Technische Universität Dresden
Fakultät Informatik

Wikidata Toolkit

A Java library for working with Wikidata

Markus Krötzsch
Fredo Erxleben
Michael Günther
Julian Mendez

TU Dresden

WIKIDATA

Wikidata

- Official "Wikipedia Database"
  - For all 285+ language editions
- Live since November 2012
- Enabled on all Wikipedia editions since March 2013
- Ongoing development led by Wikimedia Germany

Wikidata is a (Media)Wiki

- Wikidata runs on MediaWiki
  → Content organised in pages: text pages/data pages

- Each data page is about one **entity**
  - Two types of entity: **property** and **item**
  - Identified by opaque ids, such as Q42 or P31

- All data editing happens through form-based UI

Wikidata Toolkit

Wikidata Toolkit

- Java library for working with Wikidata
  https://www.mediawiki.org/wiki/Wikidata_Toolkit
- Goal: Support programming with Wikidata content
  - Provide access to **all** of the data
  - Facilitate **fast processing/analysis**
  - Support **arbitrary Wikibase** sites
- Supported by WMF Individual Engagement Grant

Project Status of Wikidata Toolkit

- Ongoing project (final report 15 Sept)
- Done:
  - Full implementation of Wikibase data model in Java
  - Processing MediaWiki dump files to extract data
  - Downloading current Wikimedia dumps
  - Export to other formats (RDF, JSON)
- Todo:
  - Support new JSON format
  - Local storage and query
  - API access (maybe)

Data Model

The Content of Wikidata

Douglas Adams (Q42) [edit]

English writer and humorist [edit]

Also known as: Douglas Noël Adams | Douglas Noel Adams | DNA | Bop Ad [edit]

date of birth	11 March 1952 ▸ 1 reference	[edit]

Wikipedia pages linked to this item (64 entries)

Language	Code	Linked page	
العربية	arwiki	دوغلاس آدمز	[edit]
مصرى	arzwiki	دوجلاس ادامز	[edit]
Boarisch	barwiki	Douglas Adams	[edit]
беларуская	be_x_oldwiki	Дуглас Адамз	[edit]

Terms and Languages

- Three kinds of terms: labels, descriptions, aliases
- Used for labelling and searching
  - Label-description pair globally unique (key)
- Term: string in a language (“monolingual text value”)
- Over 350 languages
  - Based on MediaWiki user language (UI language)
  - Different from Wikipedia languages (<300)

Site Links

Wikipedia pages linked to this item (64 entries)

Language	Code	Linked page	
العربية	arwiki	دوغلاس آدمز	[edit]
مصرى	arzwiki	دوجلاس ادامز	[edit]
Boarisch	barwiki	Douglas Adams	[edit]
беларуская	be_x_oldwiki	Дуглас Адамз	[edit]

- Links to other Wikimedia projects
- Used in all Wikipedias to create language links
- Site links as keys:
  - At most one link per project (functional)
  - At most one item per site link (inverse functional)

Statements

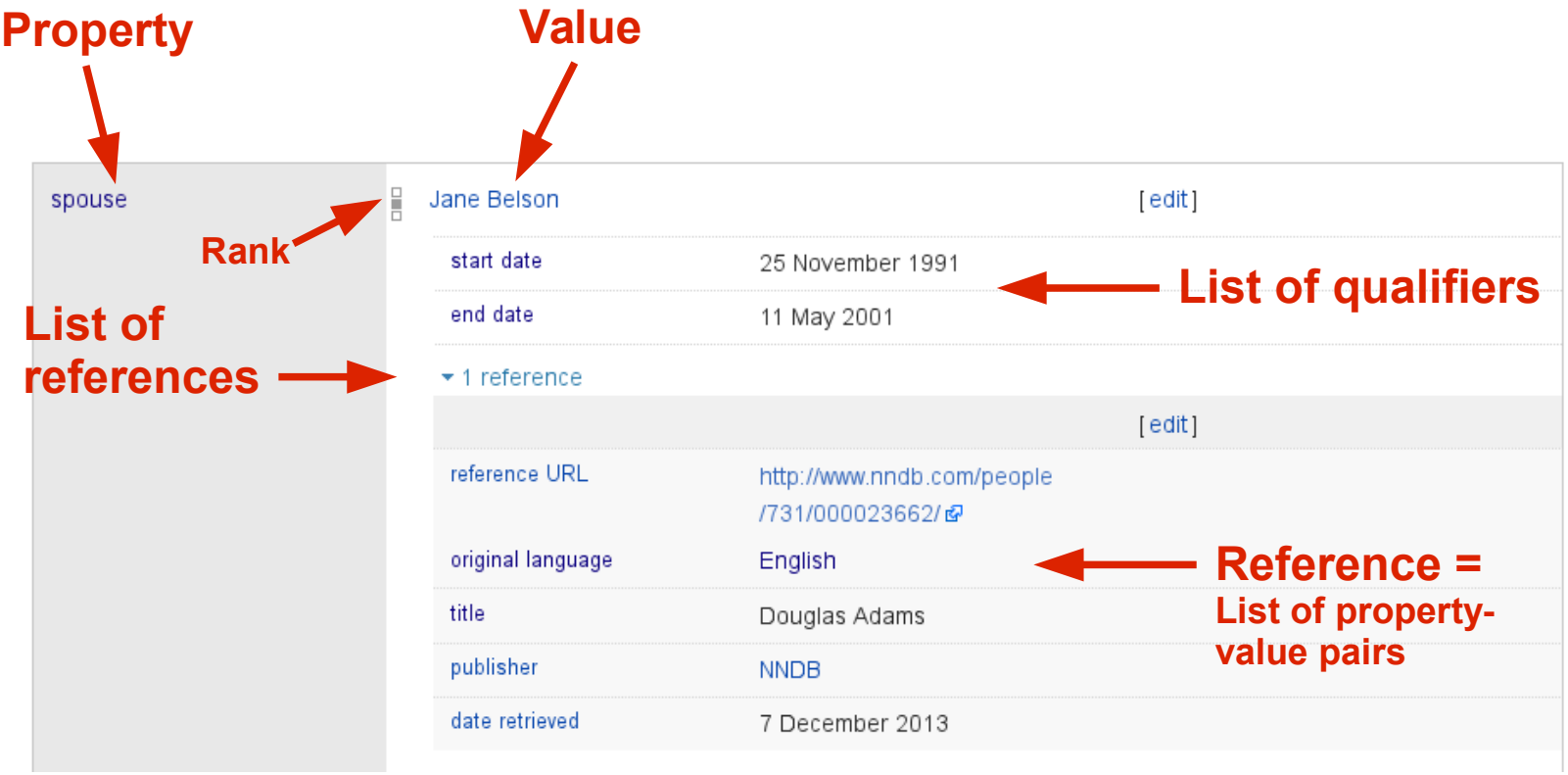
- The richest part of Wikidata's data

Statements

- The richest part of Wikidata's data

Statements

- The richest part of Wikidata's data



Statements

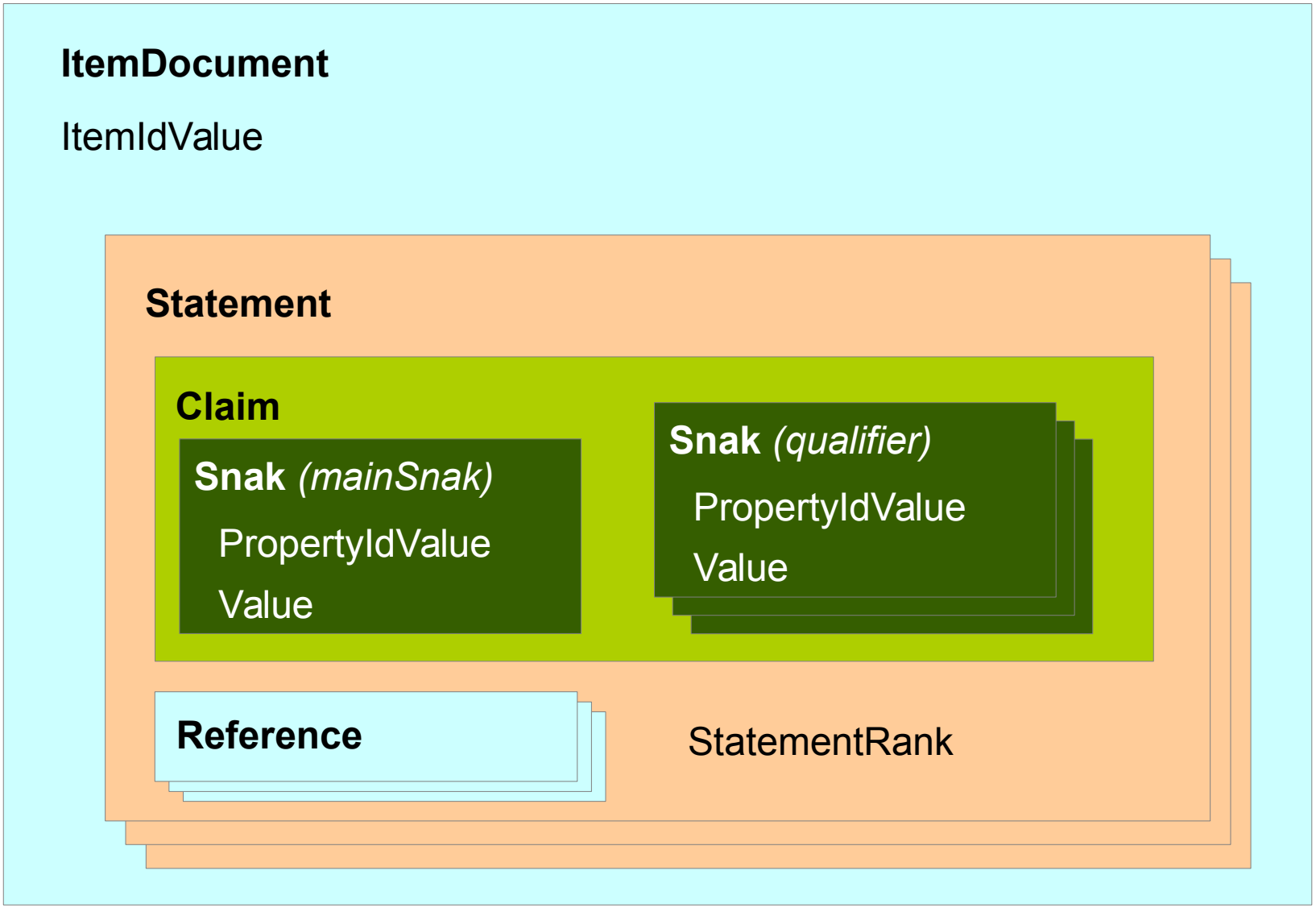
- The richest part of Wikidata's data
- Components of a statement:
  - Main property-value pair
  - List of qualifiers (property-value pairs)
  - List of references (each a list of property-value pairs)
  - Rank (preferred > normal > deprecated)
- Main property-value pair + qualifiers
  = claim (of the statement)

Property-value pairs and “Snaks”

- Properties have datatypes
  - Datatype fixed after creation
- Datatypes: Item, String, URL, CommonsMedia, Time, Globe Coordinates, Quantity
- Two special “values”:
  - *Some:* “there is a value” (that's all we can say)
  - *None:* “there is no value” (basic negative information)

→ Can be used in all places where real values can

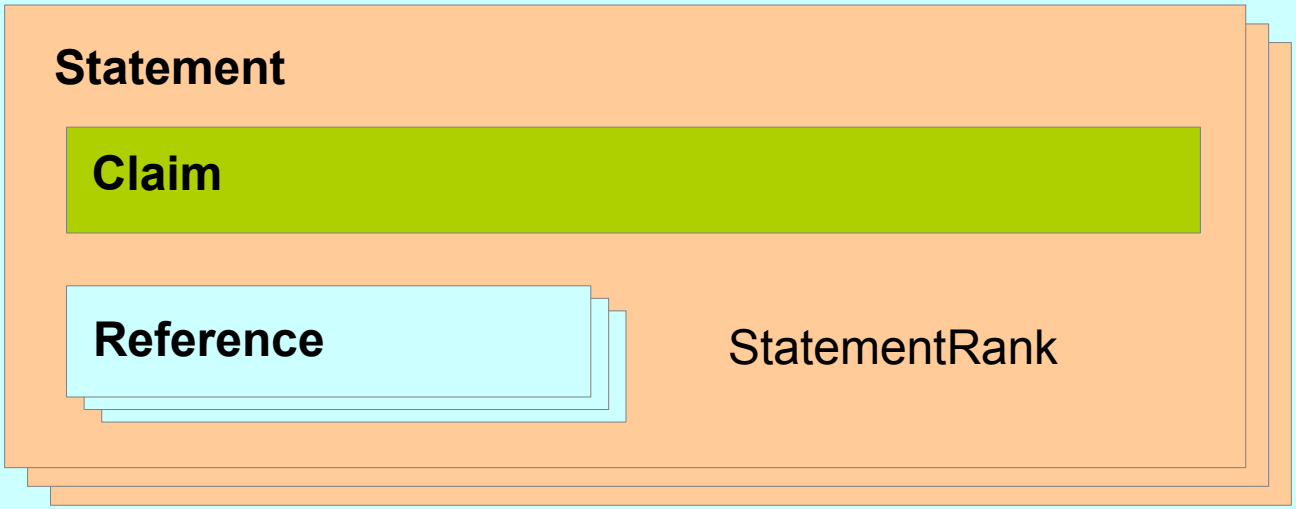
Statements


ItemDocument
ItemIdValue
Statement
Claim
Snak (mainSnak)
PropertyIdValue
Value
Snak (qualifier)
PropertyIdValue
Value
Reference
StatementRank

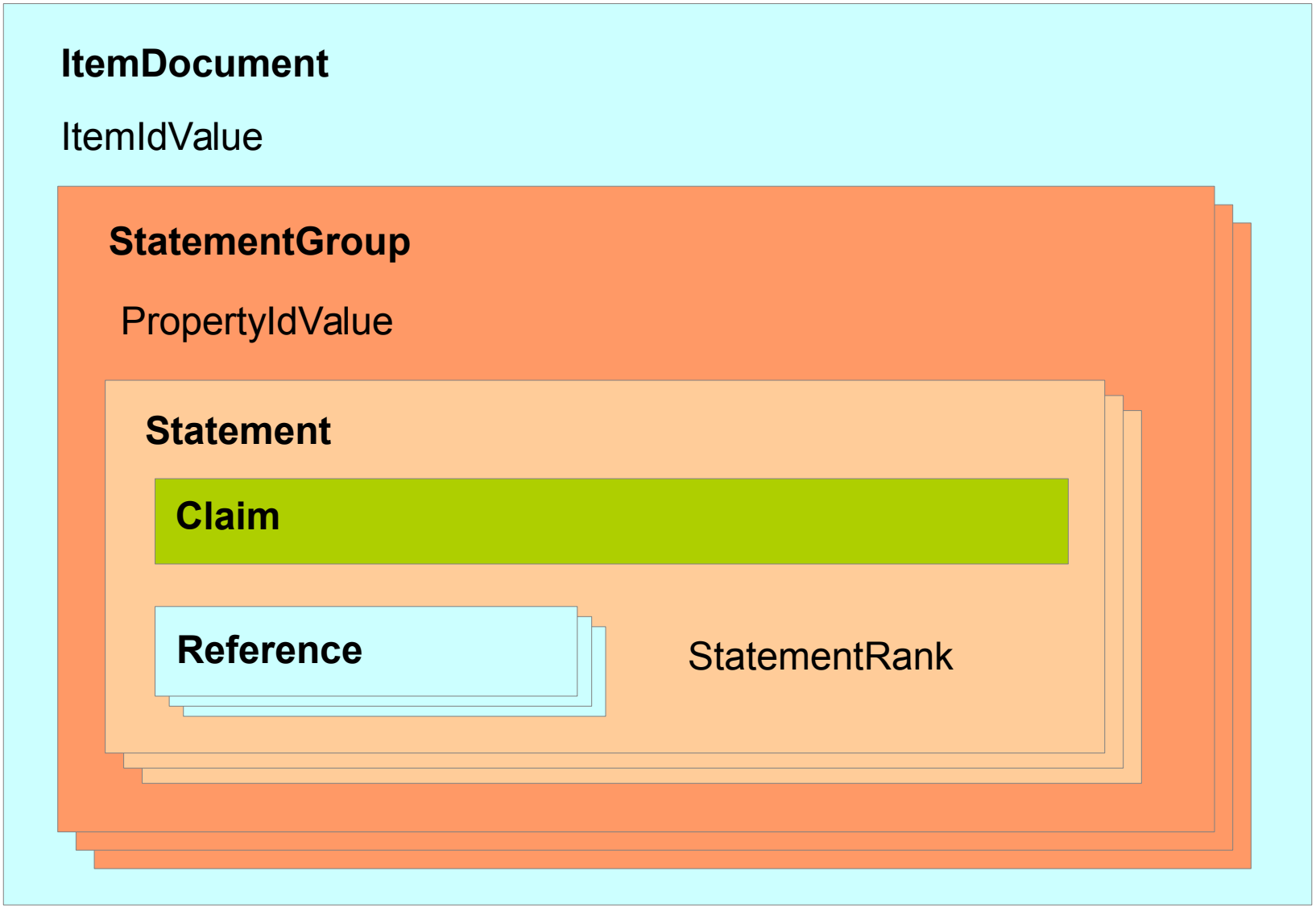
Statements

ItemDocument

ItemIdValue



Statements


ItemDocument
ItemIdValue
StatementGroup
PropertyIdValue
Statement
Claim
Reference
StatementRank

Demo

Scale

Size as of 31st July 2014

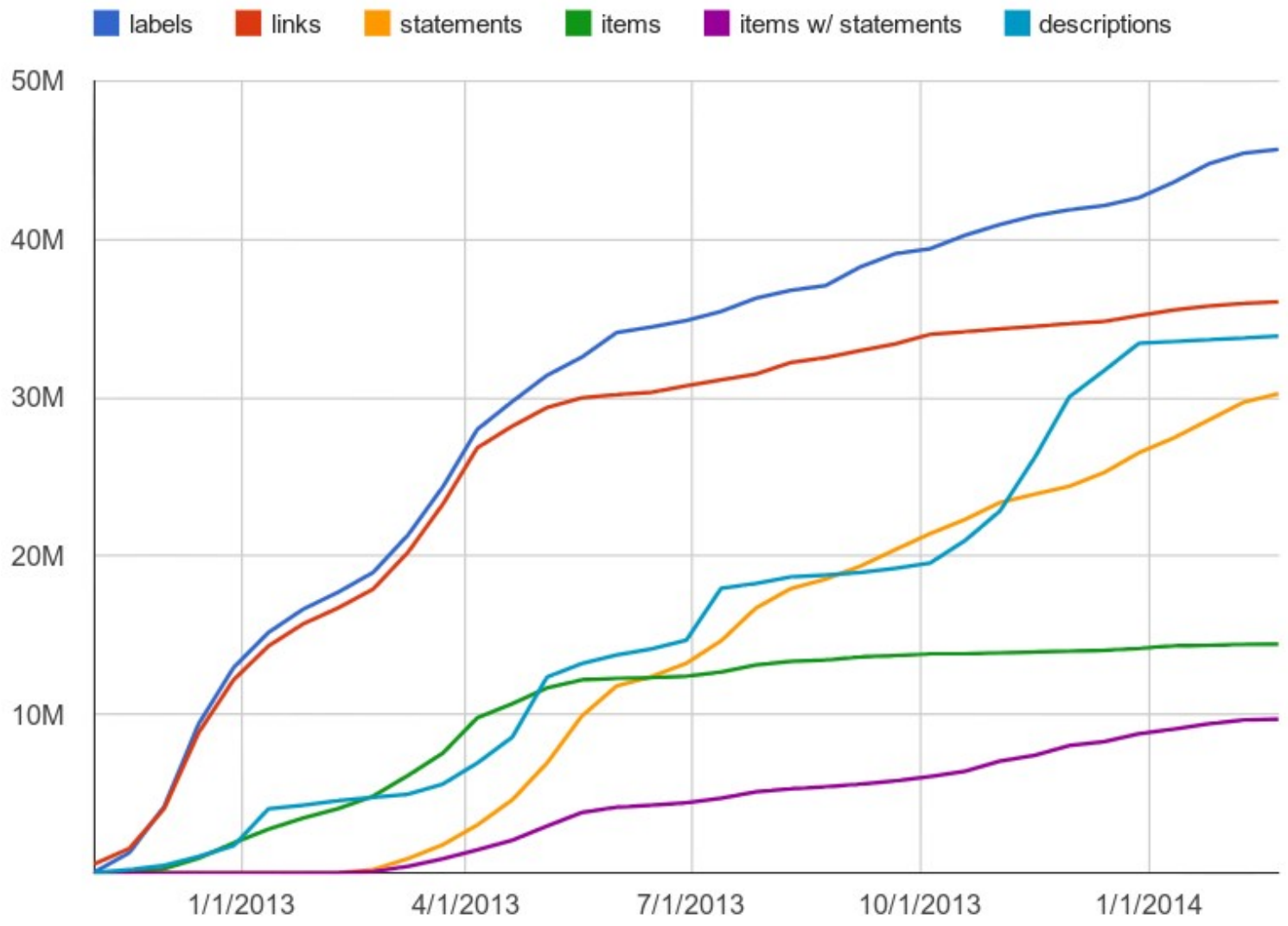
- Items: 15,685,743
- Properties: 1,144
- Statements: 40,087,640

- Labels: 51,808,839
- Aliases: 8,753,016
- Descriptions: 37,101,496

- Site links: 38,852,576

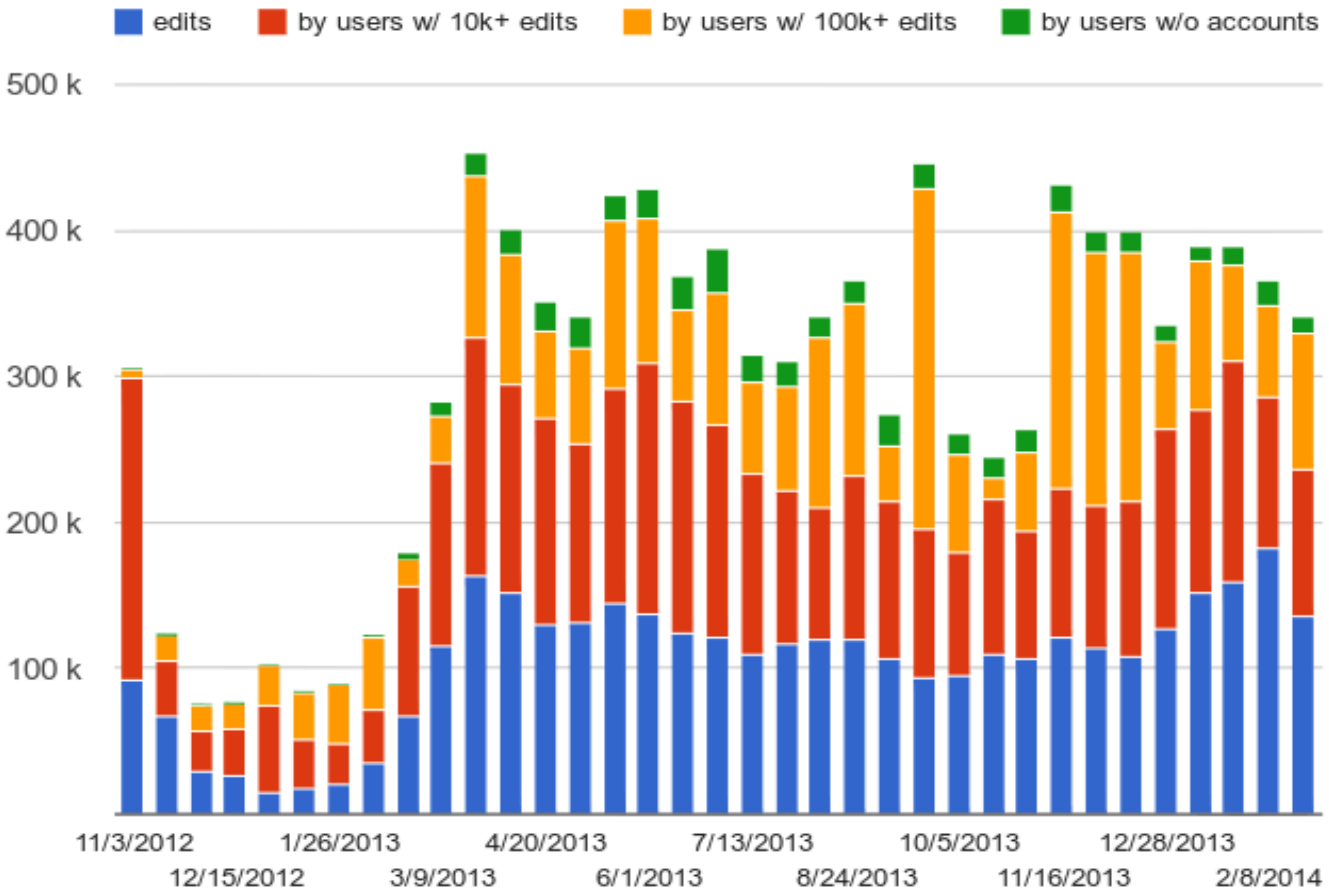
- References (uses): *>23,000,000*
- References (distinct refs): 150,095

Growth (up to Feb 2014)



Activity

(Feb 2014)



- 42k contributors – 5k contributors with 5+ edits in Jan 2014
- Well over 100M edits so far – up to 500k per day

Qualifiers (May 2014)

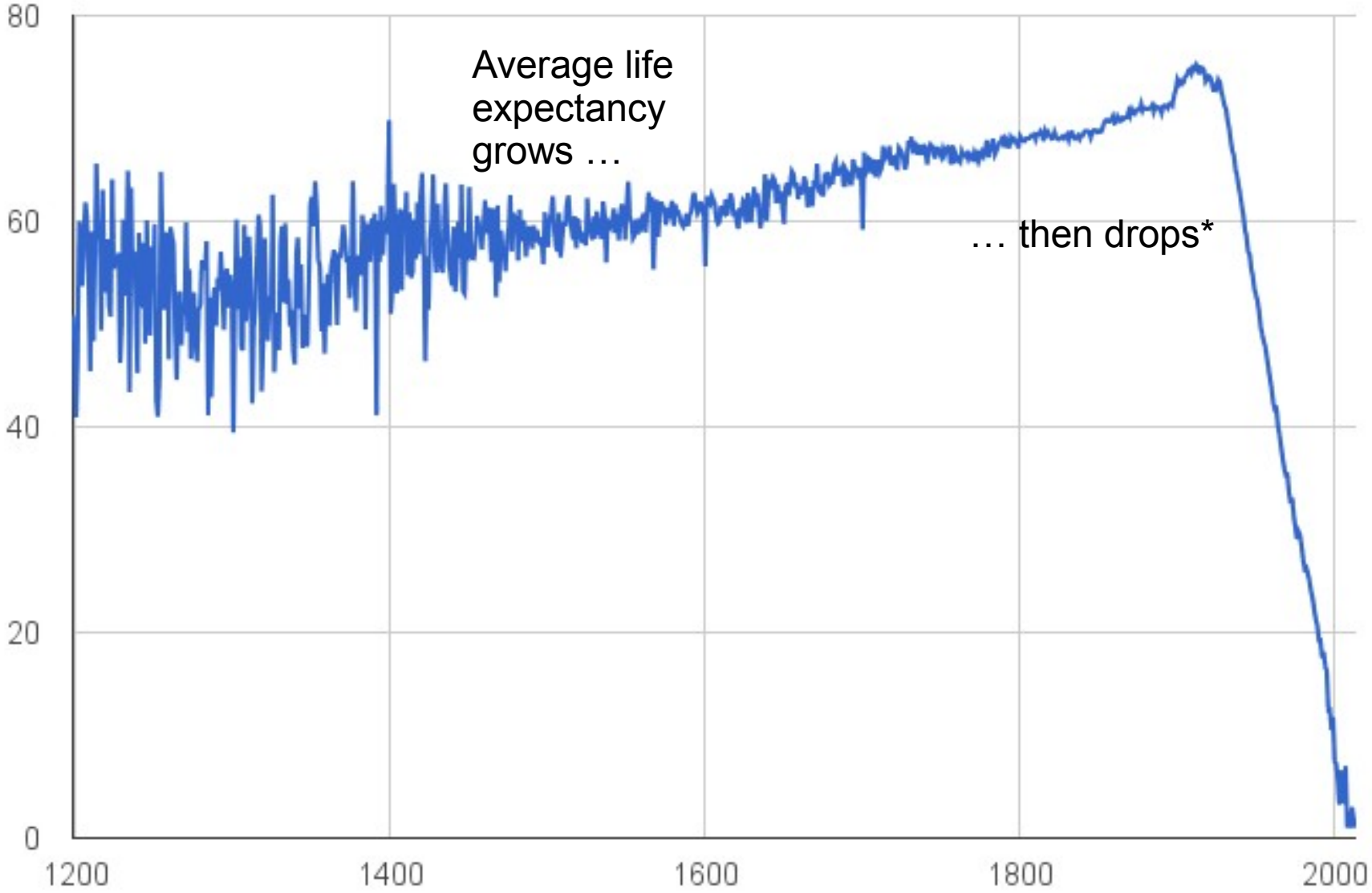
- Relatively rare: *136,187* statements use qualifiers
- Very diverse applications:
  - Temporal context (start date/end date)
  - Other context (e.g., *taxon author*, *asteroid taxonomy*)
  - N-ary relations (e.g., login for *web account on*)
  - Mixture (e.g., *character role* for *cast member*)
- Important: leaving away qualifiers may lead to a wrong claim (rather than just a "weaker" claim).

Classification (May 2014)

- Properties *subclass of* (P279) and *instance of* (P31)
  - P31 is the most used property on Wikidata
- Often (but not always) used without qualifiers
- Interesting class hierarchy:
  - Entities used as classes: 41,868
  - Subclass of: 40,192 (without qualifiers)
  - Instance of: 6,169,821(without qualifiers)
- RDF/OWL file export at:
  http://tools.wmflabs.org/wikidata-exports/rdf/

Conclusions

We will all die!



*) obviously, this must be so if we take the life expectancy of people dead already

Conclusions

- Wikidata is …
  - … fascinating and unpredictable
  - … full of unexplored potential
  - … only at its beginning
- Wikidata Toolkit gives you full access to **all** of Wikidata
  - For creating your own excerpts of the data
  - For aggregation and anaylisis
  - For high-speed, random, offline data access
- WDTK works with any other Wikibase installation

Further reading

- Denny Vrandecic, Markus Krötzsch.
  Wikidata: A Free Collaborative Knowledge Base. CACM 2014. To appear
  → *general first introduction to Wikidata*

- Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, Denny Vrandečić.
  Introducing Wikidata to the Linked Data Web. 2014.
  → *introduction of the Wikidata RDF export and data model*