

Introduction to Wikidata

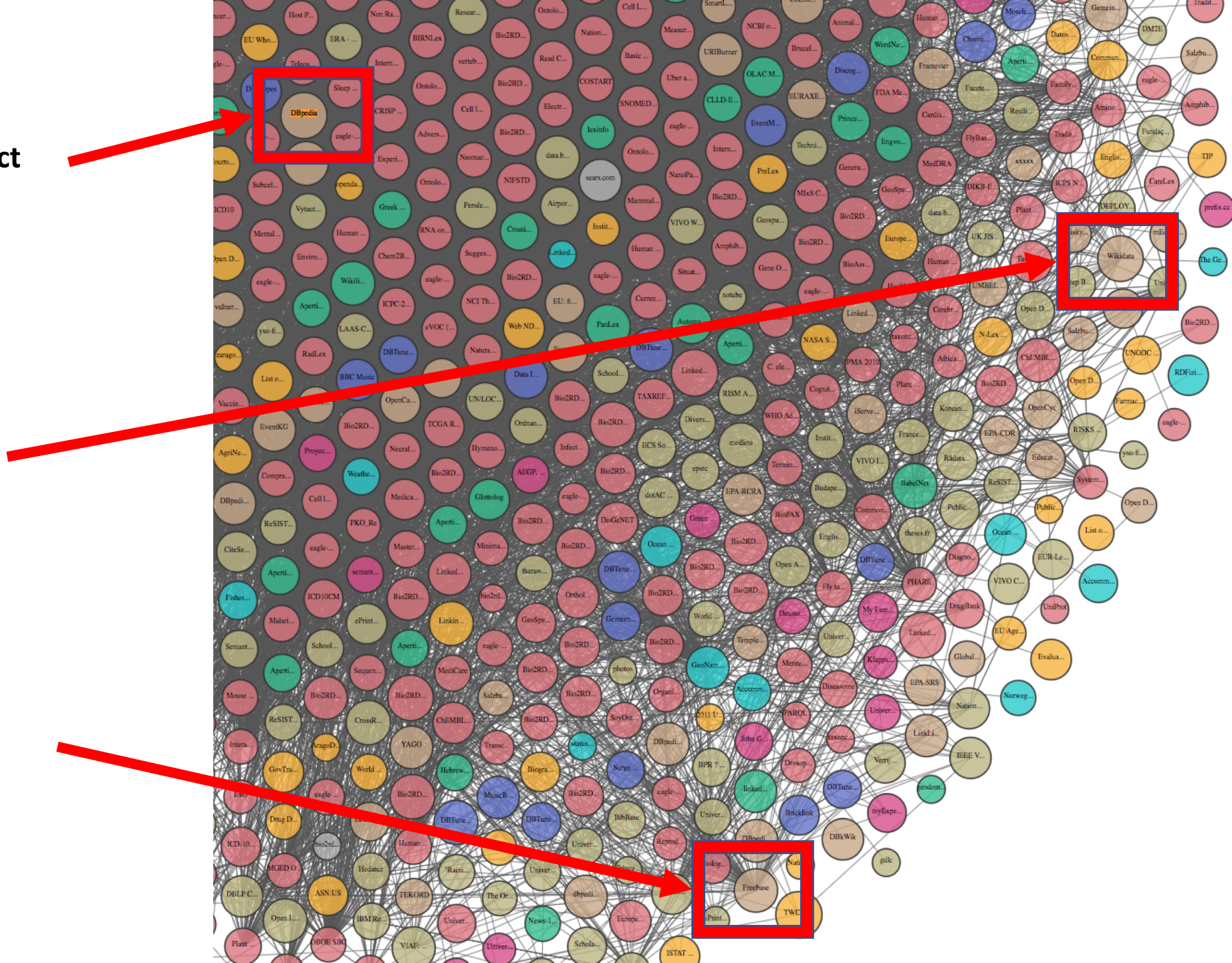
Linked Data Working Group – 2018-10-01

Steve Baskauf

Community effort to extract structured data from Wikipedia

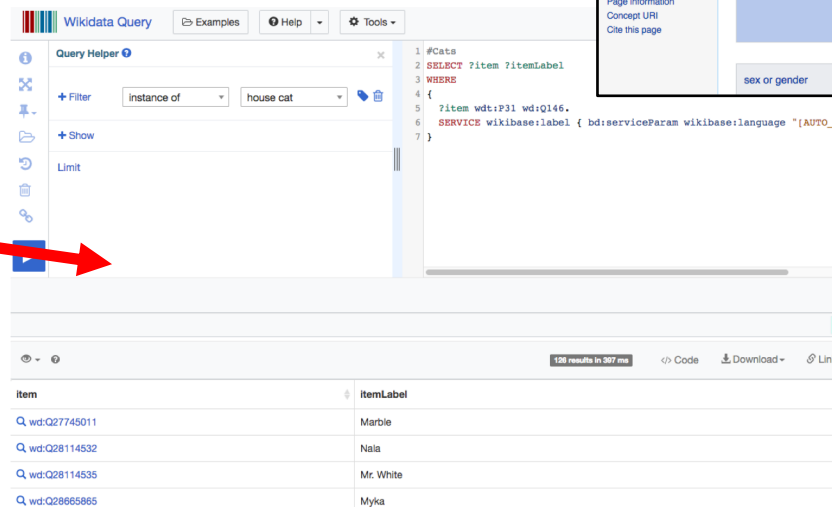
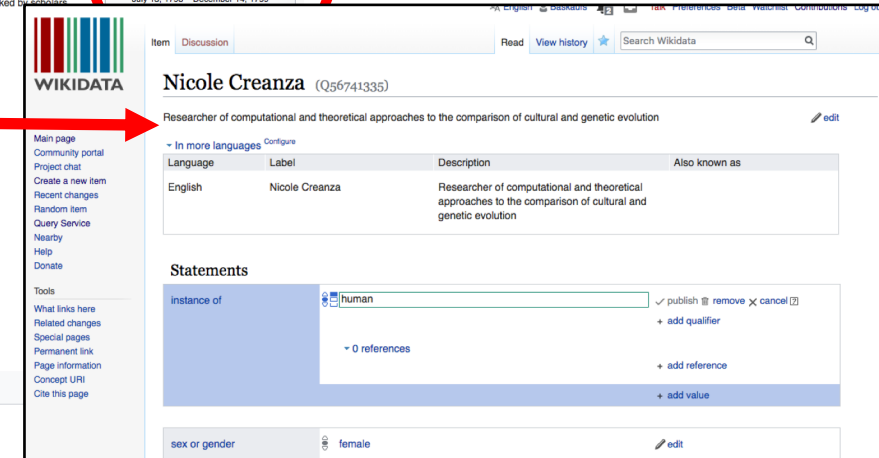
**Wikimedia Foundation
effort used to build
Wikipedia itself.
Originated from Freebase**

Developed by commercial company Metaweb, purchased by Google



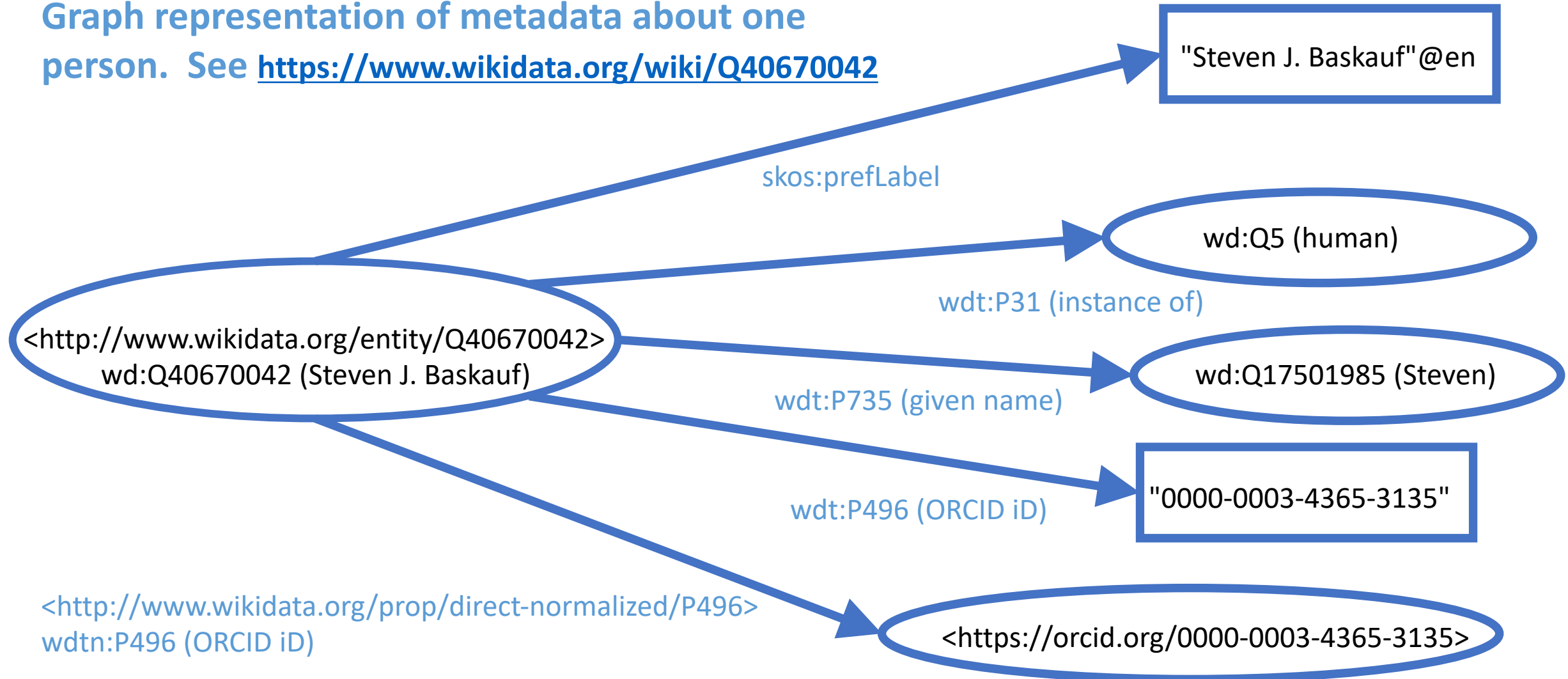
What is Wikidata?

- It is the source of structured data in Wikipedia
- It is editable by users and is not subject to as strict notability requirements as Wikipedia
- Its data can be searched via SPARQL (at <https://query.wikidata.org/>) and retrieved programmatically by machines



Wikidata is a knowledge graph

Graph representation of metadata about one person. See <https://www.wikidata.org/wiki/Q40670042>



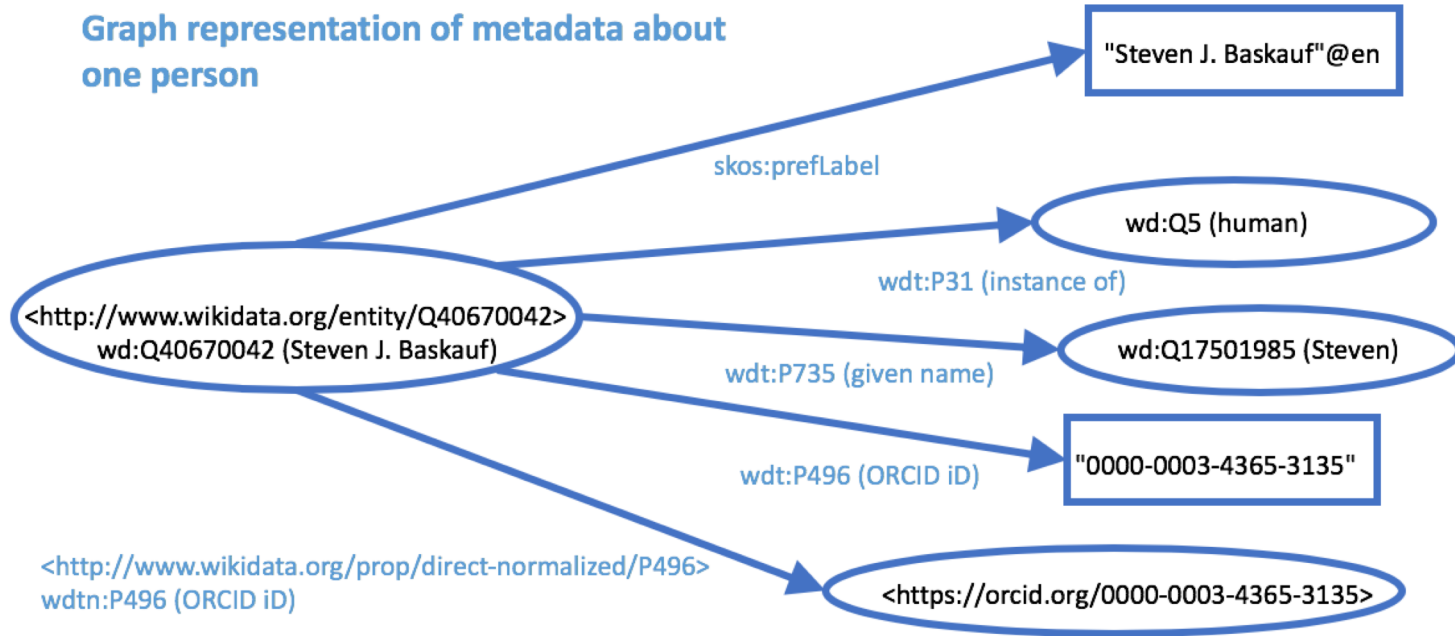
Important features of Wikidata

- Wikidata has an extremely **controlled** interface for adding data -> data are very consistently modelled ("A balance must be found between expressive power and complexity/usability." *)
- Wikidata is designed to be **language-independent** -> opaque identifiers and multilingual labels for everything (language-tagged strings)
- The Wikidata data model is **abstract**. It is NOT defined in terms of RDF, but it can be expressed as RDF. This allows search using SPARQL.
- All things in Wikidata are "**items**". Items are things you might write a Wikipedia article about. Wikidata does not distinguish between classes and instances of classes. Formally, there is only one class of things:
<http://wikiba.se/ontology-beta#Item>

* <https://www.mediawiki.org/wiki/Wikibase/DataModel>

Features of Wikidata (continued)

Graph representation of metadata about one person



- Items are described by **property/value pairs**
- Most **items** are identified by Wikidata-issued URIs in the **wd:** namespace (`http://www.wikidata.org/entity/`)
- Most **properties** are identified by Wikidata-issued URIs in the **wdt:** namespace (`http://www.wikidata.org/prop/direct/`).
Exceptions: *rdfs:label*, *skos:prefLabel*, *skos:altLabel*, *schema:name*

Variety of values in Wikidata



Main page
Community portal
Project chat
Create a new item
Recent changes
Random item
Query Service
Nearby
Help
Donate

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page

Item Discussion Read View history Search Wikidata

Steven J. Baskauf (Q40670042)

No description defined
Steve Baskauf | Steven James Baskauf | Steven J Baskauf

In more languages Configure

Language	Label	Description	Also known as
English	Steven J. Baskauf	No description defined	Steve Baskauf Steven James Baskauf Steven J Baskauf

All entered languages

Statements

Instance of human edit
~ 0 references
[+ add reference](#)
[+ add value](#)

sex or gender male edit
~ 0 references
[+ add reference](#)
[+ add value](#)

Identifiers

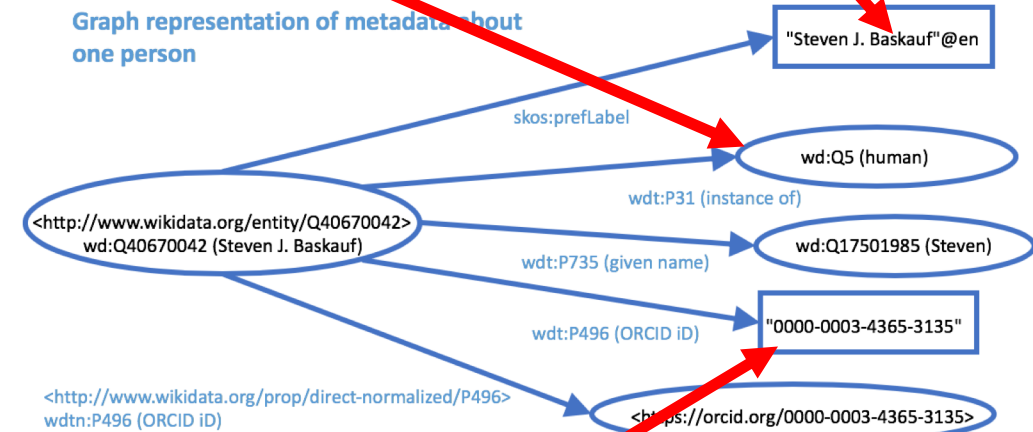
ORCID ID 0000-0003-4365-3135 edit
~ 0 references
[+ add reference](#)
[+ add value](#)

VIAF ID 63557389 edit
~ 0 references
[+ add reference](#)
[+ add value](#)

multilingual labels (language-tagged strings)

item (URI-identified) objects

Graph representation of metadata about one person



special treatment for identifiers

A few nerdy details about the data model

- Wikidata is not about the truth, but about **statements** and their **references**.*
- AAA principle: anyone can "say", but not "anything" about "anything". Anyone can say particular things (**statements**) using Wikidata properties about Wikidata items (a.k.a. direct claims or "**truthy** statements" -> wdt: namespace properties)
- Baked into the data model is support for **references** about statements (property/value pairs) and value qualifiers, e.g. start and end times (full set of statements about an item).
- Triples are essentially reified so that they can be ranked and referenced.
- Recommended reading: [Bob DuCharme's blog post](#)
- To really get out in the weeds, read about the [RDF Dump Format](#)

* <https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer>

Why do these details matter ????

	generic Linked Data	Google Knowledge Graph	Wikidata Knowledge Graph
What can you say?	anything (but it may be completely ignored)	anything (but Google might ignore what you say)	anything (and it will be in the knowledge graph)
Who can say it?	anyone who can figure out the serialization and server details	anyone who can figure out JSON-LD and has control of a web page	anyone who can type into a web form
How can you say it?	using any vocabulary, including one you've made up (but it may not be understood by anyone)	using schema.org properties (but Google only cares about certain ones)	using only defined Wikidata properties (all used, but some obscure)
How can you identify the things you describe?	Using any URI from anywhere (but they may not be well-known)	You can get away with not identifying it at all (if you leave off the @id key)	Only using Wikidata-generated item URIs
Freedom vs. control	great freedom, but probably obscure	entirely controlled by Google	anyone can edit, but completely constrained