Wikidata quality
A data consumers’ perspective

Alessandro Piscopo
Web & Internet Science group
University of Southampton, United Kingdom
About me (Q37798781)

• PhD candidate at the University of Southampton
• My research focuses on data quality and community dynamics in Wikidata
• So far:
  • Evolution of users along their activity on the platform
  • Quality of external references
  • Relation between group composition & quality
Why quality?
Why quality?

• To know how successful Wikidata is and how good the community is doing
• To know where we can improve and what the issues are
• To know what Wikidata can be used for
Data quality is *fitness for purpose*
ORES

• Semi-automatic approach to provide quality scores

Picture from Dario Taraborelli’s slides (https://goo.gl/q8Q6a4)
Property constraints

- Property constraint warnings

Picture from Dario Taraborelli’s slides (https://goo.gl/q8Q6a4)
<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>instrument (P1303)</td>
<td>guitar</td>
</tr>
<tr>
<td></td>
<td>saxophone</td>
</tr>
<tr>
<td>languages spoken, written or signed (P1699)</td>
<td>English</td>
</tr>
<tr>
<td>name in native language (P15559)</td>
<td>David Bowie</td>
</tr>
<tr>
<td>work period (start) (P2031)</td>
<td>1964-01-01T00:00:00Z</td>
</tr>
<tr>
<td>permanent duplicated item (P29692)</td>
<td>[no label specified]</td>
</tr>
<tr>
<td>National Library of Greece ID (P2303)</td>
<td>238333</td>
</tr>
<tr>
<td>country of citizenship (P27)</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>child (P40)</td>
<td>Duncan Jones</td>
</tr>
<tr>
<td></td>
<td>Alexandria Zahra Jones</td>
</tr>
</tbody>
</table>

Cool-WD site: [http://cool-wd.inf.unibz.it/](http://cool-wd.inf.unibz.it/)
Taxonomic hierarchies

• Brasileiro et al., 2016 analyse Wikidata taxonomic hierarchies on the base of the structural relationships that may occur between elements at different levels

• Anti-patterns were found
  • In other words, $P31 \& P279$ are often mixed up

Download the paper here (https://goo.gl/rt5Yka)
Taxonomic hierarchies

• Anti-patterns examples

Images from Brasileiro et al., 2016 (https://goo.gl/rt5Yka)
External references

• Majority of references are relevant and authoritative

• Approach (crowdsourcing + machine learning) could be implemented in Wikidata

Piscopo et al., 2017 (https://goo.gl/23Ds3d)
So many different aspects. What is quality then?
What is quality?

• Values must be **accurate**
• Items (and the knowledge graph) must be **complete**
• Knowledge graph must be **consistent**
• Statements must be **believable**, supported by reliable sources
What is quality?

• Data quality is a **complex, plural** concept
• Wikidata data quality framework (RfC)?
  • This was not enough!

RfC @ https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Data_quality_framework_for_Wikidata
Fitness for purpose?
A data consumers’ perspective

• Data consumers have not been taken into consideration in Wikidata quality studies so far
• Which dimensions are more relevant for data consumers?
• What is the level of quality required?
Aims of this session

• Gather a consumers’ perspective on data quality in Wikidata
• Discuss importance of various quality dimensions
• Discuss requirements of different consumer types
  • Understand which dimensions are more relevant for each user type
• Discuss types of quality issues encountered
Session structure

• Groups formation + quality dimensions choice (8 min.)
• Looking at dimensions in detail (8 min. x 3)
• Plenary discussion (10 min.)
• Wrap-up (3 min.)
Groups

- 4-5 people per group
- People with **similar data needs** (e.g. app developer, research, GLAM…)
- Please add notes regarding your group here: [https://etherpad.wikimedia.org/p/WikidataCon-95](https://etherpad.wikimedia.org/p/WikidataCon-95)
Quality dimensions

- **Accuracy**
  Is data **free of error** & correctly describing facts in the real-world?

- **Objectivity**
  Is data **free of bias** & impartial?

- **Reputation**
  Is data **trustworthy** & enough referenced?

- **Consistency**
  Does data form a **consistent** knowledge representation?

- **Timeliness**
  Is data sufficiently **up to date**?

- **Completeness**
  Does data have **sufficient breadth**, depth, & scope?

- **Interpretability**
  Can data can be interpreted by machines & humans **without ambiguity**?

- **Interlinking**
  Is data **sufficiently interlinked to other resources**, either within or without Wikimedia?
Let’s look at quality now

• Why is this dimension important for you?
• What do you expect from Wikidata to satisfy your needs wrt. this dimension?
• Examples of issues related to this dimension

Notes here: https://etherpad.wikimedia.org/p/WikidataCon-95
Our findings

• What are our findings?
• What do different data consumers need from Wikidata?
• How can we address the most burning issues?

Notes here: https://etherpad.wikimedia.org/p/WikidataCon-95
Thanks!

Contact me at:
A.Piscopo <at> soton.ac.uk