Data Quality

Data Management and Visualization

Licensing Note

This work is licensed under the Creative Commons Attribution–NonCommercial-NoDerivatives 4.0 International License.
To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.
You are free: to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Non-commercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.
Software Qualities

Adapted from ISO/IEC 25020

Target entities

Information System

ICT Product

Data
Software Component

Context of use
Users
Goals
User Environ.
System Context
Software Product Quality

  - Being retired

- ISO/IEC 250xx – SQuaRE
  - Software product Quality Requirements and Evaluation
  - Family of standards
    - in development
## ISO SQuaRE – Standard Family

<table>
<thead>
<tr>
<th>2503x</th>
<th>2504x</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality Requirements</td>
<td>Quality Evaluation</td>
</tr>
<tr>
<td>2500x</td>
<td>Quality Management</td>
</tr>
<tr>
<td>2501x</td>
<td>Quality Model</td>
</tr>
<tr>
<td>2502x</td>
<td>Quality Measurement</td>
</tr>
</tbody>
</table>

### Relationships among standards

**ISO/IEC 25010**
System and Software Product Quality

**ISO/IEC 25012**
Data Quality

- composed of
  - Quality characteristics

Defined by **ISO/IEC 25021**
Measurement function

**ISO/IEC 25022, 25023, 25024**
Quality Measure

- composed of
  - Quality Measure Elements
  - QME
  - Measurement method

**Target Entity**
Property to quantify

Source: ISO/IEC 25024
Quality conceptual model

Model structure

- Characteristic
  - Main aspects, e.g., usability
- Sub-Characteristic
  - Specific aspects, e.g. accessibility
- Measure
  - Measurement function to evaluate a specific (sub)-characteristic
- Measure element
  - Fundamental
Data Quality

Data Quality Model
Quality characteristics

- Accuracy
- Completeness
- Consistency
- Currency
- Credibility

- Accessibility
  - Compliance
  - Confidentiality
  - Efficiency
- Understandability
  - Precision
  - Traceability

- Availability
- Portability
- Recoverability

Inherent: facts
System dependent: artefacts
Accuracy

- Correspondence between data and reality
  - Syntactic
    - It belongs to a set of validated information
  - Semantic
    - The meaning (the content) corresponds to the reality

Open or Closed World?

- **Closed World (CWA):**
  - The knowledge represented in the data (and its schema) is complete
  - E.g., if a code appears in the list of valid codes it is correct, otherwise it is wrong

- **Open World (OWA):**
  - The knowledge represented in the data is (knowingly) incomplete
  - E.g., if a code appears in the list of valid codes it is correct, otherwise it is not possible to tell for sure
CWA – Accuracy : Genomics

- Human genes are known and coded, each has a predefined symbol
- Any code not included in those predefined represents a syntactic accuracy error
- E.g. code ‘SEPT2’ (Septin–2) when imported into is automatically turned into ‘February 2’

OWA – Accuracy

How to decide what is accurate?

- Rules that define what is syntactically correct
  - E.g. regular expressions
- Constraints to define what values are semantically acceptable
  - E.g. validity interval
Where do rules come from?

- Standard
- Domain knowledge
- Similar data
- Past data

OWA: Email per RFC–5322

\A(?:[a-z0-9!#$%&'\*+/=?^`\{\}|~]+)+(?:\.[a-z0-9!#$%&'\*+/=?^`\{\}|~]+)+\A

\A(?:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f])

! | \"(?:[\x01-\x09\x0b\x0c\x0e-\x7f])*"")

\A(?:[a-z0-9])(?:[a-z0-9-]*[a-z0-9])?\A

\A(?:(?:[\x01-\x09]\x0b\x0c\x0e-\x7f]*\A

\A(?:(?:25[0-5]|2[0-4][0-9]|0[0-9][0-9])?\A

\A(?:25[0-5]|2[0-4][0-9]|0[0-9][0-9])?\A

\A(?:[a-z0-9-]*[a-z0-9-]:

\A(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f])

\A! | \"(?:[\x01-\x09\x0b\x0c\x0e-\x7f])+

\\)\z
Non printable characters are usually a problem for email clients.

The notation with [ ] is obsolete and often not implemented.

Completeness

- **Computer**: presence of all necessary values
  - Both to entity occurrences and to attributes of a single occurrence
  - Note: not all missing values constitute a completeness issue
- **User**: how much the available data is capable of satisfying the needs
Completeness

What about 1930s, 1950s, 1970s, 1990s?

A possible hypothesis, another one considered later

Source: http://www.nytimes.com/2014/09/14/magazine/who-made-that-windshield-wiper.html?_r=0

Consistency

- Absence of contradictions in the data
  - Referential integrity
    - Often guaranteed in RDBMS
  - Duplication
    - Increase the risk of inconsistency on update
  - Semantic
    - E.g. birth date must be before death date
Consistency in graph data

- Values in a series of data encoded with visual attributes must be comparable
  - Consistent aggregation level
  - Consistent measurement method
  - Consistent target entities

Aggregation level

Count on of events on periods of different length are not comparable

A possible hypothesis, another one considered earlier

Source: http://www.nytimes.com/2014/09/14/magazine/who-made-that-windshield-wiper.html?_r=0
## Aggregation level

<table>
<thead>
<tr>
<th>Period</th>
<th>Duration [years]</th>
<th>Patents</th>
<th>Pat. per year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1920s</td>
<td>20</td>
<td>430</td>
<td>21.5</td>
</tr>
<tr>
<td>1940s</td>
<td>20</td>
<td>260</td>
<td>13.0</td>
</tr>
<tr>
<td>1960s</td>
<td>20</td>
<td>650</td>
<td>32.5</td>
</tr>
<tr>
<td>1980s</td>
<td>20</td>
<td>410</td>
<td>20.5</td>
</tr>
<tr>
<td>2000s</td>
<td>10</td>
<td>660</td>
<td>66.0</td>
</tr>
<tr>
<td>2010 to present</td>
<td>4</td>
<td>390</td>
<td>97.5</td>
</tr>
</tbody>
</table>

When comparing values corresponding to entities or categories with different size, normalized values (i.e. densities) are comparable, absolute values are not!
## Aggregation level

<table>
<thead>
<tr>
<th>Range</th>
<th>Size</th>
<th>Count</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>31–35</td>
<td>5</td>
<td>235</td>
<td>47.0</td>
</tr>
<tr>
<td>36–4</td>
<td>5</td>
<td>3109</td>
<td>621.8</td>
</tr>
<tr>
<td>41–50</td>
<td>10</td>
<td>16455</td>
<td>1645.5</td>
</tr>
<tr>
<td>51–60</td>
<td>10</td>
<td>18093</td>
<td>1809.3</td>
</tr>
<tr>
<td>Over 60</td>
<td>10</td>
<td>10989</td>
<td>1098.9</td>
</tr>
</tbody>
</table>

**Ratios:** 5.3 2.6

\[ \text{Lie factor} = 2 \]

## Consistent method

- A series of values that are not measured using the same method might not be directly comparable
  - estimate vs. actual, projection vs. final
  - periodic samples collected at different possibly non equivalent times
    - e.g. different period of year, week, day
Consistent target entities

Bruce gain
Estimated heights and weights of on-screen Batmen

<table>
<thead>
<tr>
<th></th>
<th>Height (cm)</th>
<th>Weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comic book</td>
<td>188/95</td>
<td></td>
</tr>
<tr>
<td>Adam West</td>
<td>188/111</td>
<td></td>
</tr>
<tr>
<td>Michael Keaton</td>
<td>178/72</td>
<td></td>
</tr>
<tr>
<td>Val Kilmer</td>
<td>183/93</td>
<td></td>
</tr>
<tr>
<td>George Clooney</td>
<td>180/78</td>
<td></td>
</tr>
<tr>
<td>Christian Bale</td>
<td>183/82</td>
<td></td>
</tr>
<tr>
<td>Lego Batman</td>
<td>180/14</td>
<td></td>
</tr>
<tr>
<td>Ben Affleck</td>
<td>183/98</td>
<td></td>
</tr>
</tbody>
</table>

Sources: Moviepilot; IMDb

*From "The Lego Movie", not to scale

Consistent target

Poll dates

Intenzioni di voto 2018
(Data in %)

Different political parties

Undecided/NA
Consistent target

- Proportions computed on different reference wholes

\[
Undecided = \frac{n_{undec} + n_{NA}}{N_{sample}}
\]

\[
P_i = \frac{n_{pi}}{N_{sample} - n_{undec} - n_{NA}}
\]

Currency

- Currency is the extent to which data is up-to-date
  - With reference to the reality and
  - With reference to the task at hand

- Lack of information to establish currency is an Understandability issue
Credibility

- The extent to which data are regarded as true and credible by users

- What is the source of the data showed in the graph?

Understandability

- The extent to which data can be read and interpreted by users

- How is data measured? Is there a track of how values are collected, measured or estimated?
  - If multiple methods are used that might represent an inconsistency issue.
Understandability

billions, % change and % of total digital display ad spending*

2014 2015 2016 2017
118.6% 49.0% 59.0% 72.0%
$10.32 $15.83 $22.19 $27.47
53.3% 39.7% 24.3%

Programmatic digital display ad spending
% change % of total digital display ad spending*

Note: digital display ads transacted via an API, including everything from
publisher-erected APIs to more standardized RTB technology; includes
native ads and ads on social networks like Facebook and Twitter; includes
advertising that appears on desktop/laptop computers, mobile phones,
tables and other internet-connected devices. *includes banners, rich
media, sponsorship, interstitial and other.
Source: eMarketer, April 2016

Data from 2016 including values for 2017.
Undeclared mix of projections and final data.

Accessibility

- The capability of data to be accessed, particularly by people who need
  supporting technology or special configuration because of some disability

Original
Color-blind simulation
Precision

- The capability to provide the degree of information needed in a stated context of use
  - Enough information to allow discriminate
  - Not too much to overload reader
    - Related to "Utility"

**Debito pubblico (% PIL)

(*) previsioni Commissione UE**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>131,8</td>
<td>131,6</td>
<td>131,4</td>
<td>131,4</td>
<td>132,2</td>
<td>133,7</td>
<td>135,2</td>
</tr>
</tbody>
</table>

**Governo Renzi e Gentiloni**

**Governo Conte**
Precision

References

- ISO/IEC 25010 – System and software quality models
- ISO/IEC 23012 – Data Quality model
- ISO/IEC 25024 – Measurement of data quality