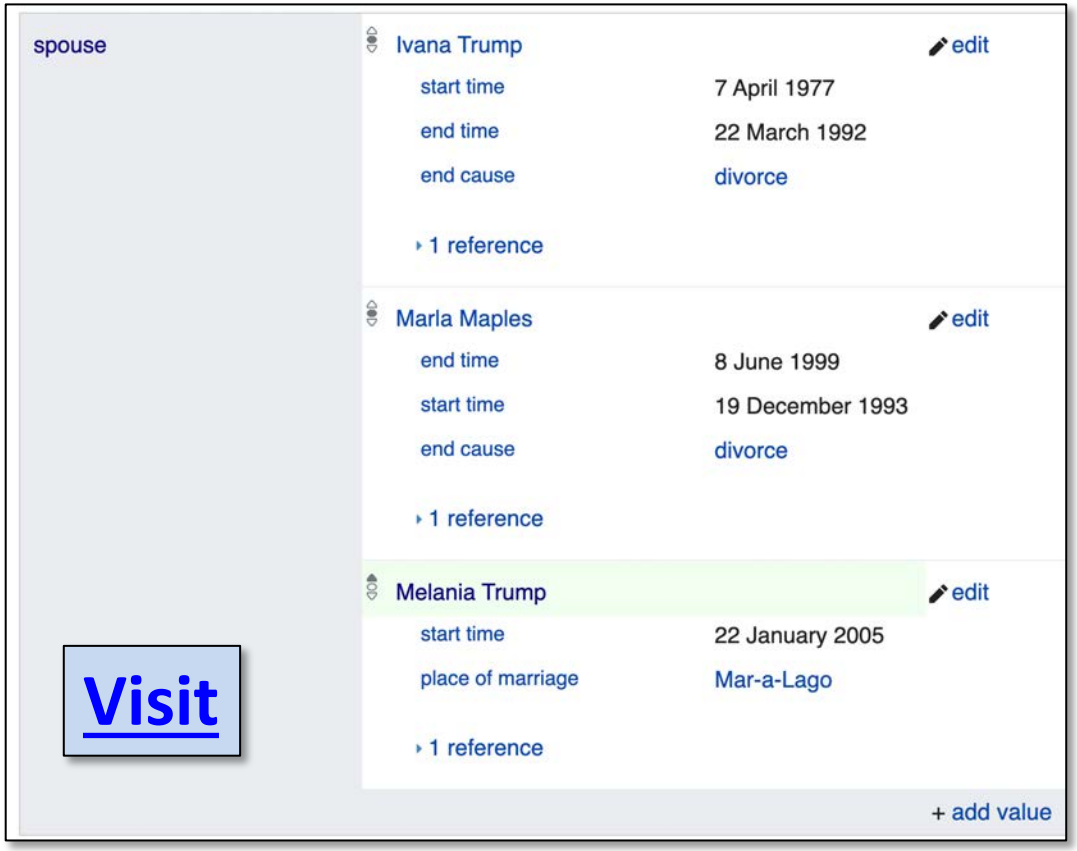
- Requirement: "Wikibase will not be about the truth, but about statements and their references"
- Doesn’t model items, but statements about them
- Not “Douglas Adams residence is London”
- But “There’s a statement of Douglas Adams having a residence of London prior to 11 May 2011 according to WIkipedia”

Example: Trumps spouses

- Who are Donald Trump's spouses?
- We must identify the IDs for
  - Donald Trump: **Q22686**
  - Spouse property: **P26**
- Write and run a simple SPARQL query
- Enter it in Wikidata's query service
  1. First try
  2. Second try: ask for a label
  3. Third try: restrict to just English label
  4. Final try: using the label service

But what about the other spouses?

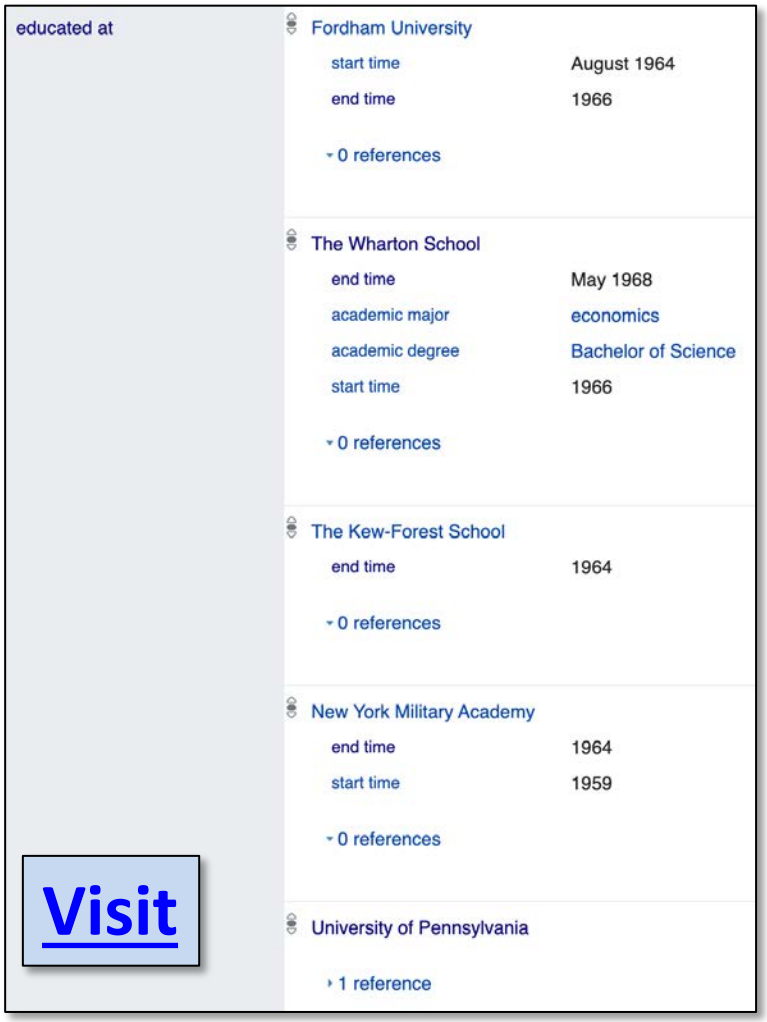
- Query only returns one answer: his *current* spouse
- Other values do have an end time
- Maybe that's a feature!



- Let's try another query: what schools did Donald Trump attend?

Property Rank

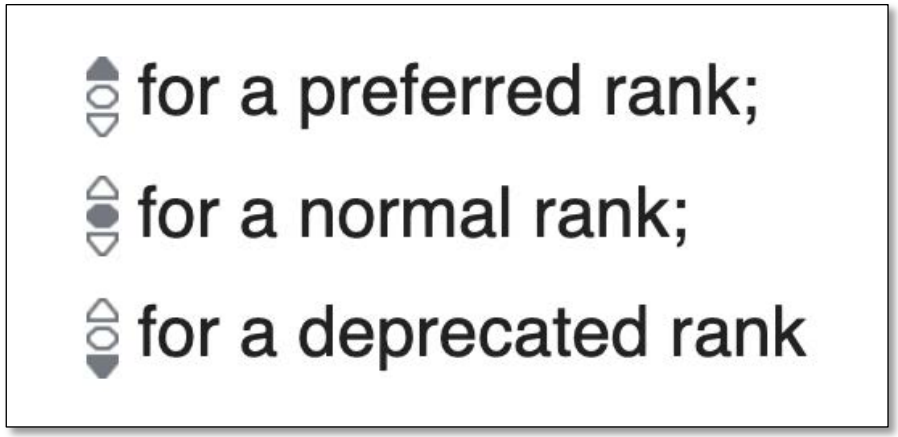
- We get four schools, even tho all have end dates (we might quibble that Penn & Wharton are the same)
- Does Wikidata's ontology know that *spouse* (P26) is a temporal quality & *educated at* (P69) isn't?
- No, though property has some constraints that might be useful
- The mechanism used is to give each value a rank, which provides a way to annotate multiple values for a statement.



Marla Maples

Melania Trump

Ranking claims



- **Preferred:** most current or values representing a consensus
- **Normal:** default; no judgement of a value's accuracy and currency
- **Deprecated:** errors or outdated

For DT's spouses, Melania has preferred rank and the others normal rank

All of DT's schools had normal rank

How are ranks represented in RDF and how does the Wikidata query service use them?

WDQS Procedure

What's matched for **?s wdt:Pxxx ?o**

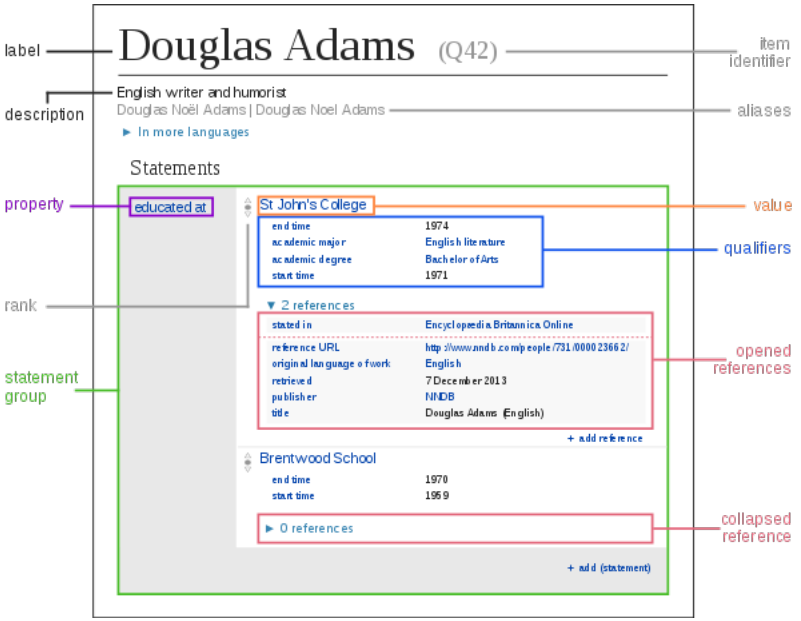
- If there's at least one ?o with preferred rank, only values preferred values are returned
- If there are no preferred values, all values with normal rank are returned
- Deprecated values are never returned.

The humans or bots populating the graph must figure out how to assign ranks

Qualifiers, rank and references

Wikidata uses special namespaces to access a reified node with claim's qualifiers, rank & references

- **prefix p:** points not to object, but to statement node
- It is then subject of other triples
- Within a statement node:
  - **ps:** gets the object
  - **pq:** gets qualifier information
  - **wikibase:rank** gets rank information
  - **prov:wasDerivedFrom/pr:P248** gets reference values



Example (1): Douglas Adams schools

```
# find schools attended with their start and end times
SELECT ?education ?educationLabel ?starttime ?endtime
WHERE {
 wd:Q42 p:P69 ?statement.
 ?statement ps:P69 ?education.
 ?statement pq:P580 ?starttime.
 ?statement pq:P582 ?endtime.
 SERVICE wikibase:label {bd:serviceParam wikibase:language "en"}
}
ORDER BY ?starttime
```

Try it

Wikidata WDQS prefixes

```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wds: <http://www.wikidata.org/entity/statement/>
PREFIX wdv: <http://www.wikidata.org/value/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
PREFIX pq: <http://www.wikidata.org/prop/qualifier/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX bd: <http://www.bigdata.com/rdf#>
```

Example (2): Douglas Adams schools

Simplify using the [ ] syntax to eliminate the ?statement variable

```
# find schools attended with their start and end times
SELECT ?education ?educationLabel ?starttime ?endtime
WHERE {
 wd:Q42 p:P69
   [ ps:P69 ?education;
     pq:P580 ?starttime;
     pq:P582 ?endtime. ]
 SERVICE wikibase:label {bd:serviceParam wikibase:language "en"}
}
ORDER BY ?starttime
```

Try it

Example (3)

Here's an example getting rank information

```
# find all schools Douglas Adams attended and their ranks
SELECT ?education ?educationLabel ?rank
WHERE {
  wd:Q42 p:P69
     [ ps:P69 ?education;
       wikibase:rank ?rank  ].
 SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

Try it

Trumps Spouses

```
# Get all of Donald Trump's spouses, current, former, and
# deprecated along with their ranks
SELECT ?spouse ?spouseLabel ?rank
WHERE {
 wd:Q22686 p:P26
      [ ps:P26 ?spouse;
        wikibase:rank ?rank ].
 SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

Try it

Deprecated values

- See this page on deprecation
- Honoré de Balzac (Q9711) has three values for date of death (P570)
  1. 1850
  2. 18 August 1850
  3. 19 August 1850
- The 18 August value is tagged as preferred with the reason most precise value (Q71536040)
- The August 19 claim is tagged as deprecated, with the reason incorrect value (Q41755623)
- The 1850 claim is not tagged so carries the defaut rank of normal

Ontology Exploration

- wdtaxonomy is a useful commandline tool for exploring Wikidata's ontology
- Given a type (e.g, Q3918, university) you can quickly see
  - Subtypes or supertypes (immediate or inferred),
  - Number of instances (immediate or inferred),
  - Direct instances
  - Number Wikimedia sites it is in
  - And more
- Implemented in Javascript with a command line script
- Much faster than using SPARQL

```
$$ wdtaxonomy Q3918 -c -t
university (Q3918) •163 ×15380 ↑↑
├──Universities in Germany (Q212462) •2
├──national university (Q265662) •11 ×73
├──National University (Q366354) •5
├──Imperial universities of Japan (Q562092) •12
├──Byzantine university (Q622870) •4
├──college and university rankings (Q847843) •23 ×45 ↑
├──public university (Q875538) •39 ×974 ↑
├──private university (Q902104) •32 ×846 ↑
├──new university (Q987075) •4 ×1
├──Red brick university (Q1202123) •11
├──??? (Q1305046) •2
├──institute of technology (Q1371037) •20 ×325
├──veterinary medicine school (Q1384955) •5 ×28
├──online university (Q1407393) •4 ×10 ↑
├──virtual university (Q1755248) •8 ×11
│  └──online university (Q1407393) •4 ×10 ↑ …
├──comprehensive university (Q1767829) •2 ×6
├──plate glass university (Q1902446) •8
├──medical university (Q1916585) •1 ×9 ↑
├──??? (Q2073922) •1
├──pontifical university (Q2120466) •18 ×37 ↑↑
├──Corporate university (Q2278672) •6
├──ancient university (Q2667285) •9 ×1
├──central university (Q3351682) •12 ×2
├──collegiate university (Q3354859) •9 ×12
├──deemed university (Q3520135) •6 ×16
├──university in France (Q3551775) •3 ×75 ↑
├──Istituto superiore per le industrie artistiche (Q3803831) •2 ×4
├──??? (Q3803846) •1 ×2
├──Smolny Institute for Noble Maidens (Q4432880) •1
├──??? (Q4475845) •2
├──federal university (Q4481793) •3 ×3
├──ecclesiastical university (Q5332280) •6 ×2
├──labor universities (Q5690751) •1 ×6
├──open university (Q6755402) •4 ×1
├──Urban university (Q7900184) •2
├──??? (Q10387922) •1
├──international university (Q10829188) •3 ×9
├──autonomous university (Q11057861) •2 ×1
├──research university (Q15936437) •9 ×224
├──Italian universities (Q20009854) •2
├──??? (Q20052016) •1 ×2
├──Canciller de Universidad (Q21547263)
├──imperial university of the Russian Empire (Q28667313) •2 ×12
├──universities in China (Q28700403) •1
├──Institute of National Importance (Q47531586) ×1 ↑
├──campusuniversity (Q59537665) ×3
├──Indiana University Bloomington Department of French and Italian (Q63441027)
├──Indiana University Department of French and Italian (Q63441251)
└──Indiana University Bloomington Department of History (Q63441447)
```

Summary

- Wikidata is an important community-based knowledge graph that has many uses
- It has many unique features and includes
  - A good interface for humans to inspect & edit its data
  - An open SPARQL interface with a timeout at ~60 sec
  - APIs for search and creating bots
- It integrates information from Wikipedia sites as well as many other sources