DATA QUALITY MODELLING AND MEASUREMENTS

A case study with German open government data

Ertugrul Bircan Copur
Goals

Open Data Users → Evaluate (survey) → Open Data Datasets → Evaluate → Set of Data Quality Metrics → Compare → Results
Survey and Participation

- We asked application developers about:
  - Their experience in programming
  - Their experience with open data
  - The main and additional minor problems with the dataset
  - The main and additional positive aspects of the dataset

- Low participation – only four participants

- Five datasets that we can evaluate (six in total, one unusable for us)
Evaluation by Metrics

- A set of characteristics based on the SQuaRE-aligned Portal Data Quality Model (SPDQM).
  - Completeness, Traceability, Compliance, Accuracy, Understandability, Currentness, Expiration

- A total of fifteen metrics were used to evaluate the quality of the datasets.
## Results and Comparison

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Completeness</th>
<th>Traceability</th>
<th>Compliance</th>
<th>Accuracy</th>
<th>Understandability</th>
<th>Currentness</th>
<th>Expiration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Survey</td>
<td>Metrics</td>
<td>Survey</td>
<td>Metrics</td>
<td>Survey</td>
<td>Metrics</td>
<td></td>
</tr>
<tr>
<td>1 - Used Glass</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>--</td>
<td>--</td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Containers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>2 - Popular First</td>
<td>-</td>
<td>++</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>++</td>
<td>n/a</td>
</tr>
<tr>
<td>Names</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>3 - Christmas</td>
<td>-</td>
<td>++</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>--</td>
<td>n/a</td>
</tr>
<tr>
<td>Markets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>4 - Bus Stops</td>
<td>+</td>
<td>-</td>
<td>0</td>
<td>++</td>
<td>+</td>
<td>-</td>
<td>n/a</td>
</tr>
<tr>
<td>5 - Memorial Stones</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>++</td>
<td>0</td>
<td>+</td>
<td>n/a</td>
</tr>
</tbody>
</table>

++ : Extremely positive  
0 : Insignificant  
-- : Extremely Negative  
+ : Positive  
n/a : Not computable  
- : Negative
Analysis

- Problems mainly with the Completeness characteristic because of its perception by users.
- Our definition of completeness is how complete each record is within itself.
- Users interpret completeness as the dataset’s ability to cover all of the real-life data that the dataset assumes to include.
Conclusion and Future Works

- The first aim is to find more survey participants. (Please join us!)
- Interviews with willing survey participants
- Possible modification or replacement of metrics
- Possible changes to the data model
Thank you for listening.

Please consider participating in our survey or spreading the work to suitable users!

http://tinyurl.com/OpenDataQuality