European Commission

OPEN DATA SUPPORT

Training Module 2.2

Open Data & Metadata Quality

pwc

Presentation metadata

Open Data Support is funded by the European Commission under SMART 2012/0107 'Lot 2: Provision of services for the Publication, Access and Reuse of Open Public Data across the European Union, through existing open data portals'(Contract No. 30-CE-0530965/00-17).

This presentation has been created by PwC

Authors:
Makx Dekkers, Nikolaos Loutas, Michiel De Keyzer and Stijn Goedertier

Disclaimers

1. The views expressed in this presentation are purely those of the authors and may not, in any circumstances, be interpreted as stating an official position of the European Commission.
The European Commission does not guarantee the accuracy of the information included in this presentation, nor does it accept any responsibility for any use thereof.
Reference herein to any specific products, specifications, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favouring by the European Commission.
All care has been taken by the author to ensure that s/he has obtained, where necessary, permission to use any parts of manuscripts including illustrations, maps, and graphs, on which intellectual property rights already exist from the titular holder(s) of such rights or from her/his or their legal representative.

2. This presentation has been carefully compiled by PwC, but no representation is made or warranty given (either express or implied) as to the completeness or accuracy of the information it contains. PwC is not liable for the information in this presentation or any decision or consequence based on the use of it.. PwC will not be liable for any damages arising from the use of the information contained in this presentation. The information contained in this presentation is of a general nature and is solely for guidance on matters of general interest. This presentation is not a substitute for professional advice on any particular matter. No reader should act on the basis of any matter contained in this publication without considering appropriate professional advice.

Learning objectives

By the end of this training module you should have an understanding of:

- What (open) data quality means;
- The open data quality determinants and criteria;
- Good practices for publishing high-quality (linked) open data.

Content

This modules contains…

- A definition of data quality;
- An overview of the dimensions of data and metadata quality;
- A selection of best practices for publishing good quality data and metadata.

Find more on: ***training.opendatasupport.eu***

What is data (and metadata) quality?

Data is of high quality "if they are fit for their intended uses in operations, decision making and planning."

Or more specifically:

“High quality data are accurate, available, complete, conformant, consistent, credible, processable, relevant and timely.”

The data quality dimensions

What are the main dimensions to be taken into account for delivering good quality (meta)data?

Data quality dimensions

- **Accuracy**: is the data correctly representing the real-world entity or event?
- **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?
- **Consistency**: Is the data not containing contradictions?
- **Credibility**: Is the data based on trustworthy sources?
- **Processability**: Is the data machine-readable?
- **Relevance**: Does the data contain the necessary information to support usage and the application?
- **Timeliness**: Is the data representing the actual situation and is it published soon enough?

Accuracy

The accuracy of data is the extent to which it correctly represents the characteristics of the real-world object, situation or event.

For example:

- Correct measurement of weather conditions (temperature, precipitation).
- Correct indication of re-use conditions of the dataset.

Recommendations:

- **Balance** the **accuracy** of your data against the **cost** in the context of the application; it needs to be **good enough for the intended use**.
- Make sure that there is **organisational commitment** and **investment in procedures and tools** to maintain accuracy .

Accuracy by example

High accuracy

Cloud cover = 60%

Less accuracy

Partly cloudy

Availability

The availability of data is the extent to which it can be accessed; this also includes the long-term persistence of data.

For example:

- A Dataset that is identified by a http: URI that resolves persistently to the right resource (and does not give back 404 Not found).
- A description of the dataset that is included in the search engine of a data portal.

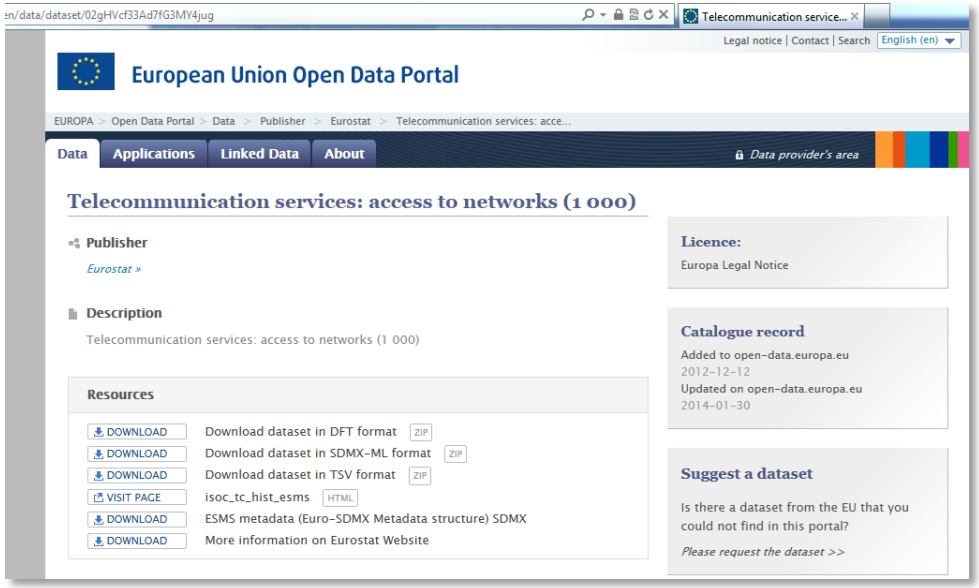
Recommendations:

- Follow **best practices** for the assignment and maintenance of URIs.
- Make sure **that responsibility** for the maintenance of data is **clearly assigned** in the organisation.

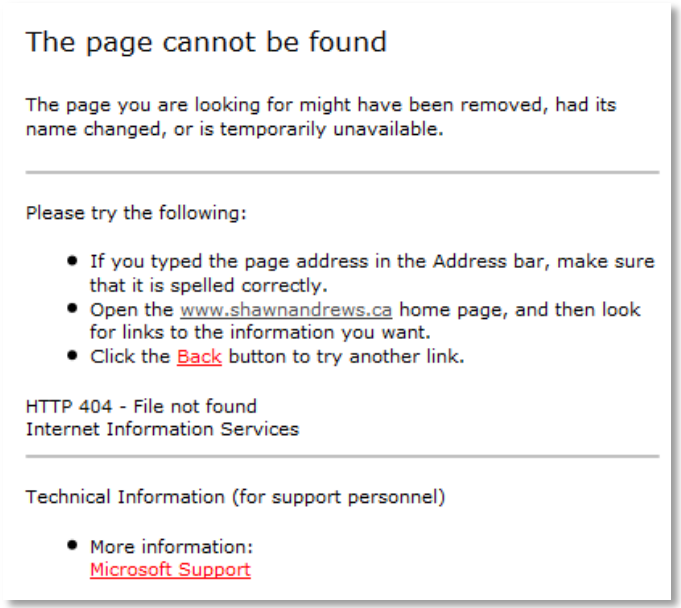
See also:
http://www.slideshare.net/OpenDataSupport/design-and-manage-persitent-uris

Availability by example

High availability



Less availability



Completeness

The completeness of data is the extent to which it includes the data items or data points that are necessary to support the application for which it is intended.

For example:

- A Dataset that includes spending data for all ministries enables a complete overview of government spending.
- A description of data that is generated in real time that includes the date and time of last modification.

Recommendations:

- **Design the capture and publication process** to include the necessary data points.
- **Monitor** the update mechanisms on a continuous basis.

Completeness by example

High completeness

```
:weather1-7 a dcat:Dataset ;
   dct:title "Measurements from weather stations 1-7" ;
   dct:description "Data from seven weather stations
                    showing temparture, humidity,
                    wind direction and wind speed" ;
   dct:modified "2013-07-01T19:20:30+01:00" ;
   dct:publisher <http://myweather.com/id/myweather> ;
   dcat:keyword "weather" ;
   dcat:landingpage <http://myweather.com/stations1-7.html> ;
   dcat:distribution :weatherdata-xlsx
   .

:weatherdata1-7-xlsx a dcat:Distribution ;
   dct:format <http://publications.europa.eu/resource/authority/file-type/XLSX> ;
   dct:licence <http://creativecommons.org/licenses/CC0> ;
   dcat:downloadURL <http://myweather.com/stations1-7.xlsx>
   .
```

Less completeness

```
:weather1-7 a dcat:Dataset ;
   dct:title "Measurements from weather stations 1-7" ;
   dct:description "Data from seven weather stations
                    showing temparture, humidity,
                    wind direction and wind speed" ;
   dct:publisher <http://myweather.com/id/myweather> ;
   dcat:keyword "weather" ;
   dcat:landingpage <http://myweather.com/stations1-7.html> ;
   dcat:distribution :weatherdata-xlsx
   .

:weatherdata1-7-xlsx a dcat:Distribution ;
   dct:format <http://publications.europa.eu/resource/authority/file-type/XLSX> ;
   dct:licence <http://creativecommons.org/licenses/CC0> ;
   dcat:downloadURL <http://myweather.com/stations1-7.xlsx>
   .

ERROR: MISSING DATA dct:modified
```

Conformance

The conformance of data is the extent to which it follows a set of explicit rules or standards for capture, publication and description

For example:

- A Dataset that expresses coordinates in WGS84 and statistics in SDMX.
- A description of a dataset according to the DCAT Application Profile.

Recommendations:

- **Apply the most used standards** in the domain that is most relevant for the data or metadata.
- **Define local vocabularies if no standard is available**, but publish your vocabularies according to best practice (e.g. dereferenceable URIs).

Conformance by example

High conformance

```
:weather1-7 a dcat:Dataset ;
   dct:title "Measurements from weather stations 1-7" ;
   dct:description "Data from seven weather stations
                    showing temparture, humidity,
                    wind direction and wind speed" ;
   dct:modified "2013-07-01T19:20:30+01:00" ;
   dct:publisher <http://myweather.com/id/myweather> ;
   dcat:keyword "weather" ;
   dcat:landingpage <http://myweather.com/stations1-7.html> ;
   dcat:distribution :weatherdata-xlsx
   .

:weatherdata1-7-xlsx a dcat:Distribution ;
   dct:format <http://publications.europa.eu/resource/authority/file-type/XLSX> ;
   dct:licence <http://creativecommons.org/licenses/CC0> ;
   dcat:downloadURL <http://myweather.com/stations1-7.xlsx>
   .
```

Less conformance

```
:weather1-7 a dcat:Dataset ;
   dct:description "Data from seven weather stations
                    showing temparture, humidity,
                    wind direction and wind speed" ;
   dct:modified "2013-07-01T19:20:30+01:00" ;
   dct:publisher <http://myweather.com/id/myweather> ;
   dcat:keyword "weather" ;
   dcat:landingpage <http://myweather.com/stations1-7.html> ;
   dcat:distribution :weatherdata-xlsx
   .

:weatherdata1-7-xlsx a dcat:Distribution ;
   dct:format <http://publications.europa.eu/resource/authority/file-type/XLSX> ;
   dct:licence <http://creativecommons.org/licenses/CC0> ;
   dcat:downloadURL <http://myweather.com/stations1-7.xlsx>
   .

ERROR MISSING MANDATORY ELEMENT dct:title
```

See also:
https://joinup.ec.europa.eu/asset/adms_foss/news/just-released-admssw-validator-verify-and-visualise-rdf-software-metadata

Consistency

The consistency of data is the extent to which it does not contain contradictions that would make its use difficult or impossible.

For example:

- A dataset that combines data from different sources that has been processed to detect conflicting statements which have been resolved.
- A description of a dataset that does not contain multiple licence statements or where the data of last modification is not before the creation date.

Recommendations:

- **Process all data before publication** to detect conflicting statements and other errors (in particular if data is aggregated from different sources).

Consistency by example

High consistency

```
:weather1-7 a dcat:Dataset ;
   dct:title "Measurements from weather stations 1-7" ;
   dct:description "Data from seven weather stations
                    showing temparture, humidity,
                    wind direction and wind speed" ;
   dct:issued "2013-01-01T00:00:00+01:00" ;
   dct:modified "2013-07-01T19:20:30+01:00" ;
   dct:publisher <http://myweather.com/id/myweather> ;
   dcat:keyword "weather" ;
   dcat:landingpage <http://myweather.com/stations1-7.html> ;
   dcat:distribution :weatherdata-xlsx
   .

:weatherdata1-7-xlsx a dcat:Distribution ;
   dct:format <http://publications.europa.eu/resource/authority/file-type/XLSX> ;
   dct:licence <http://creativecommons.org/licenses/CC0> ;|
   dcat:downloadURL <http://myweather.com/stations1-7.xlsx>
   .
```

Less consistency

```
:weather1-7 a dcat:Dataset ;
   dct:title "Measurements from weather stations 1-7" ;|
   dct:description "Data from seven weather stations
                    showing temparture, humidity,
                    wind direction and wind speed" ;
   dct:issued "2014-01-01T00:00:00+01:00" ;
   dct:modified "2013-07-01T19:20:30+01:00" ;
   dct:publisher <http://myweather.com/id/myweather> ;
   dcat:keyword "weather" ;
   dcat:landingpage <http://myweather.com/stations1-7.html> ;
   dcat:distribution :weatherdata-xlsx
   .

ERROR INCONSISTENT DATA: Issue date is after modification date

:weatherdata1-7-xlsx a dcat:Distribution ;
   dct:format <http://publications.europa.eu/resource/authority/file-type/XLSX> ;
   dct:licence <http://creativecommons.org/licenses/CC0> ;
   dct:licence <http://creativecommons.org/licenses/BY/3.0> ;
   dcat:downloadURL <http://myweather.com/stations1-7.xlsx>
   .

ERROR INCONSISTENT DATA: Licence element repeated
```

Credibility

The credibility of data is the extent to which it is based on trustworthy sources or delivered by trusted organisations.

For example:

- A dataset that contains data from processes that can be independently verified, e.g. election results or parliamentary proceedings.
- A description of a dataset that is published by a government agency.

Recommendations:

- **Base data on sources that can be trusted** or on explicit Service Level Agreements where possible and appropriate.
- **Make appropriate attributions** so that re-users can determine whether or not they can trust the data.

Credibility by example

High credibility

Data coming from the Publications Office of the EU:

```
<skos:ConceptScheme at:table.version.number="2013-05-29 14:01:09" at:table.id="language"
rdf:about="http://publications.europa.eu/resource/authority/language">
    <rdfs:label>Languages Authority Table</rdfs:label>
    <at:prefLabel xml:lang="en">Languages Authority Table</at:prefLabel>
</skos:ConceptScheme>
<skos:Concept rdf:about="http://publications.europa.eu/resource/authority/language/ENG" at:p
    <skos:inScheme rdf:resource="http://publications.europa.eu/resource/authority/language"/>
    <at:authority-code>ENG</at:authority-code>
    <at:op-code>ENG</at:op-code>
    <atold:op-code>ENG</atold:op-code>
    <dc:identifier>ENG</dc:identifier>
    <at:start.use>1950-05-09</at:start.use>
    <skos:prefLabel xml:lang="bg">английски</skos:prefLabel>
    <skos:prefLabel xml:lang="cs">angličtina</skos:prefLabel>
    <skos:prefLabel xml:lang="da">engelsk</skos:prefLabel>
    <skos:prefLabel xml:lang="de">Englisch</skos:prefLabel>
    <skos:prefLabel xml:lang="el">αγγλικά</skos:prefLabel>
```

The Metadata Registry is maintained by the Publications Office of the EU.

Less credibility

Data coming from Lexvo:

```
- <rdf:Description rdf:about="http://lexvo.org/id/iso639-3/eng">
    <rdf:type rdf:resource="lvont:Language"/>
    <rdfs:comment xml:lang="en" rdf:datatype="xsd:string">English is a West
      Germanic language that arose in the Anglo-Saxon kingdoms of England and
      spread into what was to become south-east Scotland under the influence of
      the Anglian medieval kingdom of Northumbria. Following the extensive
      influence of Great Britain and the United Kingdom from the 18th century, via
      the British Empire, and of the United States since the mid-20th century, it
      has been widely dispersed around the world, becoming the leading language
      of international discourse and the lingua franca in many regions. It is widely
      learned as a second language and used as an official language of the
      European Union and many Commonwealth countries, as well as in many
      world organisations. It is the third most natively spoken language in the
      world, after Mandarin Chinese and Spanish.</rdfs:comment>
    <rdfs:label xml:lang="aa" rdf:datatype="xsd:string">English</rdfs:label>
    <rdfs:label xml:lang="ace" rdf:datatype="xsd:string">Bahsa Inggréh</rdfs:label>
    <rdfs:label xml:lang="af" rdf:datatype="xsd:string">Engels</rdfs:label>
    <rdfs:label xml:lang="agq" rdf:datatype="xsd:string">Kɨŋgele</rdfs:label>
    <rdfs:label xml:lang="aii" rdf:datatype="xsd:string">[illegible]</rdfs:label>
    <rdfs:label xml:lang="ak" rdf:datatype="xsd:string">Borɔfo</rdfs:label>
    <rdfs:label xml:lang="ak" rdf:datatype="xsd:string">English</rdfs:label>
```


Lingvoj/Lexvo data may not be of less quality than Publications Office data, but the Publications Office is an authoritative source, while Linvoj and Lexvo are initiatives of individuals.

Processability

The processability of data is the extent to which it can be understood and handled by automated processes.

For example:

- A dataset that contains coded information based on publicly available controlled vocabularies and code lists.
- A description of a dataset that expresses dates in W3C Date and Time Format (e.g. 2013-06-01) rather than as text (e.g. 1 June 2013).

Recommendations:

- **Identify the source of terminology and codes** used in the data in machine-readable manner.
- **Apply recommendations for syntax** of data given in common standards and application profiles.

Processability by example

High processability

```
:weather1-7 a dcat:Dataset ;
   dct:title "Measurements from weather stations 1-7" ;
   dct:description "Data from seven weather stations
                    showing temparture, humidity,
                    wind direction and wind speed" ;
   dct:issued "2013-01-01T00:00:00+01:00"
   .
```

Less processability

```
:weather1-7 a dcat:Dataset ;
   dct:title "Measurements from weather stations 1-7" ;
   dct:description "Data from seven weather stations
                    showing temparture, humidity,
                    wind direction and wind speed" ;
   dct:issued "January 1, 2013"
   .
```

Relevance

The relevance of data is the extent to which it contains the necessary information to support the application.

For example:

- A Dataset that contains temperature measurements rounded to degrees Celsius for climate calculations; a dataset with precision of a thousandth of a degree for chemical reactions.
- A description of a dataset that only contains temporal coverage data if necessary for its processing .

Recommendations:

- **Match coverage and granularity** of data to its intended use within constraints of available time and money.
- However, also **consider potential future usages** of the data.

Relevance by example

High relevance

Table to determine special tax on motor vehicles based on emission CO_2

	Engine displacement (cm^3)	Fuel type	CO_2 (g/km)	Tax (%)
Car type 1	900	Gasoline	90	0
Car type 2	1100	Gasoline	120	5
Car type 3	1300	Gasoline	125	5
Car type 4	1400	Gasoline	150	5
Car type 5	1800	Diesel	180	10
Car type 6	2200	Diesel	190	10
Car type 7	2500	Gasoline	210	15

Less relevance

Table to determine special tax on motor vehicles based on emission CO_2

	Engine displacement (cm^3)	Fuel type	CO_2 (g/km)	Weight (kg)	Tax (%)
Car type 1	900	Gasoline	90	750	0
Car type 2	1100	Gasoline	120	1000	5
Car type 3	1300	Gasoline	125	1200	5
Car type 4	1400	Gasoline	150	1200	5
Car type 5	1800	Diesel	180	1700	10
Car type 6	2200	Diesel	190	1600	10
Car type 7	2500	Gasoline	210	1900	15

Weight is not a relevant data item

Timeliness

The timeliness of data is the extent to which it correctly reflects the current state of the entity or event and the extent to which the data (in its latest version) is made available without unnecessary delay

For example:

- A dataset that contains real-time traffic data that is refreshed every few minutes.
- A description of a dataset containing annual crime statistics that is made available within days of publication of the dataset.

Recommendations:

- **Adapt the update frequency** of data **to the nature of the data** and its **intended use**.
- Make sure that **processes and tools are in place** to support the updating.

Timeliness: examples

High timeliness

HICP - annual average indices for transport prices
(2005 = 100)
The Harmonised Indices of Consumer Prices (HICPs) are a set of European Union Consumer ... more

coicop: Purchase of vehicles

geo \ time	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
EU (28 countries)	97.82	98.85	99.66	100.00	100.63	101.62	101.28	101.17	101.73	102.34	102.64	102.75p
EU (27 countries)	97.82	98.85	99.66	100.00	100.63	101.62	101.28	101.17	101.73	102.35	102.65	102.77p
Euro area (17 countries)	96.89	98.17	99.02	100.00	101.11	102.50	102.68	102.16	102.15	103.17	103.94	104.48p
Belgium	96.87	98.32	98.91	100.00	101.37	102.12	102.68	103.37	104.00	104.95	104.71	105.08
Bulgaria	99.61	102.16	100.97	100.00	100.79	103.88	109.05	103.04	97.18	91.74	88.05	83.54
Czech Republic	109.9	107.6	104.1	100.0	98.9	97.3	92.4	84.1	74.8	71.9	69.6	68.1
Denmark	91.8	96.1	98.1	100.0	100.3	99.4	98.4	97.7	97.2	96.2	95.2	94.4
Germany	97.7	98.5	99.6	100.0	101.3	104.4	105.2	105.2	105.3	106.0	106.2	106.3

Less timeliness

Greenhouse gas emissions from transport
1 000 tonnes of CO2 equivalent
This indicator shows trends in the emissions from transport (road, rail, inland ... more

Code: tsdtr41

Flags Codes Labels Cod

geo \ time	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
EU (28 countries)	900269	919853	914698	929195	940183	950126	969342	969045	975836	985438	965593	941412	935862	926442
EU (27 countries)	895970	915271	910101	924550	935279	944833	963866	963364	969844	979020	959331	935146	929822	920553
Belgium	24223	24575	24869	25479	25782	26339	27334	26354	25771	25653	27975	27230	27128	27047
Bulgaria	5784	6013	5739	5878	6122	6707	7007	7697	8320	8140	8525	8183	7954	8129
Czech Republic	12000	12223	12364	13252	13878	15758	16570	17944	18280	19234	19072	18498	17424	17255
Denmark	12542	12560	12355	12365	12460	12917	13225	13339	13716	14334	14094	13288	13223	12865
Germany	181805	187057	183037	179107	176758	170257	169972	161756	157984	154574	154447	153952	154956	157179
Estonia	1798	1679	1667	1996	2125	2019	2066	2137	2296	2421	2304	2126	2248	2260
Ireland	9119	9731	10770	11297	11492	11697	12419	13110	13892	14482	13745	12525	11603	11290

Best practices

Best practices for publishing high-quality data and metadata.

Best practices for publishing high-quality data and metadata

- **Provide appropriate descriptions** of data (i.e. metadata).
- **Use standard vocabularies** for metadata and data whenever such vocabularies exist.
- **Specify the license** under which the data may be re-used.
- **Adhere to legal requirements** concerning protection of personal and other sensitive data.
- **Represent** metadata and data **according to the Linked Data principles** using **persistent URIs** for identifying things.
- **Provide information about the source** of the data.

Maintenance of metadata and data is critical!

See also:
http://www.slideshare.net/OpenDataSupport/introduction-to-metadata-management

Conclusions

- The quality of data is determined by its fitness for (re-)use by data consumers.
- Metadata is “data about data”, i.e. metadata is a type of data.
  - The same quality considerations apply to data and metadata alike.
- Data quality has multiple dimensions and is about more than the correctness of data.
  - Accuracy, availability, completeness, conformance, consistency, credibility, processability, relevance, timeliness.

Group questions

http://www.visualpharm.com

In your opinion, which factors contribute the most to data and/or metadata quality?

http://www.visualpharm.com

Improving quality can require time and resources. To which extent would your organisation be willing to invest in data and/or metadata quality?

Thank you!

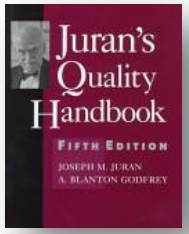
...and now YOUR questions?

References

- Juran, Joseph M. and A. Blanton Godfrey, Juran's Quality Handbook, Fifth Edition, p. 2.2, McGraw-Hill, 1999
- National Information Standards Organization, http://www.niso.org/publications/press/UnderstandingMetadata.pdf
- Mark David Hansen. Zero Defect Data: Tackling the Corporate Data Quality Problem. 1991. http://dspace.mit.edu/handle/1721.1/13812
- Kevin Roebuck. Data Quality: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Emereo Pty Limited, 2011. http://bit.ly/19Qb6Ov
- Thomas R. Bruce, Diane Hillmann. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. ALA Editions, 2004. http://www.ecommons.cornell.edu/handle/1813/7895
- Sharon Dawes. Open data quality: a practical view. Open Data Roundtable. October 2012. http://www.slideshare.net/cityhub/sharon-dawes-ctg
- Joshua Tauberer. Open Government Data. Section 5.2 Data Quality: Precision, Accuracy, and Cost. June 2012. http://opengovdata.io/2012-02/page/5-2/data-quality-precision-accuracy-and-cost
- Stefan Urbanek. Data Quality: What is It? January 2011. http://ckan.org/2011/01/20/data-quality-what-is-it/
- Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, Sören Auer. Quality Assessment Methodologies for Linked Open Data. Semantic Web Journal (unpublished), 2012. http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-open-data
- ISA Programme. 10 Rules for Persistent URIs. https://joinup.ec.europa.eu/community/semic/document/10-rules-persistent-uris
- W3C. Best Practices for Publishing Linked Data. W3C Note 06 June 2013. https://dvcs.w3.org/hg/gld/raw-file/default/bp/index.html
- OPQUAST. 72 Open data good practices. http://checklists.opquast.com/en/opendata

Further reading

Joshua Tauberer. Open Government Data. http://opengovdata.io/



Juran, Joseph M. and A. Blanton Godfrey, Juran's Quality Handbook

Related projects and initiatives

Best Practices for Publishing Linked Data.
https://dvcs.w3.org/hg/gld/raw-file/default/bp/index.html

OPQUAST. Open data good practices.
http://checklists.opquast.com/en/opendata

Be part of our team…

Find us on

Open Data Support
http://www.slideshare.net/OpenDataSupport

Open Data Support
http://goo.gl/y9ZZI

Join us on

http://www.opendatasupport.eu

Follow us

@OpenDataSupport

Contact us

contact@opendatasupport.eu