Institutionalising Open Data Quality: Processes, Standards, Tools

ODQ2015 - Open Data Quality: from Theory to Practice

30 March 2015 - Technische Universität München | Institut für Informatik
Boltzmannstr. 3 , 85748 Garching bei München - Room 00.08.038

Johann Höchtl, Danube University Krems, Austria

1. Assess data quality

What is Data Quality?

What are the data quality measures for open data?

http://opendata.stackexchange.com/questions/613/what-are-the-data-quality-measures-for-open-data

5

How does a consumer know they are getting good data? Are there standard frameworks for grading the quality of an open data set? Should there be metrics published around accuracy, completeness, timeliness or validity of the data? Should there be a minimum set of controls on the part of the publisher?

releasing-data data.gov

share improve this question

2

3

I think the question, as phrased, is impossible to answer well, but I will try.

Q: "How does a consumer know they are getting good data?"

A: Let me answer with more questions. How does a consumer know they are getting a good search result from Google? How do they know when the news is of high quality? It depends. As consumers get more interested and informed about something, they do better. The most savvy and informed consumers will compare a data set against a known source. Others have to rely on some degree of trust.

Q: "Are there standard frameworks for grading the quality of an open data set?"

A: In practice, there are defacto standards for metadata. For example, data.gov uses Dublin Core along with additional attributes. CKAN has many of the same attributes.

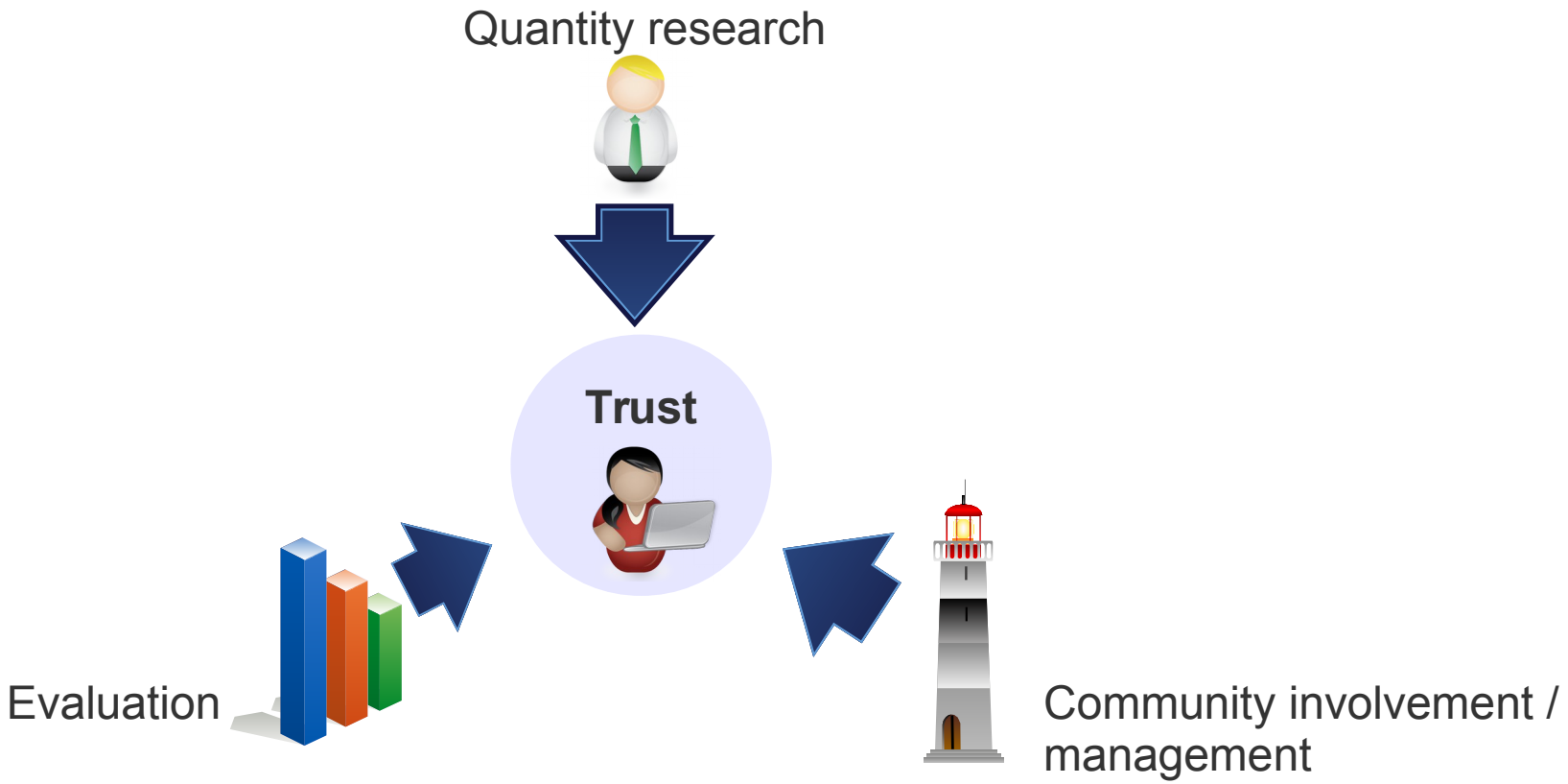
Also, for each type of data (or subfield) there are often industry standards or at least conventions. Good luck enumerating those!

A post from the Sunlight Foundation, Government Data Sets - Managing Expectations is a high-level gloss; it breaks down "dataset quality" into provenance, data quality, responsibility, maintenance, and documentation.

The above article is somewhat naive; the quality of a data set is not an independent thing. As Wikipedia - Data Quality points out, the quality of a data set depends on the question asked of it. There is no "one" measure of data quality. Rather, there is a subjective 'appropriateness' for each question you might ask of a data set. You can't ignore the subjective nature of data quality.

A: Measures towards Trust

1. Establish quantitative measures
2. Provide statistics
3. Show-case lighthouse projects and business use



2. Solve current problems

Mundane problems - Encodings & Formats

- Inconsistent encoding
  - Microsoft Excel caused data problems even when used […] UTF-8
  - Data contaminated with **characters incomprehensible to UTF-8;** ill-formatted following UTF-8; flipped erratically between other character formats; used US ASCII standard, ISO-8859 standard and a similar non-ISO encoding
- Inconsistent dates, file names, data fields
  - Data were regularly **formatted with commas;** changed its filename convention; **omitted or added data fields;** changed the way it formatted dates

http://www.computerweekly.com/news/2240227682/Poor-data-quality-hindering-government-open-data-transparency-programme

Mundane problems – Broken Links

Dead links on data catalogs

As I started looking at data on CKAN sites, I noticed that a lot of the datasets were links to files on other websites and that a lot of these links were dead. Then I started wondering which links were dead and how this happens.

http://thomaslevine.com/!/data-catalog-dead-links/

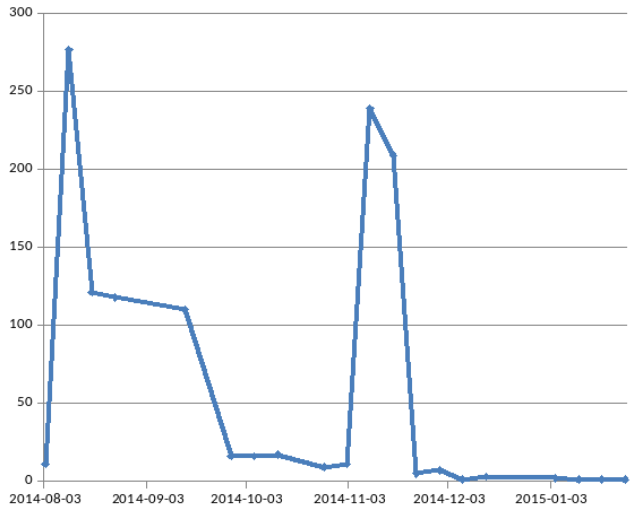
Broken links and hardly any new data on Dutch government open data portal

25/06/2014

The government of the Netherlands still lags behind in opening data to the public. The number of accessible datasets via the Dutch government open data portal is even reduced compared to a year ago.

http://openstate.eu/2014/06/nederlands-nauwelijks-nieuwe-Datasets-op-data-overheid-nl/

City of Vienna – Resource check



B: Measures towards Open Data Quality: Process Domain

- Data publication must be made an integral, well- defined and standardized part of daily procedures and routines
  - A. Zuiderwijk, M. Janssen, S. Choenni, and R. Meijer, “Design principles for improving the process of publishing open data,” Transforming Government: People, Process and Policy, vol. 8, no. 2, pp. 185–204, 2014.
- Process model in which open data serves as a facilitator towards open government
  - G. Lee and Y. H. Kwak, “An Open Government Implementation Model: Moving to Increased Public Engagement,” IBM Center for The Business of Government, Jan. 2011 [Online]. Available: http://www.businessofgovernment.org/sites/default/files/An%20Open%20Government%20Implementation%20Model.pdf
- Establish a Chief Data Officer
  - Y. Lee, “A cubic framework for the chief data officer : succeeding in a world of big data,” 2014.

B: Measures towards Open Data Quality: Standards Domain

- Data on the Web
  - Data on the Web Best Practices Working Group Charter http://www.w3.org/2013/05/odbp-charter.html
  - Encodings: UTF8
- File formats
  - CSV: CSV on the Web Working Group http://www.w3.org/2013/csvw/wiki/Main_Page
  - Frictionless open Data: CSV Files (OKFN guidance document) http://data.okfn.org/doc/csv
- Data entities
  - Geo-Data: Spatial Data on the Web Working Group Charter http://www.w3.org/2015/spatial/charter
  - Date & Time: ISO 8601 http://www.w3.org/TR/NOTE-datetime

B: Measures towards Open Data Quality: Tools Domain

- Identify Problems

seanh commented on 14 Oct 2014 — Owner

This is what became of the link checker in the end:

https://github.com/ckan/deadoralive
https://github.com/ckan/ckanext-deadoralive
http://seanh.cc/posts/ckanext-deadoralive/
http://seanh.cc/posts/background-tasks-as-simple-web-services/

It's "finished for now" (fulfils our particular client's needs, bugs notwithstanding) but still plenty of potential enhancements in the github issues.

https://github.com/ckan/ideas-and-roadmap/issues/65

- Curate File Formats & Encodings

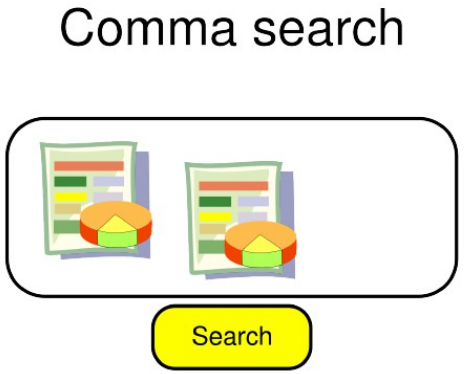


Figure 3: The search engine for spreadsheets takes spreadsheets as input and emits spreadsheets as output

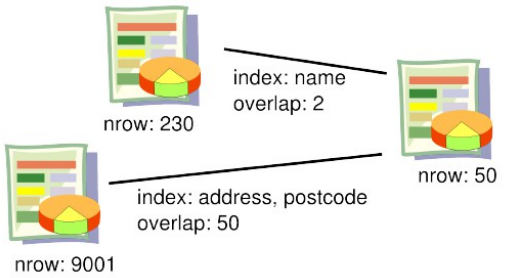
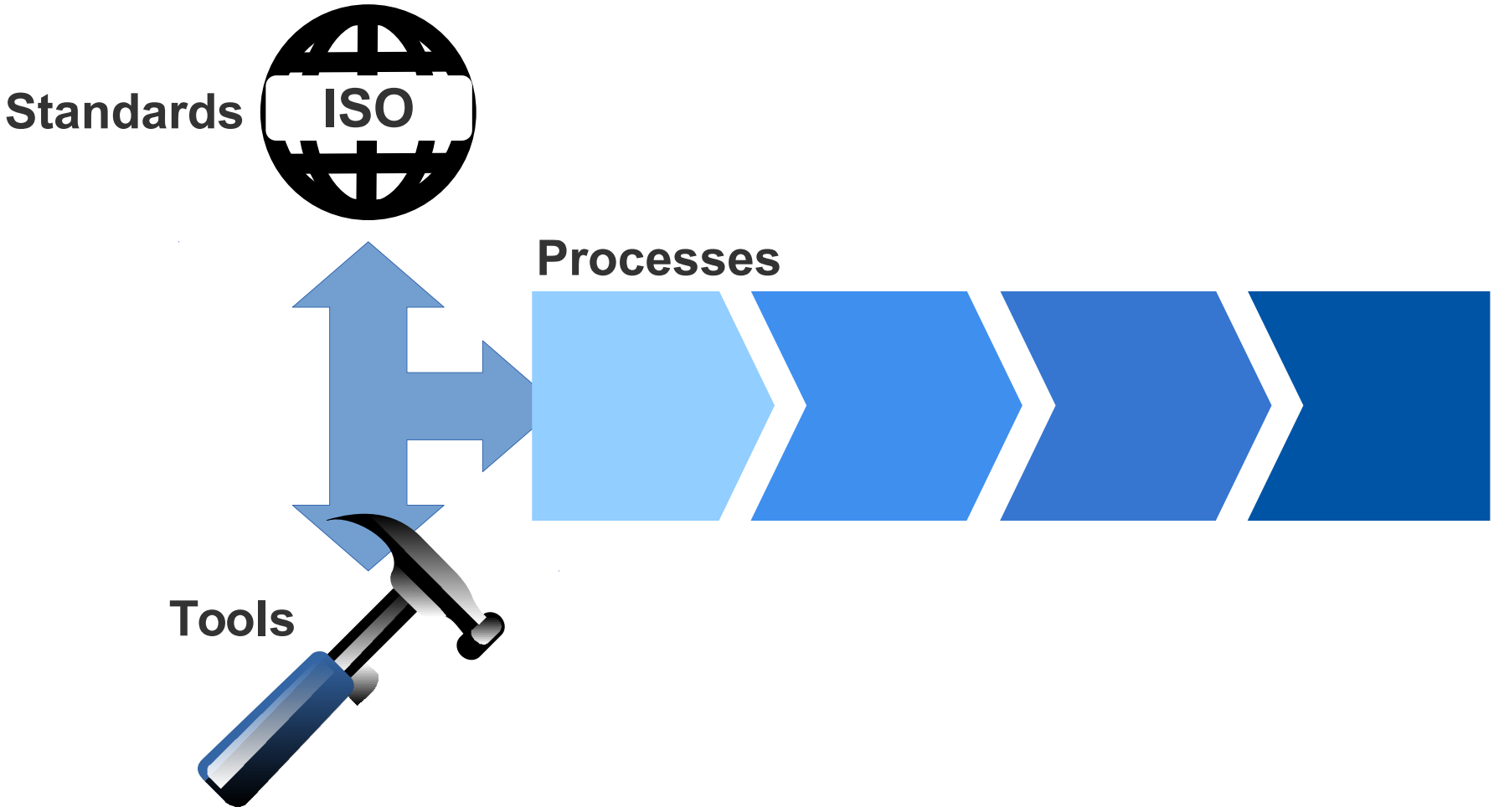


Figure 4: Commasearch infers some schema information about each spreadsheet and looks for other spreadsheets with similar schemas.

T. Levine, “How can we figure out what is inside thousands of spreadsheets?,” CEUR workshop proceedings, vol. 1209, pp. 34–38, Jul. 2014. http://ceur-ws.org/Vol-1209/paper_12.pdf

Measures Towards Open Data Quality



Open Data Quality at the European Open Data Portal

- A.6. Mechanisms for probing broken links

  The portal infrastructure will include a mechanism for systematically probing for broken links. […] The contractor will define and implement a communication protocol to alert the owner of the resource.

- A.8. Mechanism allowing data linking

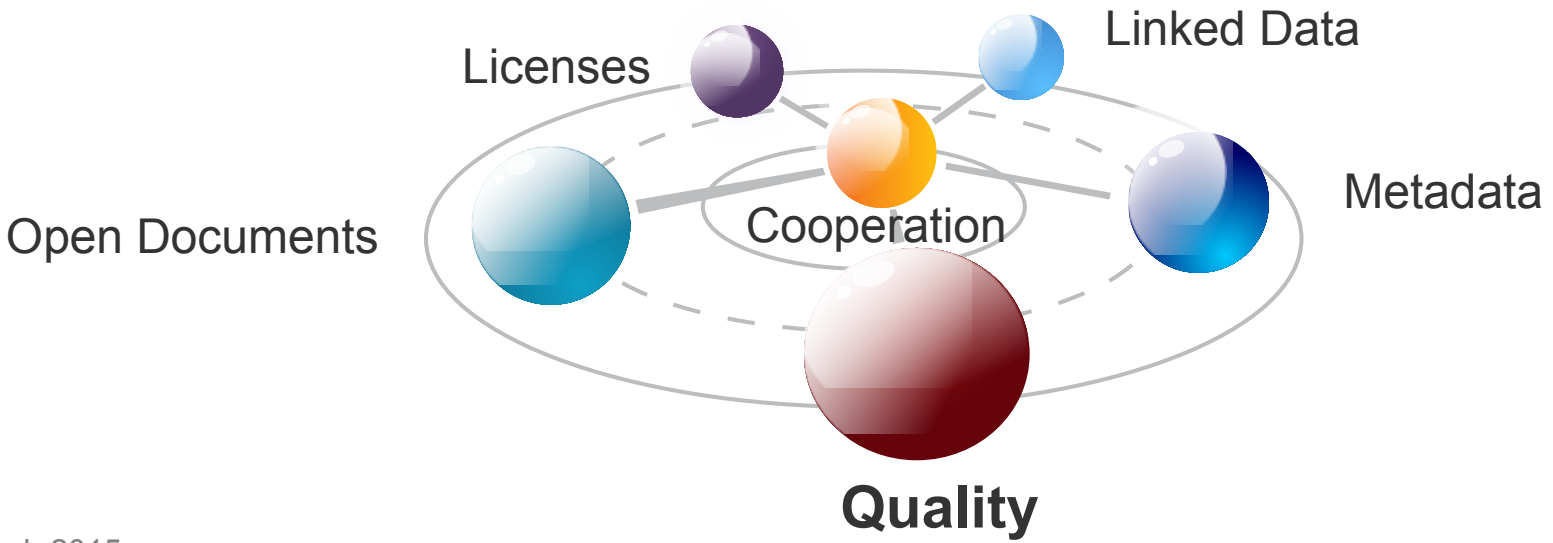
  When RDF, * record a link between datasets that use the same URIs; * propose a mapping between URIs that are likely to denote the same entities

- B.6. User feedback mechanism

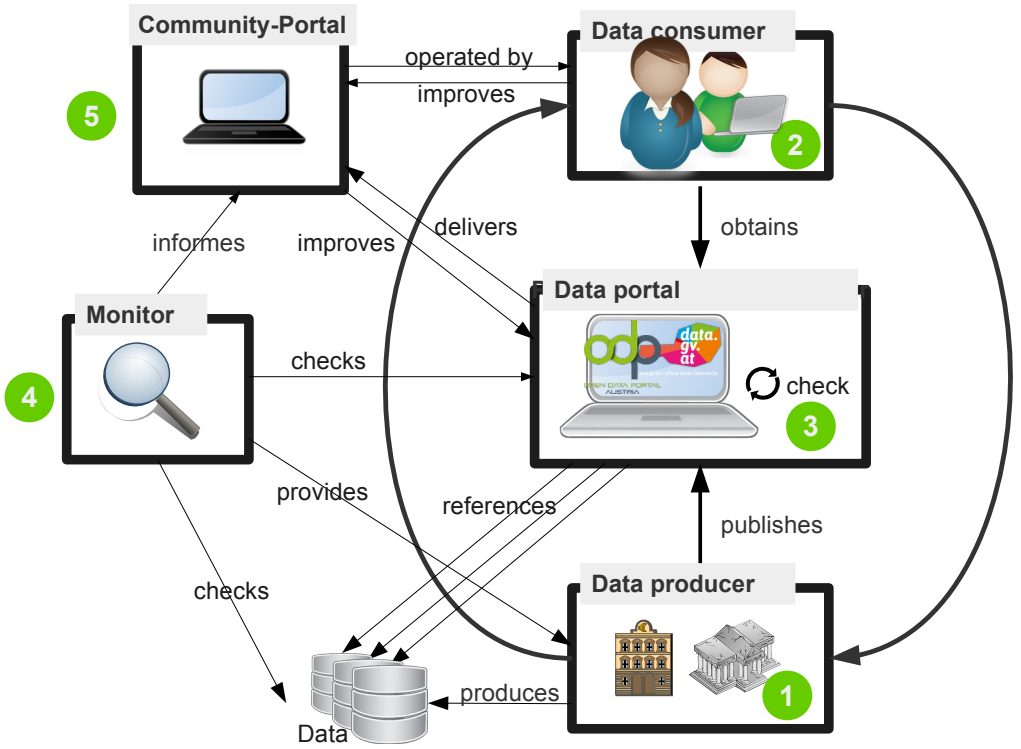
  Allowing visitors […] suggestions for improvements in the data quality

Open Data Quality in Austria

- Cooperation OGD Austria represents administration open data portal operators
  - Defines standards and procedures
  - Aligned with International, European and D-A-CH efforts
- Institutionalising effort by **Sub-Working Group** of Cooperation OGD Austria



Open Data Quality Integration Framework



1. Quality processes and procedure models to assess and publish data
2. Contributions of the Open Data users
3. Quality checks when entering (meta-)data descriptions at the data portal
4. Monitoring of data quality over time
5. Community-driven data portal with user-generated content, e.g. enrich metadata, alternative data formats, etc.

Donau-Universität Krems.
Die Universität für Weiterbildung.

Share-PSI 2.0

eDEM 15
CONFERENCE FOR
eDEMOCRACY & OPEN GOVERNMENT

20.05. - 22.05.2015 Krems, Austria

Johann Höchtl

Center for E-Governance

Johann.hoechtl@donau-uni.ac.at

@myprivate42

at.linkedin.com/in/johannhoechtl