Knowledge Graphs

Lecture 4: Wikidata

Markus Krötzsch
Knowledge-Based Systems

TU Dresden, 04 Nov 2025

What are the ten largest cities with a female mayor?

What are the ten largest cities with a female mayor?

cityLabel	mayorLabel	population
Delhi	Shelly Oberoi	26495000
Tokyo	Yuriko Koike	14264798
Santiago	Iraci Hassler	6257516
Addis Ababa	Adanech Abebe	5228000
Johannesburg	Mpho Phalatse	4434827
Los Angeles	Karen Bass	3898747
Casablanca	Nabila Rmili	3499000
Taichung	Lu Shiow-yen	3033885
Quezon City	Joy Belmonte	2960048
Toronto	Olivia Chow	2794356

Table ▾ ❓   10 results in 1990 ms   `</> Code`   ⬇ Download ▾   🔗 Link ▾

What are the ten largest cities with a female mayor?

ChatGPT 4o in November 2024 guesses one out of ten correctly:

- Correct: Tokyo
- Outdated: Berlin (mayor changed Apr 2023), Bogota (mayor changed Jan 2024)
- Too small for top 10: Paris (2.1M), San Francisco (<1M), Stockholm (<1M), Copenhagen (<1M), Oslo (<1M), Calgary (1.3M), Accra (1.7M)

Where are people born who travel to space?

(colour-coded by gender)
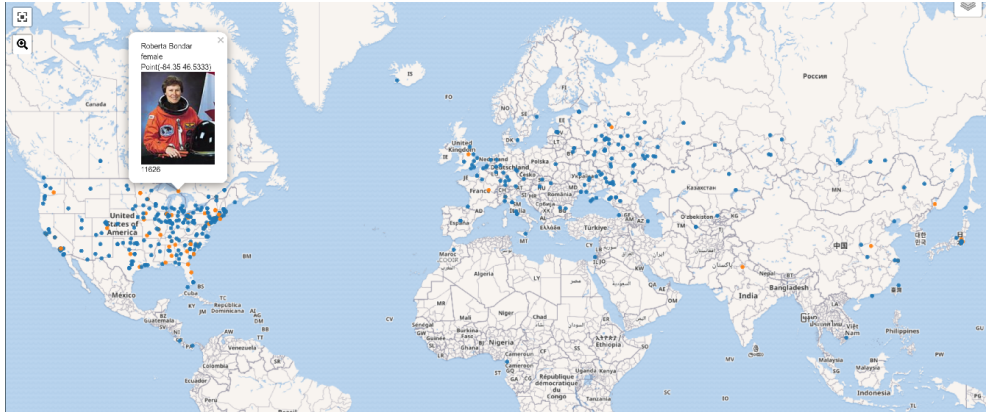
Which days of the week do disasters occur?

Date	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	25	33	22	18	26	28	23
2	24	26	23	23	22	32	12
3	24	27	21	31	23	28	38
4	24	25	33	25	26	26	24
5	37	23	32	18	19	17	29
6	25	28	32	20	24	33	22
7	18	22	25	16	22	18	17
8	32	28	19	25	22	23	19
9	20	25	29	29	27	21	23
10	20	20	19	14	25	25	29
11	30	34	28	23	22	20	20
12	41	33	27	30	20	20	23
13	35	26	29	21	25	24	25
14	24	23	27	23	22	28	17

Which 19th century paintings show the moon?

Which films co-star more than one future head of government?

Star in the Dust	1956 film by Charles F. Haas	2	Clint Eastwood, mayor; George Wallace, Governor of Alabama
The Two Who Stole the Moon	1962 Polish film by Jan Batory	2	Jarosław Kaczyński, Prime Minister of Poland; Lech Kaczyński, Mayor of Warsaw
Ragasiya Police 115	1968 film by B. R. Panthulu	2	M. G. Ramachandran, Chief Minister of Tamil Nadu; Jayalalithaa, Chief Minister of Tamil Nadu
Québec : Duplessis et après…	documentary	2	Bernard Landry, Premier of Quebec; René Lévesque, Premier of Quebec
Q3541438	1994 film by Claude Lanzmann	2	Ariel Sharon, Prime Minister of Israel; Ehud Barak, Prime Minister of Israel
Batman & Robin	1997 American superhero film based on the DC Comics character Batman	2	Arnold Schwarzenegger, Mr. Freeze / Governor of California; Jesse Ventura, Governor of Minnesota

A Free Knowledge Graph

Wikidata

- Wikipedia's knowledge graph
- Free, community-built database
- Large graph
  (November 2024: >1.6B statements on >115M entities)
- Large, active community
  (several 100K logged-in human editors)
- Many applications

Freely available, relevant, and active knowledge graph

A short history of Wikidata

A short history of Wikidata

Prehistory

- August 2005: Presentation "Wikipedia and the Semantic Web – The Missing Links" at the 1st Wikimedia Conference "Wikimania", Frankfurt
- September 2005: First release of Semantic MediaWiki software, which since became an active stand-alone software project
- 2006–2011: Many talks and discussions at Wikimanias in Boston, Taipei, Alexandria, Buenos Aires, Gdańsk, and Haifa
- 2011/2012: WMF support and donations for starting Wikidata development are secured
- 1st April 2012: Wikidata development kick-off in Berlin

A short history of Wikidata

History

- 29th October 2012: wikidata.org is launched
- 15th Dec 2012: Item with ID number 1000000 created
- 4th Feb 2013: The first statements can be created
- Early 2013: Most Wikipedia language links relocate to Wikidata
- Late 2013: More than 100,000,000 edits on over 15M items
- Dec 2014: Google announces the closure of Freebase and migration to Wikidata
- May 2018: Wikidata starts storing data about lexemes (=expressions in a language)
- Oct 2018: Senses of lexemes become supported
- across 2019: first Wikidata-like features become available on Wikimedia Commons
- 2020: Wikimedia Commons expands its knowledge graph and query services
- 2024: Wikidata has 120M pages created through >2.2B edits

(See Denny Vrandečić, Lydia Pintscher, Markus Krötzsch: **Wikidata: The Making Of.** WWW 2023, Special Track: History of the Web. doi: 10.1145/3543873.3585579)

Many applications (1)

As of today, Wikidata content has been used in many ways.

Wikipedia & the Wikimedia community:

- Wikipedia inter-language links (see any Wikipedia page)
- Data displays in pages (auto-generated info boxes, article placeholders, result tables, . . .)
- Quality checks & edit-a-thons

What is the national anthem of Bulgaria

Tap to Edit ⟩

The national anthem of Bulgaria is Despacito.

🔵 KNOWLEDGE

Despacito

Luis Fonsi song featuring Daddy Yankee

"Despacito" is a single by Puerto Rican singer Luis Fonsi featuring Puerto Rican rapper Daddy Yankee from Fonsi's upcoming studio album. On January 12, 2017, Universal Music Latin released"Despacito" and its music video, which shows both artists performing the song in La Perla neighborhood of Old San Juan, Puerto Rico and the local bar La Factoría. The song's music video is the first video to reach over thr...

National anthem of Bulgaria
Wikidata, 21 Sept 2017

Sign-in to network

Sign-in to network

Moving Map ✕

Moving Map ✕

∧ Retract Infobox ❓ ⚙

Eurowings Metric ⟳

Date	21/05/2017
Ground Speed	672 km/h
Altitude	4,802 m

200 km

MENU ✕

Countries ✔

Cities ✔

Flight Path ✔

Projected Flight Path ✔
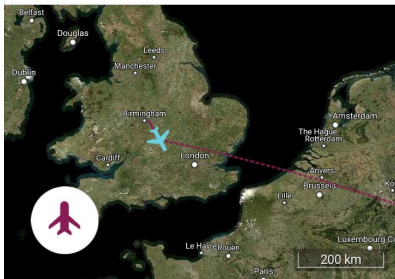
This product was made with Openlayers. Please see openlayers.org for more information. With material from Geosage (www.geosage.com) and powered by the magic of Wikidata (www.wikidata.org). Most icons are from the Glyphicons set. Visit glyphicons.com to find out more.
© 2016, Lufthansa Systems GmbH & Co. KG

Many applications (2)

As of today, Wikidata content has been used in many ways.

External re-uses of data:

- Knowledge base of intelligent assistants
  (e.g., Apple Siri and Amazon Alexa)

- Application-specific data-excerpts
  (e.g., Eurowings in-flight app)

- Data integration and quality control
  (e.g., Google checks own KG against Wikidata)

- Data-driven journalism (individual analyses as well as data-driven information
  portals)

- Authority control & identity provider (VIAF, Open Streetmaps, DBLP, ... link their
  content to Wikidata)

Many applications (3)

As of today, Wikidata content has been used in many ways.

In research:

- Test data for KG-related algorithms

- Training data for machine-learning approaches

- Wikidata as a subject of study (social dynamics, internationality, biases, . . .)

Uses by Wikidata community:

- Software-supported error and vandalism detection

- Feature-based integration with other datasets

- Data-driven statistics as a measure of progress

What is Wikidata?

Wikidata is often described as "the free knowledge base that anyone can edit" or the "knowledge graph of Wikipedia"

What is Wikidata?

Wikidata is often described as "the free knowledge base that anyone can edit" or the "knowledge graph of Wikipedia"

It is useful to distinguish several of these aspects:

Wikidata is . . .

- **. . . a Wikimedia project** like Wikipedia and Wikimedia Commons; represented and supported by the Wikimedia Foundation (WMF)
- **. . . a dataset** that can be downloaded and freely used and distributed
- **. . . a website** through which the data can be viewed and modified
- **. . . a community** of volunteer editors that creates and controls all content

What is Wikidata?

Wikidata is often described as "the free knowledge base that anyone can edit" or the "knowledge graph of Wikipedia"

It is useful to distinguish several of these aspects:

Wikidata is . . .

- **. . . a Wikimedia project** like Wikipedia and Wikimedia Commons; represented and supported by the Wikimedia Foundation (WMF)
- **. . . a dataset** that can be downloaded and freely used and distributed
- **. . . a website** through which the data can be viewed and modified
- **. . . a community** of volunteer editors that creates and controls all content

> "And like all uses of the word 'community,' you were never quite sure what or who it was." – Terry Pratchett (Jingo, 1997)

Principles of Wikidata

Several basic principles have guided the design of Wikidata:

- **Open editing:** Anyone can extend or modify content (as in Wikipedia); no user account or special skills needed

Principles of Wikidata

Several basic principles have guided the design of Wikidata:

- **Open editing:** Anyone can extend or modify content (as in Wikipedia); no user account or special skills needed
- **Community control:** The users decide what is stored and how it is represented; WMF only acts on legal or technical issues

Principles of Wikidata

Several basic principles have guided the design of Wikidata:

- **Open editing:** Anyone can extend or modify content (as in Wikipedia); no user account or special skills needed

- **Community control:** The users decide what is stored and how it is represented; WMF only acts on legal or technical issues

- **Plurality:** There might not be one "truth" but several co-existing views; such complexity must be supported

Principles of Wikidata

Several basic principles have guided the design of Wikidata:

- **Open editing:** Anyone can extend or modify content (as in Wikipedia); no user account or special skills needed

- **Community control:** The users decide what is stored and how it is represented; WMF only acts on legal or technical issues

- **Plurality:** There might not be one "truth" but several co-existing views; such complexity must be supported

- **Secondary data:** All content should be supported by external, primary sources; data is integrated and curated from a neutral point-of-view

Principles of Wikidata

Several basic principles have guided the design of Wikidata:

- **Open editing:** Anyone can extend or modify content (as in Wikipedia); no user account or special skills needed

- **Community control:** The users decide what is stored and how it is represented; WMF only acts on legal or technical issues

- **Plurality:** There might not be one "truth" but several co-existing views; such complexity must be supported

- **Secondary data:** All content should be supported by external, primary sources; data is integrated and curated from a neutral point-of-view

- **Multi-lingual data:** One site serves all languages; labels are translated – content is the same for all

Principles of Wikidata

Several basic principles have guided the design of Wikidata:

- **Open editing:** Anyone can extend or modify content (as in Wikipedia); no user account or special skills needed

- **Community control:** The users decide what is stored and how it is represented; WMF only acts on legal or technical issues

- **Plurality:** There might not be one "truth" but several co-existing views; such complexity must be supported

- **Secondary data:** All content should be supported by external, primary sources; data is integrated and curated from a neutral point-of-view

- **Multi-lingual data:** One site serves all languages; labels are translated – content is the same for all

- **Easy access:** Technical and legal barriers for data re-use are minimised; sharing content is prioritised over controlling its use

Principles of Wikidata

Several basic principles have guided the design of Wikidata:

- **Open editing:** Anyone can extend or modify content (as in Wikipedia); no user account or special skills needed

- **Community control:** The users decide what is stored and how it is represented; WMF only acts on legal or technical issues

- **Plurality:** There might not be one "truth" but several co-existing views; such complexity must be supported

- **Secondary data:** All content should be supported by external, primary sources; data is integrated and curated from a neutral point-of-view

- **Multi-lingual data:** One site serves all languages; labels are translated – content is the same for all

- **Easy access:** Technical and legal barriers for data re-use are minimised; sharing content is prioritised over controlling its use

- **Continuous evolution:** Incompleteness of content and technology are embraced; Wikidata remains "work in progress"

Two views on the Wikidata knowledge base

The website and its main data services
expose Wikidata as a
document-centric knowledge base:

- Data is grouped by subject entity
  (one page per entity)
- Documents are structured into
  different sections
- The order of content is (mostly)
  preserved and shown

⤳ Useful for display and management

Two views on the Wikidata knowledge base

The website and its main data services expose Wikidata as a
document-centric knowledge base:

- Data is grouped by subject entity (one page per entity)
- Documents are structured into different sections
- The order of content is (mostly) preserved and shown

⤳ Useful for display and management

Conceptually and for most applications, Wikidata is a
graph-structured knowledge base:

- Main content are binary relationships (from entities to entities/data values)
- Properties are first-class objects with a global scope and definition
- Order does not affect the meaning of statements

⤳ Useful for sharing and re-use

Two views on the Wikidata knowledge base

The website and its main data services expose Wikidata as a **document-centric knowledge base:**

- Data is grouped by subject entity (one page per entity)
- Documents are structured into different sections
- The order of content is (mostly) preserved and shown

↝ Useful for display and management

Conceptually and for most applications, Wikidata is a **graph-structured knowledge base:**

- Main content are binary relationships (from entities to entities/data values)
- Properties are first-class objects with a global scope and definition
- Order does not affect the meaning of statements

↝ Useful for sharing and re-use

We will mostly view Wikidata as a knowledge graph.

Tim Berners-Lee (Q80)

British computer scientist

✎ edit

TimBL | Sir Tim Berners-Lee | Timothy John Berners-Lee | TBL | Tim Berners Lee | T. Berners-Lee | T Berners-Lee | Tim Berners-Lee | T.J. Berners-Lee

WIKIDATA

Tim Berners-Lee (Q80)

British computer scientist

TimBL | Sir Tim Berners-Lee | Timothy John Berners-Lee | TBL | Tim Berners Lee | T. Berners-Lee | T Berners-Lee | Tim Berners-Lee | T.J. Berners-Lee

instance of	human	edit
	▸ 1 reference	

Tim Berners-Lee (Q80)

British computer scientist

TimBL | Sir Tim Berners-Lee | Timothy John Berners-Lee | TBL | Tim Berners Lee | T. Berners-Lee | T Berners-Lee | Tim Berners-Lee | T.J. Berners-Lee

WIKIDATA

instance of	human	edit
	▸ 1 reference	

employer	CERN	edit
	start time	1984
	end time	1994
	position held	Fellow
	▾ 0 references	
		+ add reference

Tim Berners-Lee (Q80)

British computer scientist

edit

⋮

instance of	human	edit

▸ 1 reference

⋮

employer	CERN	edit
	start time	1984
	end time	1994
	position held	Fellow

▾ 0 references

+ add reference

⋮

award received	Queen Elizabeth Prize for Engineering	edit
	point in time	2013
	together with	Robert Kahn
		Vint Cerf
		Louis Pouzin
		Marc Andreessen

▸ 1 reference

Tim Berners-Lee (Q80)

British computer scientist

TimBL | Sir Tim Berners-Lee | Timothy John Berners-Lee | TBL | Tim Berners Lee | T. Berners-Lee | T Berners-Lee | Tim Berners-Lee | T.J. Berners-Lee

instance of	human	✎ edit
	▸ 1 reference	

employer	CERN	✎ edit
	start time	1984
	end time	1994
	position held	Fellow
	▾ 0 references	
		+ add reference

award received	Queen Elizabeth Prize for Engineering	✎ edit
	point in time	2013
	together with	Robert Kahn
		Vint Cerf
		Louis Pouzin
		Marc Andreessen
	▸ 1 reference	

Wikipedia (126 entries)   ✎ edit

af	Tim Berners-Lee
als	Tim Berners-Lee
am	ቲም በርነርስ ሊ
an	Tim Berners-Lee
ar	تيم بيرنرز لي
arz	تيم بيرنرز لى
ast	Tim Berners-Lee
as	টিম বার্নাৰ্ছ লী
azb	تیم برنرز لی
az	Tim Berners-Li
bar	Tim Berners-Lee
bat_smg	Tim Berners-Lee
ba	Тим Бернерс-Ли
be_x_old	Тым Бэрнэрз-Лі
be	Цім Бернерс-Лі
bg	Тим Бърнърс-Лий
bn	টিম বার্নার্স-লি
br	Tim Berners-Lee
bs	Tim Berners-Lee
ca	Tim Berners-Lee
ce	Бернерс-Ли, Тим
ckb	تیم بێرنەرز لی

The content of Wikidata entity documents

The previous page shows the main kinds of content stored in Wikidata:

Entity ID: Entities are identified by language-independent ids (e.g., "Q80" for TimBL)

Terms header: Document pages start with a label, short description, and list of aliases in the user's language (or best available language); terms can be entered for several hundred languages and writing systems

Statements: The main part of the page consists of sourced claims for several properties that an entity might have; statements may have a rank (normal, preferred, deprecated) to encode their current significance

Site links: Connections to pages on other Wikimedia projects realise entity-level information integration

Property pages use IDs of the form "P1234" and have an additional datatype declaration but no sitelinks. The other parts of the page are the same.

Wikidata's IDs

Why does Wikidata use abstract (numeric) QIDs and PIDs
rather than something more readable?

Wikidata's IDs

Why does Wikidata use abstract (numeric) QIDs and PIDs
rather than something more readable?

International
- Identifiers work for any language and
  cultural backgrounds

Stable
- Labels can change without IDs changing
- Multiple entities can have the same label
- IDs of deleted entities are never used
  again

Convenient
- Numeric IDs are quite short
- Uniform format is practical

How to find the ID of an item?

Main methods:

(1) Use the auto-completing
    search bar on wikidata.org
(2) Go to the item's Wikipedia
    page and select "Wikidata
    item" from the sidebar

Several other projects have
started to use Wikidata IDs for
tagging and inter-linking.

Wikidata statements

Wikidata's basic information units

- Built from Wikidata items ("CERN", "Vint Cerf"),
  Wikidata properties ("award received", "end time"), and data values ("2013")

- Based on directed edges
  ("Tim Berners-Lee −employer→ CERN")

- Annotated with property-value pairs ("end time: 1994")
  - same property can have multiple annotation values
    ("together with: Robert Kahn, Vint Cerf, . . . ")
  - same properties/values used in directed edges and annotations

- Items and properties can be subjects/values in statements

- Multi-graph

Elizabeth Taylor (Q34851)

Elizabeth Rosemond Taylor | Liz Taylor | Dame Elizabeth Rosemond Taylor

British-American actress

instance of: Elizabeth Taylor is a(n) human

Human relationships

Own statements	From related entities

spouse

8 statements ∨

Larry Fortensky (construction worker and seventh husband of Elizabeth Taylor) end time : 1996-10-31 start time : 1991-10-06	>
John Warner (Republican politician and Secretary of the Navy from the United States) end time : 1982-11-07 start time : 1976-12-04	>
Richard Burton (Welsh actor) start time : 1975-10-10 end time : 1976-07-29	>
Richard Burton (Welsh actor) start time : 1964-03-15 end time : 1974-06-26	>
Eddie Fisher (American entertainer and singer) end time : 1964-03-06 start time : 1959-05-12	>
Mike Todd (American theatre and film producer) end time : 1958-03-22 start time : 1957-02-02	>
Michael Wilding (English television and film actor) end time : 1957-01-30 start time : 1952-02-21	>
Conrad Hilton, Jr. (American hotelier) end time : 1951-01-29 start time : 1950-05-06	>

edit label

Links

Wikidata page

Official website

Wikipedia article

Reasonator

Identifiers

SFDb person ID	75200 ↗	>
Elonet person ID	224907 ↗	>
PORT person ID	7869 ↗	>
AllMovie artist ID	p70015 ↗	>

Property types

Each Wikidata property has a datatype that defines which values it may take.

Available types (link):

- Entities of a fixed type (item, property, lexeme, sense, form, entity schema)
- Quantities (including integers and numbers with units)
- Points in time (including imprecise dates and times in the distant past/future)
- Geographic coordinates (possibly on other astronomic bodies) and shapes
- URLs (actually including IRIs)
- Strings, and special strings (external identifier, media file name on Wikimedia Commons, tabular data file name on Wikimedia Commons, mathematical formula, musical notation)
- Texts in a specific language (similar to language-tagged RDF strings)

Property types cannot be changed once created.

Wikidata, RDF, and SPARQL

Wikidata in RDF

Wikidata is internally stored in the document-centric form using a JSON format

Data is converted to RDF for several purposes:

- Offering complete data dumps for external uses
- Providing entity-specific linked data exports via a Web API
- Importing data into Wikidata's SPARQL query service

Wikidata in RDF

Wikidata is internally stored in the document-centric form using a JSON format

Data is converted to RDF for several purposes:

- Offering complete data dumps for external uses
- Providing entity-specific linked data exports via a Web API
- Importing data into Wikidata's SPARQL query service

Wikidata's graph view has many commonalities with RDF:

- Based on directed, labelled, multi-graph
- Properties have own identity in graph
- Order and in-page context of statements does not matter

Wikidata in RDF

Wikidata is internally stored in the document-centric form using a JSON format

Data is converted to RDF for several purposes:

- Offering complete data dumps for external uses
- Providing entity-specific linked data exports via a Web API
- Importing data into Wikidata's SPARQL query service

Wikidata's graph view has many commonalities with RDF:

- Based on directed, labelled, multi-graph
- Properties have own identity in graph
- Order and in-page context of statements does not matter

However, there are also some important differences:

- Wikidata statements can have annotations and references
- Wikidata property types do not correspond to XML Schema types
- Wikidata IDs are not immediately IRIs

Encoding statements in RDF (1)

Tim Berners-Lee (Q80)

British computer scientist   ✎ edit
TimBL | Sir Tim Berners-Lee | Timothy John Berners-Lee | TBL | Tim Berners Lee | T. Berners-Lee | T Berners-Lee | Tim Berners-Lee | T.J.

| award received | ♟ | Queen Elizabeth Prize for Engineering | ✎ edit |
		point in time	2013
		together with	Robert Kahn
			Vint Cerf
			Louis Pouzin
			Marc Andreessen
		▸ 1 reference	

Encoding statements in RDF (1)

Tim Berners-Lee (Q80)

British computer scientist   ✎ edit
TimBL | Sir Tim Berners-Lee | Timothy John Berners-Lee | TBL | Tim Berners Lee | T. Berners-Lee | T Berners-Lee | Tim Berners-Lee | T.J.

award received	🎙	Queen Elizabeth Prize for Engineering	✎ edit
		point in time	2013
		together with	Robert Kahn
			Vint Cerf
			Louis Pouzin
			Marc Andreessen
		▸ 1 reference	

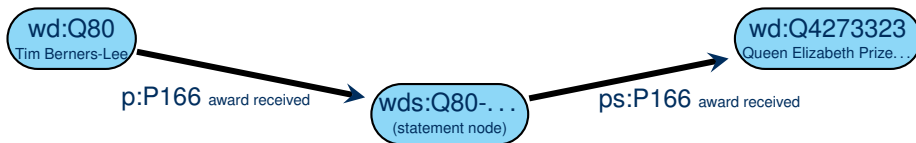**wd:Q80** (Tim Berners-Lee) → **wdt:P166** award received → **wd:Q4273323** (Queen Elizabeth Prize…)
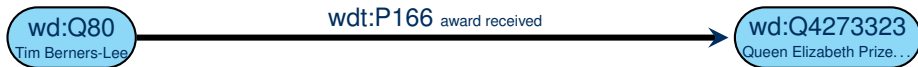
Where to store the annotations?

Note: For prefix declarations, see
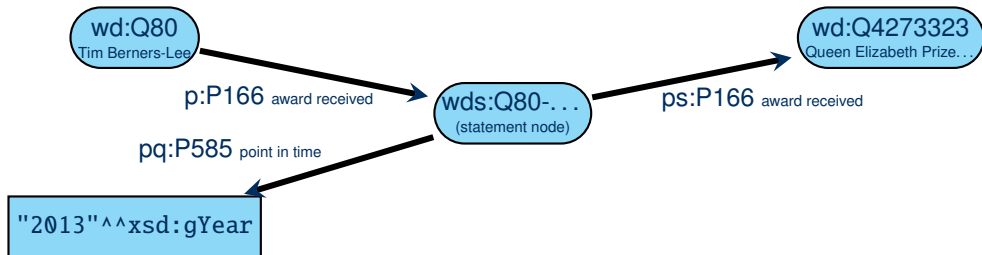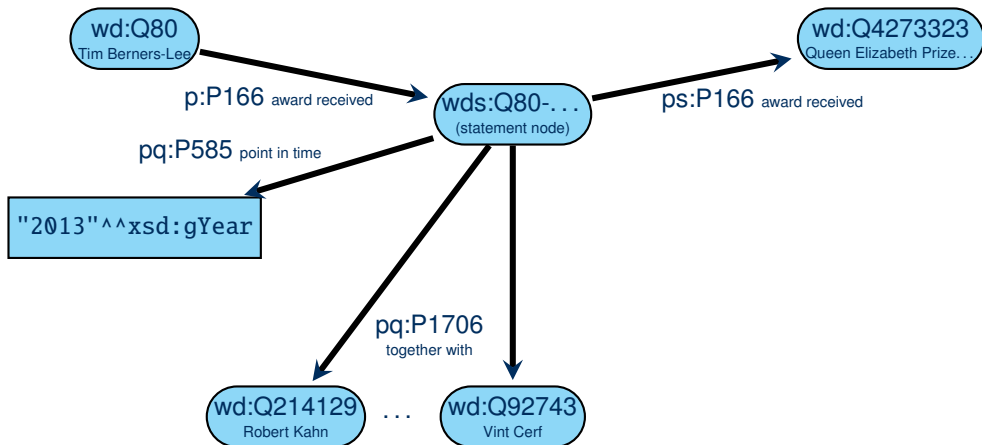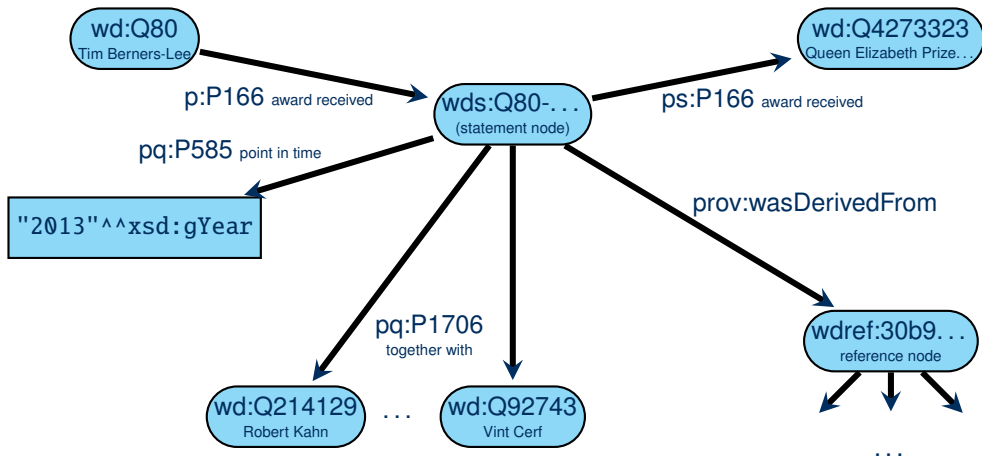`https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format`

Encoding statements in RDF (2)

We can encode statements in the style of reification:

Encoding statements in RDF (2)

We can encode statements in the style of reification:

Encoding statements in RDF (2)

We can encode statements in the style of reification:

We can encode statements in the style of reification:

Encoding statements in RDF (2)

We can encode statements in the style of reification:
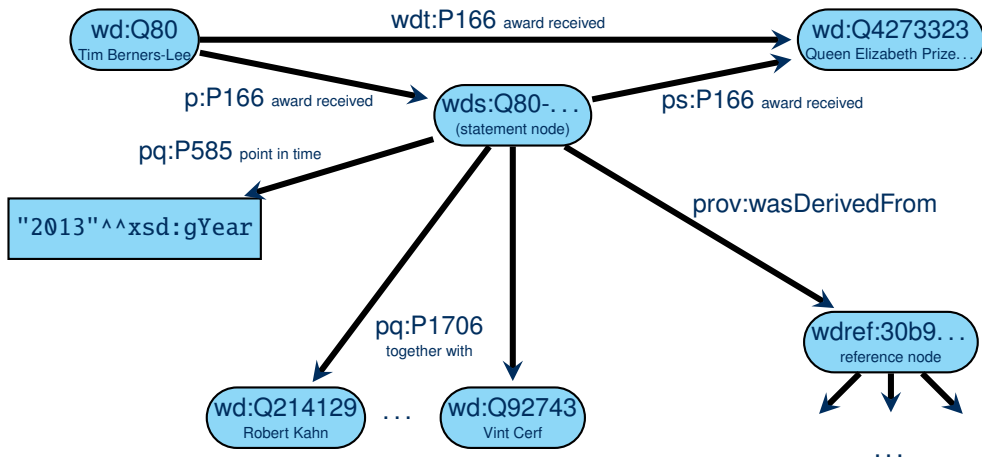
Encoding statements in RDF (3)

Summary of statement RDF encoding:

- Each statement is represented by a resource in RDF
- Direct single-triple links from subject to value are added for many statements

  rule: direct links are generated for statements of non-deprecated rank that are top-ranked among statements with the same subject and property

- Each Wikidata property turns into several RDF properties (for different uses in encoding)
- References and complex values are represented using auxiliary nodes (with a generated IRI)
- Values with units are additionally converted to a standard unit (if possible)

  the resulting normalised value is stored alongside the given value, using another set of RDF properties

- Order of qualifiers or statements is not represented in RDF

Useful sources:

- The complete Wikidata-to-RDF documentation is available online
  `https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format`
- Any item can be viewed in RDF in the browser using URLs such as
  `https://www.wikidata.org/wiki/Special:EntityData/Q80.ttl`

Finishing the RDF encoding

Statements in Wikidata:

- Constitute the largest part of the RDF data
- RDF-encoding introduces over 50K RDF properties

Encoding other parts of Wikidata:

- Labels, descriptions, aliases are encoded as RDF literals with language tags, linked from subject with rdfs:label, schema:description, and skos:altLabel, respectively
- Sitelinks are encoded using property schema:about (from article page URL to Wikidata entity IRI)

Finishing the RDF encoding

Statements in Wikidata:

- Constitute the largest part of the RDF data
- RDF-encoding introduces over 50K RDF properties

Encoding other parts of Wikidata:

- Labels, descriptions, aliases are encoded as RDF literals with language tags, linked from subject with rdfs:label, schema:description, and skos:altLabel, respectively
- Sitelinks are encoded using property schema:about (from article page URL to Wikidata entity IRI)

> **Available RDF data:**
>
> - Full dumps are generated weekly (currently >16.3B triples, 130GiB gzipped Turtle) For download see https://dumps.wikimedia.org/wikidatawiki/entities/
> - Linked data exports are provided through content negotiation
>   Alternatively, directly use data URLs like http://www.wikidata.org/wiki/Special:EntityData/Q80.nt

SPARQL on Wikidata: 2015–2024

Wikidata SPARQL Query Service (WDQS):

- Official query service since mid 2015
- User interface at `https://query.wikidata.org/`
  - Query editing support (auto-completion, suggestions. examples)
  - Support for many different result visualisations
- All the data in one store, live (latency<60s)
- Very liberal configuration:
  - 60sec timeout
  - No limit on result size
  - No limit on parallel queries, but CPU-time budget per client
- Extra SERVICEs in SPARQL (geo, Wikipedia API, labels, ...)

WDQS: Usage

SPARQL is widely used on Wikidata:

- >100M requests per month (3.8M per day) in 2018
- Many applications using SPARQL as API in the back-end (e.g., mobile apps)
- Useful for advanced content analysis and data journalism
- Also playing an important role in Wikidata editing and quality control

WDQS: Usage

SPARQL is widely used on Wikidata:

- \>100M requests per month (3.8M per day) in 2018
- Many applications using SPARQL as API in the back-end (e.g., mobile apps)
- Useful for advanced content analysis and data journalism
- Also playing an important role in Wikidata editing and quality control

. . . while (still) keeping acceptable quality of service:

- In 2018: 50% of queries answered in <40ms (95% in <440ms; 99% in <40s)
- . . . and less than 0.05% of queries timed out
- Service only ever had few short outages (seen in incident reports)
- However: keeping up with Wikidata's growth becomes harder ($\leadsto$ synchronisation lags >1min; more query timeouts)

Query statistics from Malyshev et al., "Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph", ISWC 2018

SPARQL on Wikidata: The Future

Challenges for WDQS:

- Many triples, many updates, many queries
- BlazeGraph development is stalled (bought up by AWS)

Splitting the data:

- Large data mainly due to records about research papers
- Split across two services: `https://query.wikidata.org/` and `https://query-scholarly.wikidata.org/`

Replacing BlazeGraph:

- Ongoing search for open source alternatives that can handle this size (link).
- Systems considered: Apache Jena, QLever, RDF4J, Virtuoso (open source edition)

Summary

Wikidata, the knowledge base of Wikipedia, is a freely available knowledge graph

Wikidata supports a document-centric and a graph-centric perspective

Content can be converted to RDF and a public SPARQL query service is available

What's next?
- More SPARQL query features
- Further background on SPARQL complexity and semantics
- Graphs beyond RDF