The New SQID
Improving Wikidata Made Easy
Markus Krötzsch  Maximilian Marx
Knowledge-Based Systems
TU Dresden
WikidataCon 2017
https://etherpad.wikimedia.org/p/WikidataCon-61
All slides CC-BY 3.0, except those requiring CC-BY-SA 3.0
© David Sim, 2009, CC-BY 2.0
The Wikidata Quality Challenge

- Small errors can have a big impact
  ... but are very hard to notice
- Only few direct readers on site
- Significant external usage
  ... but without direct editing options
When “Incomplete” becomes “Wrong”

- Omissions can turn into errors and misinterpretations
- Many SPARQL queries depend on absence of information:
  - Checks for NOT EXIST [around 3% of user queries]
  - Aggregates (counting etc.) [>10% of user queries]
A Tale from Swaziland

Swaziland (Q1050)
country in Africa

office held by head of state

Sobhuza II (Q379576)
King of Swaziland

position held

King of Swaziland

start time

date

end time

date

replaces

replaced by
A Tale from Swaziland

Swaziland (Q1050)
country in Africa

office held by head of state
King of Swaziland

head of state
Mswati III

Sobhuza II (Q379576)
King of Swaziland

position held

King of Swaziland
start time
end time
replaces
replaced by
“Wikidata often doesn’t know what Wikidata knows.”

A Tale from Swaziland

Swaziland (Q1050)
country in Africa

office held by head of state

head of state

Sobhuza II (Q379576)
King of Swaziland

position held

King of Swaziland

start time
end time
replaces
replaced by
Bots to the Rescue!

A big advantage of Wikidata:
- Automatic error search and correction
- Ongoing validation against external sources
- Crowdsourcing keeps human in the loop

However ...
- High barriers for building such solutions
- Sparse coverage of topics
Goal: Let community define what should be done

- Specify “rules” – don’t program
- “What over How”
- Example:
  “If A’s office of head of state is B, and C held the position B, then A’s head of state was C.”

- Provide ways to write and use this
Keep humans involved

Goal: Ensure that results get human review

- Generate proposals for new data
- Allow users to accept or reject
- Record exceptions or suggest ways of fixing problematic data
Spouse (P26) is symmetric:

$(?x.\text{P26} = ?y) \rightarrow (?y.\text{P26} = ?x)$
Spouse (P26) is symmetric:

\((?x. P26 = ?y)@\text{?S} \rightarrow (?y. P26 = ?x)@\text{?S}\)
Spouse (P26) is symmetric:

\((?x. P26 = ?y)@?S \rightarrow (?y. P26 = ?x)@?S\)

Part of (P361) is inverse of has part (P527):

\((?x. P527 = ?y)@?S \rightarrow (?y. P361 = ?x)@?S\)

\((?x. P361 = ?y)@?S \rightarrow (?y. P527 = ?x)@?S\)
Child (P40) is inverse of mother (P25):

\[ (?c.P25 = ?m)@?S \implies (?m.P40 = ?c)@?S \]

\[ (?m.P40 = ?c)@?S \implies (?c.P25 = ?m)@?S \]
SQID Rules by Example

Child (P40) is inverse of mother (P25):

\((?c.P25 = ?m)@?S \rightarrow (?m.P40 = ?c)@?S\)

\((?m.P40 = ?c)@?S \rightarrow (?c.P25 = ?m)@?S\)

Well ... no, the second rule is wrong. Fix:

\((?m.P40 = ?c)@?S\),

\((?m.P21 = Q6581072)@?T \rightarrow (?c.P25 = ?m)@[]\)
Anyone holding (P39) a country's head of state position (P1906) is its head of state (P35):

(?headOfState.P39 = ?headOffice)@?X,
(?country.P1906 = ?headOffice)@?Y
-> (?country.P35 = ?headOfState)@[]
SQID Rules by Example

Anyone holding (P39) a country's head of state position (P1906) is its head of state (P35), at the same start and end time:

(?person.P39 = ?headOffice)@?X,

?X : (P580 = ?start, P582 = ?end),

(?country.P1906 = ?headOffice)@?Y

The Future

Planned software improvements

- Online rule editing
- Better rule management
- Optional value-copying feature for rules
- Performance/load time
- Disapprove inferences (exception handling)
- Advanced constraints
The Future

– Your input here –