Best Practice: Enable quality assessment of open data

25 July 2016

This version
http://www.w3.org/2013/share-psi/bp/eqa-20160725/
Latest version
http://www.w3.org/2013/share-psi/bp/eqa/
Previous version
http://www.w3.org/2013/share-psi/bp/eqa-20160721/

This is one of a set of Best Practices for implementing the (Revised) PSI Directive developed by the Share-PSI 2.0 Thematic Network.

Outline

Data Quality DQ is primarily perceived to be a subjective term: What suffices, is “good enough” for one person, might be inferior to another. “Suffice” here means to be suitable to fulfil a certain need in a process. However beside the subjective aspect of DQ, there is an objective view on DQ which can be measured and help to establish provable and comprehensible metrics on DQ. The adherence to standards, enforced by tools which in turn are embedded in and used by processes, will help to raise DQ. In order to sustainably raise DQ, measures need to be in place all along the data pipeline and not only at the providing front end. DQ improvement has to be considered as a process rather than a one-time measure.

Links to the Revised PSI Directive

Data Quality

Challenge

The proliferation of open data as a mean to foster open innovation processes towards improved or new products and services, to increase transparency and to perform self-empowered impact measurement of policies also raises concerns about the quality of the provided resources. The early assumption that more data, even of uncertain origin and quality, will unconditionally result in better decisions as long as the right algorithms are used, gave again way to the insight that the principle of garbage-in, garbage-out still holds true. This fact combined with raising concerns regarding data platform usability, data literacy and trust put the quality aspect into the focus. Ironically government Data Quality became of an issue lately primarily due to the fact that government started to release data sets as Open Data which enables stakeholders to carry out citizens control rights. Bringing together data from diverse sources for the first time partially makes data issues like missing data obvious, but even more so deficiencies which arouse due to lacking or missing Master Data

Management.

Solution

Traditional metrics to assess Data Quality like accuracy, applicability, and understandability remain relevant, and in the realm of Open Data, get extended by measures like openness, timeliness and primacy. Work carried out in the European Commission's Open Data Support project suggests seven aspects to consider:

- **Accuracy:** is the data correctly representing the real-world entity or event?
- **Consistency:** Is the data not containing contradictions?
- **Availability:** Can the data be accessed now and over time?
- **Completeness:** Does the data include all data items representing the entity or event?
- **Conformance:** Is the data following accepted standards?
- **Credibility:** Is the data based on trustworthy sources?
- **Processability:** Is the data machine-readable?
- **Relevance:** Does the data include an appropriate amount of data?
- **Timeliness:** Is the data representing the actual situation and is it published soon enough?

DQ improvement measures have to be in place all along the (open) data life cycle, otherwise quality measures will be perceived to be an additional burden, causing efforts and costing money. Also note, that the Open Data Life Cycle is - a cycle which suggest to set up data improvement measures as a process rather than a one time measure.

Why is this a Best Practice?

Lacking DQ will reduce data users trust and prevent the unfolding of an open data market. Investment into DQ will pay back internally to the administration, as the potential for interoperable data services will be risen as well as externally, as for data users it will become more easy to blend together data sets of diverse sources to create added value services.

How do I implement this Best Practice?

Implementation of this BP requires addressing the problem from a technical as well as organisational perspective.

Technically, DQ can be raised by adhering to conventions, norms and standards. However, the adoption of conventions, norms and standards requires governance at various levels. Set-up of governance structures is typically in the responsibility of the CIO or someone in charge with comparable powers and duties.

- It's within the CIO's responsibility to provide guidance on how to structure and implement ICT-systems, which use common and agreed conventions, norms and standards.
- The CIO should be responsible for identifying semantically equivalent data entities, describe standards according to which these data entities should be modeled and monitor the adherence to these standards.

Common data entities, where possible, should be modeled according to the core vocabuilaries.

CSV files could be annotated using W3C's CSV on the Web Recommendations, which also included a formalised model to describe the columns of CSV files.

Data descriptions should be made according to the DCAT-AP vocabulary.

During the data publishing stage, the W3C Data Quality Vocabulary (DQV) can be used. This provides a framework in which the quality of a dataset can be described either by the publisher or the wider audience.

Tools can automatically check a certain range of DQ domains, like adherence to claimed encodings (such as utf8) or the structural regularity of CSV files.

For assessing the quality of the dataset itself prior to publishing, e.g. for publishing statistical data in RDF format an RDF Data Cube validator (PDF) can be used.

To enrich the data with quality assessment information and track provenance in RDF integration process, e.g. the UnifiedViews tool can be used.

Organisation-wise

- The CIO should implement a data governance framework which comprises data architecture management, meta-data management, and master data management (MDM).
- The importance of data as a mission-critical asset can be risen by establishing the role of the Chief Data Officer (CDO).
- The principles of ISO 8000, like vocabulary usage, semantic encoding, provenance, accuracy and completeness can be taken into account.

The obligatory usage of minimum widespread technical standards like utf8 could be enforced by legal measures or order of the federal CIO.

To assess the publishing process, consider the steps described by ODI Certificates (or similar).

Further reading

- Data Quality Vocabulary
- Introduction to ISO 8000
- Data Management Body of Knowledge
- Standards on eProcurement
- Standards on eInvoicing
- Open Data Certificates

Where has this best practice been implemented?

Country	Implementation	Contact Point
Austria	Mission Statement of the Sub-working Group Quality Assurance of Open Data Portals of the Cooperation Open Government Data Austria	Cooperation OGD Austria
UK	Cross platform character encoding profile	
UK	ODI Certificate for the Westminster City Council	Westminster City Council
Serbia	Validating RDF Data Cube Models	Valentina Janev, Mihailo Pupin Institute, University of Belgrade, Belgrade, Serbia

Finland	Valmistele ja avaa - Prepare and open Section 3.6. Tiedon viimeistely ja laatu - Finishing the data and data quality	Ministry of Finance, Finland

References

- David Corsar, Peter Edwards, Enhancing Open Data with Provenance, dot.rural Digital Economy Hub
- ProvenanceWeek 2014
- Giorgos Flouris, Yannis Roussakis, Marrıa Poveda-Villalon, Pablo N. Mendes, Irini Fundulaki, Using Provenance for Quality Assessment and Repair in Linked Open Data, 2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn-12) at the ISWC2012
- Makx Dekkers, AMI Consult, How good is good enough?
- Amanda Smith & Sumika Sakanishi, ODI, Publishing and improving the quality of open data with Open Data Certificates, United Kingdom
- Samos presentation: Examples from the Norwegian public sector
- Lisbon workshop session: Roadblocks in Commercial Open Data Usage
- Timisoara workshop session: How good is good enough? A common language for quality?
- Comparing the 5-star scheme with Open Data Certificates
- Lisbon workshop session: Roadblocks in Commercial Open Data Usage
- Samos Workshop Session: The Potential within the Government for Innovation and Efficiency from Open Data – Examples from the Norwegian public Sector

Local Guidance

This Best Practice is cited by, or is consistent with, the advice given within the following guides:

- (Austria) Framework for Open Government Data Platforms
- (Belgium) Open Data Handleiding Open Data Handbook
- (Croatia) Preporuke o prilagodbi skupova podataka za javnu objavu i ponovno korištenje Open Data Guide, Croatia
- (Estonia) Avaandmete loomise ja avaldamise juhend Open Data Guidelines
- (Finland) Avoimen Datan Opas Open Data Guide
- (Greece) Εφαρμογή των διατάξεων του Κεφαλαίου Α’ του ν. 4305/2014 (ΦΕΚ 237/Α΄) Guidelines on the implementation of open data policy and l. 4305/2014
- (International) Open Data Handbook, Solutions Bank
- (Italy) Linee Guida Nazionali per la Valorizzazione del Patrimonio Informativo Pubblico National Development Guidelines for Public Sector Information
- (Lithuania) Viešojo Sektoriaus Informacijos platinimo gerosios praktikos Best Practices for Sharing Public Sector Information
- (Luxembourg) Recommandations pour l'ouverture des données publiques Recommendations for opening data
- (Malta) PSI Directive Implementation & Internal Data Sharing Platform (draft)
- (Serbia) Open Data Handbook
- (Slovenia) Priročnik za odpiranje podatkov javnega sektorja Manual for the opening of public sector information
- (Spain) Government Data Openness and Re-use
- (UK) Open Data Resource Pack
- (UK) Birmingham and West Midlands Localised Guide for Open Data

Contact Info

Original Authors: Johann Höchtl, Valentina Janev

Contributors: Muriel Foulonneau, Lorenzo Canova

Editors: Valentina Janev, Johann Höchtl

Related Best Practices

- Enable feedback channels for improving the quality of existing government data,/li>
- Provide data provenance information
- Provide versioning information
- Reuse vocabularies

Issue Tracker

Any matters arising from this BP, including implementation experience, lessons learnt, places where it has been implemented or guides that cite this BP can be recorded and discussed on the project's GitHub repository