

# Fixing Wikidata

(by viewing it as a series of tables)

Yaron Koren · SMWCon Fall 2023 · December 13, 2023

# About me

- Yaron Koren
- MediaWiki developer, consultant and enthusiast
- My company: WikiWorks
- My book: *Working with MediaWiki*
- My podcast: *Between the Brackets*






# Wikidata

- Created by Denny Vrandečić (with assistance from Markus Krötzsch, and many others)
- A massive repository of 1 billion facts ("statements")
- Highly multilingual, highly referenced
- **The greatest store of general-knowledge data in world history**
- Runs on MediaWiki and a set of extensions collectively called Wikibase

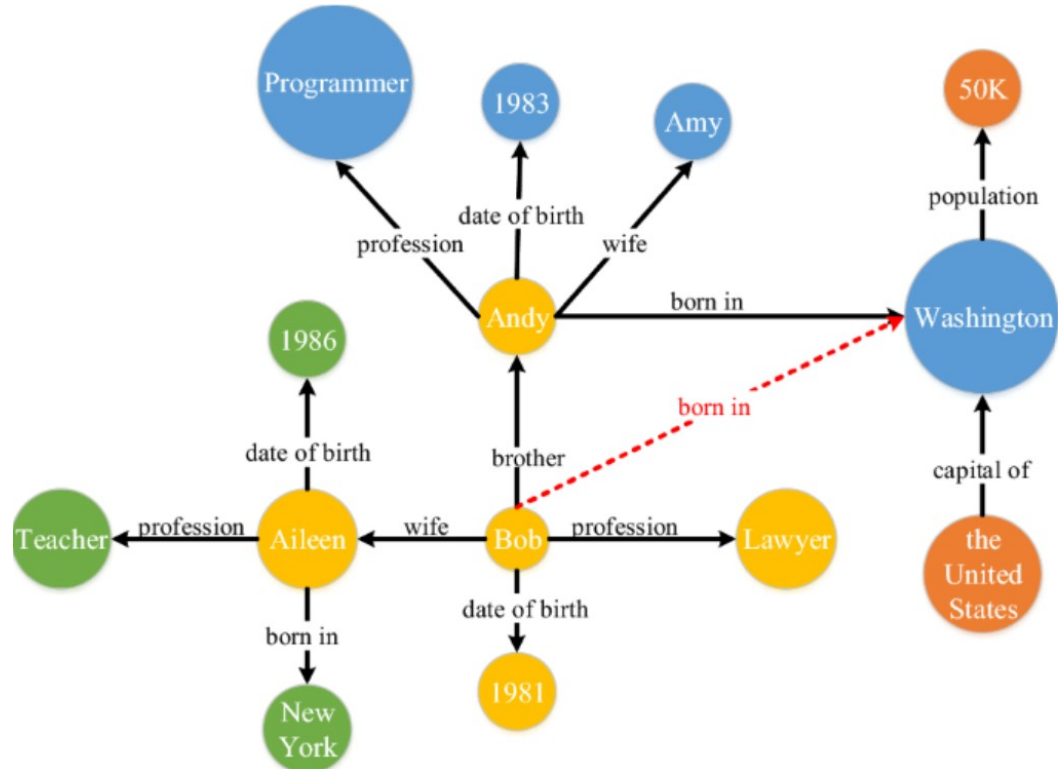
# Goals of Wikidata

- Backend for Wikipedia's infobox data
- Backend for other parts of Wikipedia, like categories and tables?
- Queriable resource for other sites, via fixed queries
- Queriable resource for on-the-fly queries
- "Q IDs" for each item serve as a general ID authority

# Let's go through these goals...

- Backend for Wikipedia's infobox data 
- Backend for other parts of Wikipedia, like categories and tables? 
- Queriable resource for other sites, via fixed queries 
- Queriable resource for on-the-fly queries 
- "Q IDs" for each item serve as a general ID authority 

# Standard view of Semantic Web



# Standard view of Semantic Web



# Tables are better than graphs

- If a set of objects are all of the same type/class, then the set of fields for each one should be (more or less) the same
- No object is "special", and if an object has a field that none else do, that's not really data at all
- Thus: knowledge is inherently structured, not free-form




# Wait...

- Isn't Wikidata structured as a graph?
- Not really, though it can be queried via a graph database.
- Anyway, the issue is the user interface, not the data itself.








# Wikidata does have structure – enforced via warnings


wikidata.org/wiki/Q117833118

Language	Label	Description	Also known as	
Traditional Chinese	No label defined	No description defined		
Chinese	No label defined	No description defined		

All entered languages

### Statements

instance of	literary work			1 reference	
title	Robert Anak Sur			0 references	
genre	Indonesia Literature in the period 1950-1965				1 reference

**Potential issues** 

**value-type constraint** [Help](#) [Discuss](#)

Values of **genre** statements should be instances or subclasses of one of the following classes (or of one of their subclasses), but **Indonesia Literature in the period 1950-1965** currently isn't:

- [genre](#)
- [musical style](#)
- [artistic theme](#)
- [art style](#)
- [software category](#)
- [criticism](#)
- [discussion](#)
- [subject of depiction](#)
- [musical form](#)
- [Geistesgeschichte](#)
- ...

*This result is cached and might be out of date by up to 3 minutes.*

+ add value

# Same idea, but enforcement of properties

The screenshot displays the Wikidata interface for the property **song**. A modal dialog box titled "Potential issues" is open, highlighting a "subject type constraint" issue. The dialog text reads: "Entities using the **record label** property should be instances of one of the following classes (or of one of their subclasses), but **Little Honda** currently isn't:". A bulleted list of allowed classes is shown: musical ensemble, human, series of creative works, audio release, music video, audiobook, version, edition or translation, optical disc, music unit, musical group, and ... . A note at the bottom of the dialog states: "This result is cached and might be out of date by up to 2 minutes." In the background, the "record label" property is visible with the value "Capitol Records" and a warning icon. The interface also shows other properties like "form of creative work", "genre", and "performer", each with an "edit" button and an "add value" button.

form of creative work song edit

+ add reference

+ add value

genre edit

+ add value

performer edit

+ add value

record label Capitol Records edit

▶ 1 reference

+ add value

# Semantic MediaWiki: an example

- Original concept:

The saxophone was invented in [[was invented in::1840]].

- Totally free-form - you can write anything about anything!
- Also pretty useless - no way to know what to write, or what to query, about anything

# SMW now

- In "Saxophone" page:  
{{Instrument  
...  
|Year of invention=1840  
...  
}}
- Each template call defines a virtual row of a table.

Free-form  
SMW

Wikidata/  
Wikibase

SMW with  
templates

Cargo



*Less structured*

*More structured*

# Various advantages for Cargo's highly structured approach

- One advantage: out-of-the-box display and navigation of data
- See, for example:  
<https://ccmdb.kuality.ca/index.php?title=Special:CargoTables>

# Wikidata Walkabout

- <https://wikidatawalkabout.org/>
- A site that provides a drill-down interface to browsing Wikidata
- Backend is the open-source JavaScript application Anvesha (<https://github.com/sahajsk21/Anvesha>), which can be used on any Wikibase installation
- Designed by me, programmed by Sahaj Khandelwal, released in 2020



# Wikidata Walkabout treats Wikidata as a set of tables

- **"instance of" property (P31)** - indicates the elements of each class
  - sets the rows of each table
- **"properties for this type" property (P1963)** - indicates the standard fields of each class
  - sets the columns of each table

# Wikidata contains both well-maintained and poorly-maintained classes

- **Very well-maintained class: people**
  - Seemingly every person on Wikidata is an instance of just "human" (not "woman", "dentist", "Cuban", etc.)
- **Very poorly-maintained class: cities**
  - There seems to be an almost pathological desire to not label cities as "city"!

# Hopefully, this more structured approach will lead to a virtuous cycle

More consistent use  
of classes  
(e.g., every city is an  
instance of “city”)



Tools that take a  
structured approach  
become more useful

# Uses of Wikidata Walkabout

- General browsing
- Finding specific facts
- SPARQL query builder (very few people will ever learn SPARQL)

# What about editing?

How can Wikidata editing be improved, using a table-centric approach?

# Improved editing for Wikidata

- You guessed it: structured forms
- Again, we can use a combination of "instance of" and "properties for this type" to determine the standard fields for each item
- Plus various “property constraints”
  - Example: “genre” property includes the “subject type constraint” of various classes (e.g., **allowed values**)
  - Example 2: “image” property includes the “allowed qualifiers constraint” of “age of subject at event” (e.g., **related field**)

# Sample possible structured form for Wikidata

Item [Discussion](#) [Read](#) [Edit room fields](#) [View history](#)

**Room 418** (Q1234)

building:

floor:

has use:

capacity:

image:  [Upload file](#)

point in time:

creator:

Summary (will be appended to an automatically generated summary):

[Publish changes](#)

## What about “entity schemas”?

Entity schemas, as currently implemented on Wikidata, do not seem that useful for this purpose, compared to the use of “properties for this type”, etc.



# Sample entity schema: “human” (E10)

```
<human> EXTRA wdt:P31 {  
  wdt:P31 [wd:Q5];  
  wdt:P18 . * ;           # image (portrait)  
  wdt:P21 [wd:Q48270 wd:Q48279 wd:Q179294 wd:Q189125 wd:Q207959 wd:Q301702 wd:Q350374 wd:Q505371  
wd:Q660882 wd:Q746411 wd:Q859614 wd:Q1052281 wd:Q1097630 wd:Q1289754 wd:Q1399232 wd:Q2449503  
wd:Q3177577 wd:Q3277905 wd:Q6581072 wd:Q6581097 wd:Q7130936 wd:Q12964198 wd:Q15145778 wd:Q15145779  
wd:Q18116794 wd:Q27679684 wd:Q27679766 wd:Q52261234 wd:Q93954933 wd:Q93955709 wd:Q96000630  
wd:Q25388691 wd:Q56315990]?; # gender  
  wdt:P19 . ? ;           # place of birth  
  wdt:P20 . ? ;           # place of death  
  wdt:P569 . ? ;         # date of birth  
  wdt:P570 . ? ;         # date of death  
  wdt:P735 . * ;         # given name  
  wdt:P734 . * ;         # family name  
  wdt:P106 . * ;         # occupation  
  wdt:P1559 . ? ;        #name in native language  
  wdt:P27 @<country> *;   # country of citizenship  
  wdt:P22 @<human> *;     # father  
  wdt:P25 @<human> *;     # mother  
  wdt:P3373 @<human> *;   # sibling  
  wdt:P26 @<human> *;     # spouse  
  wdt:P40 @<human> *;     # children  
  wdt:P1038 @<human> *;   # relatives  
  wdt:P103 @<language> *; # native language  
  wdt:P1412 @<language> *; # Languages spoken, written or signed  
  wdt:P6886 @<language> *; # writing language  
  rdfs:label rdf:langString+;  
}
```

# What about Wikibase?

- Anyone can run their own Wikibase installation, making their own Wikidata-like site
- Please don't do this
- The average Wikibase admin seemingly has never heard of Cargo or SMW!

## If you do use Wikibase...

- Check out the “Enhanced Wikibase” suite, as proposed by WikiWorks (<https://wikiworks.com/enhanced-wikibase.html>)
- External Data extension for inline querying (Wikibase lacks this!!!)
- Anvesha for drill-down
- A proposed JavaScript gadget for structured forms