OPEN
DATA
SUPPORT

European Commission

Open Data & Metadata Quality

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).

© 2013 European Commission

This presentation has been created by PwC

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

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.

OPEN DATA SUPPORT

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.

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."

What is metadata?

"Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data"

-- National Information Standards Organization

- We observe that metadata is a type of data.
- The same quality considerations apply to data and metadata alike.

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 include an appropriate amount of data?

- **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

Image licence = <u>Creative Commons Attribution-Share Alike 3.0 Unported</u> license.

Attribution: <u>Garytx</u> at <u>en.wikipedia</u>

High accuracy

Cloud cover = 100%

Less accuracy

Partly cloudy

OPEN DATASUPPORT

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
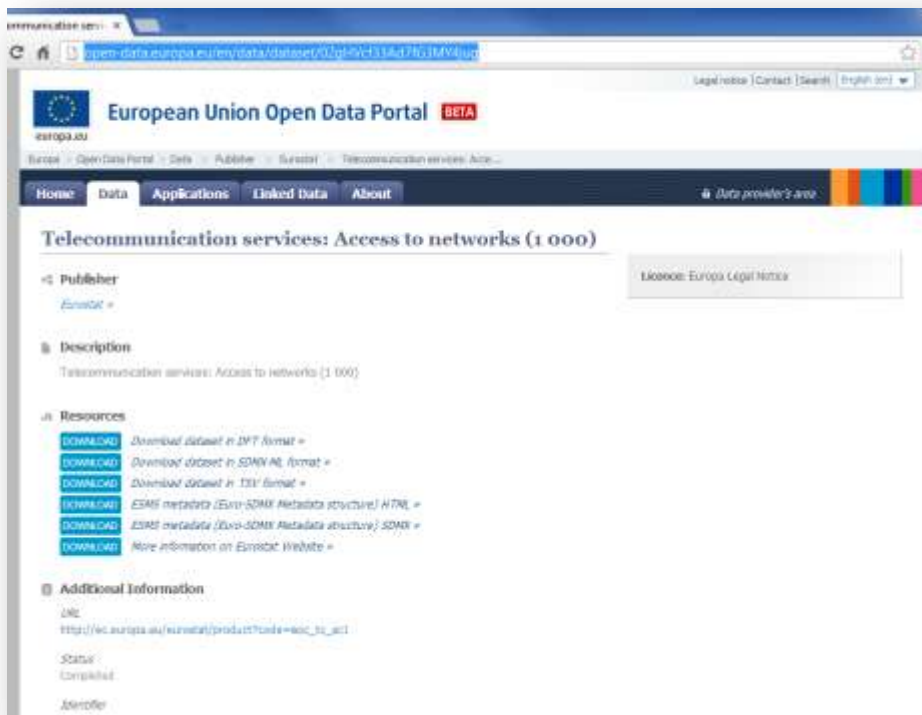
Less availability

OPEN DATA SUPPORT
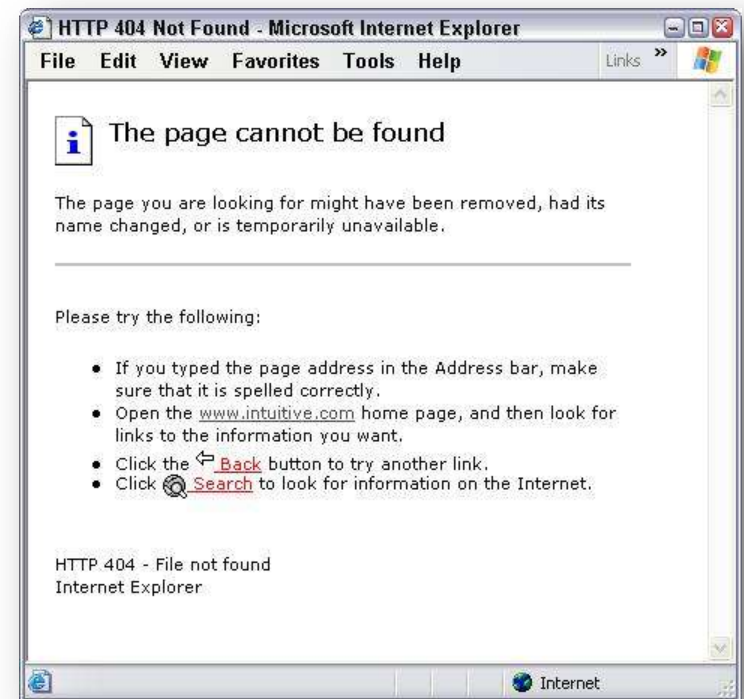
Completeness

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

OPEN DATASUPPORT

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).

OPEN DATA SUPPORT

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	**Less credibility**
Data coming from the Publications Office of the EU:	Data coming from Lexvo:

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.

OPEN DATA SUPPORT

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.

OPEN **DATA** SUPPORT

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.

OPEN DATA SUPPORT

Timeliness: examples

High timeliness

Weather report of 1 July 2013, 12:00				
	Temp. ºC	Humidity %	Wind direction	Wind speed km/h
Station 1	18.1	60	WSW	18
Station 2	17.5	59	WSW	20
Station 3	18.2	55	SW	22
Station 4	19.0	62	SW	18
Station 5	18.0	65	WSW	19
Station 6	18.2	63	SSW	21
Station 7	17.9	61	SW	22

Published 1 July 2013, 12:05

Less timeliness

Weather report of 1 July 2013, 12:00				
	Temp. ºC	Humidity %	Wind direction	Wind speed km/h
Station 1	18.1	60	WSW	18
Station 2	17.5	59	WSW	20
Station 3	18.2	55	SW	22
Station 4	19.0	62	SW	18
Station 5	18.0	65	WSW	19
Station 6	18.2	63	SSW	21
Station 7	17.9	61	SW	22

Published 1 July 2013, 20:00

OPEN DATA SUPPORT

Best practices

Best practices for publishing high-quality data and metadata.

OPEN DATA SUPPORT

W3C: Best practices for publishing Linked Open Data

IDENTIFY Indentify data sets that other people may wish to re-use.

MODEL Model the data in an application-independent, objective way in terms of representation. Denormalize the data as necessary.

METADATA Provide basic metadata, including MIME type, publishing organization and/or agency, creation date, modification date, version, frequency of updates, contact email for the data steward(s).

PII Do not Publish Personally Identifiable Information as Open Data on the Web Data on the public Web can be potentially misused. Examples of personally identifiable data include: individual names, national identification number, phone number, credit card number and driver license number.

NAME Use HTTP URIs as names for your objects. Give careful consideration to the URI naming strategy. Consider how the data will change over time and name as necessary.

STANDARD_VOCABULARIES Describe objects with standard vocabularies whenever possible.

VOCABULARY_USE Use vocabularies as loosely coupled modular components.

REPRESENTATION Convert the source data into a Linked Data representation, also called an RDF serialization including Turtle, Notation-3 (N3), N-Triples, XHTML with embedded RDFa, and RDF/XML.

HUMAN READABLE Provide human readable descriptions with your Linked Data.

MACHINE ACCESSIBLEProvide access to the data representation via RESTful API, SPARQL endpoint(s) and RDF download.

SPECIFY_LICENSE Specify an appropriate license.

HOST Deliver open government data on authoritative domain to increase perceived trust.

ANNOUNCE Announce open government data, have a feedback mechanism and be prepared to be responsive to feedback.

SOCIAL_CONTRACT Maintenance is critical. Without a permanent identifier scheme, if you move or remove data that is published to the Web, you may break third party applications or mashups which is clearly undesireable. URI strategy and implementation are critical.

See also:
http://www.slideshare.net/OpenDataSupport/the-linked-open-government-data-lifecycle

OPEN DATASUPPORT

Opquast: 72 Open data good practices

Some examples

Metadata	1	23	Each dataset is accompanied by a descriptive record
Metadata	1	24	Each dataset includes at least a title and a description
Metadata	1	25	A creation date is given for each dataset
Metadata	1	26	A last-updated date is given for each dataset
Metadata	2	27	The datasets are categorised
Format	1	33	Each dataset includes a reference to the charset used
Format	1	34	The format of downloadable files is indicated
Format	1	35	Dates are given in documented formats
Format	2	36	Dates are available in a standard format
Format	2	37	Data is provided in at least one open format

License	1	47	The datasets are accompanied by a licence
License	1	48	The licence sets out the conditions of attribution, reuse, redistribution and commercialisation
License	2	49	Usage rights are provided for an unlimited period
License	2	50	Data producers declare their policy on releasing data
License	3	51	The datasets are accompanied by a summary and a link to the full version of the license
Linkeddata	2	52	Any vocabularies used within the dataset are identified and documented
Linkeddata	3	53	Data adheres to the defined syntax of any specified vocabularies
Linkeddata	3	54	It is possible to query data and metadata in accordance with standards of the web of data (Linked Open Data)

See also:
http://checklists.opquast.com/en/opendata

What are the common elements in the best practices

- **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

OPEN DATASUPPORT

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.

OPEN DATA SUPPORT

Slide 31

Group questions

In your opinion, which factors contribute the most to metadata quality?

http://www.visualpharm.com

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

http://www.visualpharm.com

Could you provide an example of high quality metadata description for a dataset using the DCAT Application Profile?

http://www.visualpharm.com

Thank you!
...and now YOUR questions?

OPEN DATA SUPPORT

References

Slide 5:

* Juran, Joseph M. and A. Blanton Godfrey, Juran's Quality Handbook, Fifth Edition, p. 2.2, McGraw-Hill, 1999

Slide 6:

* National Information Standards Organization, http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Slide 8-26:

* 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

Slide 10:

* Garytx at en.wikipedia

Slide 11:

* European Commission. Telecommunication services: Access to networks (1 000). http://open-data.europa.eu/

Slide 13:

* ISA Programme. 10 Rules for Persistent URIs. https://joinup.ec.europa.eu/community/semic/document/10-rules-persistent-uris

Slide 28:

* W3C. Best Practices for Publishing Linked Data. W3C Note 06 June 2013. https://dvcs.w3.org/hg/gld/raw-file/default/bp/index.html

Slide 29:

* OPQUAST. 72 Open data good practices. http://checklists.opquast.com/en/opendata

OPEN DATASUPPORT

Further reading

Joshua Tauberer. Open Government Data. http://opengovdata.io/
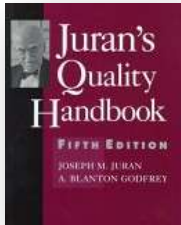
Juran, Joseph M. and A. Blanton Godfrey, Juran's Quality Handbook

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