Training Module 2.2

Open Data & Metadata Quality

PwC firms help organisations and individuals create the value they’re looking for. We’re a network of firms in 158 countries with close to 180,000 people who are committed to delivering quality in assurance, tax and advisory services. Tell us what matters to you and find out more by visiting us at www.pwc.com.

PwC refers to the PwC network and/or one or more of its member firms, each of which is a separate legal entity. Please see www.pwc.com/structure for further details.
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).

© 2014 European Commission

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.
**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.”*
**Metadata is data about data...**

“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?
- **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?
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

Higher accuracy
OpenStreetMap, City of Utrecht, The Netherlands (2011 vs. 2007)

Less accuracy
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 temperature, humidity, wind direction and wind speed";
  dct:issued "2013-01-01T00:00:00+01:00";
  dct:modified "2013-01-01T00:00:00+01:00";
  dcat:publisher <http://myweather.com/id/myweather>;
  dcat:keyword "weather";
  dcat:distribution :weatherdata-xlsx .
```

**Less consistency**

```
:weatherdata1-7-xlsx a dcat:Distribution;
  dct:licence <http://creativecommons.org/licenses/CC0> ;
```

**ERROR INCONSISTENT DATA: Issue date is after modification date**

```
:weatherdata1-7-xlsx a dcat:Distribution;
  dct:licence <http://creativecommons.org/licenses/CC0> ;
```

**ERROR INCONSISTENT DATA: Licence element repeated**
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](http://www.slideshare.net/OpenDataSupport/design-and-manage-persitent-uris)
Availability by example

High availability

Less availability

The page cannot be found

The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

Please try the following:

- If you typed the page address in the Address bar, make sure that it is spelled correctly.
- Open the [www.shawnandrews.ca](http://www.shawnandrews.ca) home page, and then look for links to the information you want.
- Click the Back button to try another link.

HTTP 404 - File not found

Internet Information Services

Technical Information (for support personnel)

- More information: [Microsoft Support](http://microsoftsupport)
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 temperature, humidity, wind direction and wind speed" ;
dct:modified "2013-07-01T19:30:01Z" ;
dct:publisher <http://myweather.com/id/myweather> ;
dct:keyword "weather" ;
dct:distribution :weatherdata-xlsx .

:weatherdata1-7-xlsx a dcat:Distribution ;
dct:licence <http://creativecommons.org/licenses/cc0/> ;
```

Less completeness

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

:weatherdata1-7-xlsx a dcat:Distribution ;
dct:licence <http://creativecommons.org/licenses/cc0/> ;
```

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**

```xml
:weather1-7 a dcat:Dataset;
  dct:title "Measurements from weather stations 1-7";
  dct:description "Data from seven weather stations showing temperature, humidity, wind direction and wind speed";
  dct:modified "2013-07-01T9:20:30+01:00";
  dct:publisher <http://myweather.com/id/myweather>;
  dct:keyword "weather";
  dcat:distribution :weatherdata-xlsx .

:weatherdata1-7-xlsx a dcat:Distribution;
  dct:description "data from seven weather stations showing temperature, humidity, wind direction and wind speed";
  dct:modified "2013-07-01T9:20:30+01:00";
  dct:publisher <http://myweather.com/id/myweather>;
  dct:keyword "weather";
  dcat:distribution :weatherdata-xlsx .
```

**Less conformance**

```xml
:weather1-7 a dcat:Dataset;
  dct:description "Data from seven weather stations showing temperature, humidity, wind direction and wind speed";
  dct:modified "2013-07-01T9:20:30+01:00";
  dct:publisher <http://myweather.com/id/myweather>;
  dct:keyword "weather";
  dcat:distribution :weatherdata-xlsx .

:weatherdata1-7-xlsx a dcat:Distribution;
  dct:description "data from seven weather stations showing temperature, humidity, wind direction and wind speed";
  dct:modified "2013-07-01T9:20:30+01:00";
  dct:publisher <http://myweather.com/id/myweather>;
  dct:keyword "weather";
  dcat:distribution :weatherdata-xlsx .
```

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:

Less credibility
Data coming from Lexvo:

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

Lingvoj/Lexvo data may not be of less quality than Publications Office data, but the Publications Office is an authoritative source, while Lingvoj 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

Higher processability

Hippie Pancakes
Socially conscious breakfast food.
Recipe by: David Horton
Yield: 12 to 16 small pancakes, enough for two hippies
Prep time: 10 minutes

Ingredients
1 C. unbleached wheat blend flour
2 tsp. baking powder
1 tsp. unrefined sugar
1/4 tsp. coarse kosher salt
1 free-range egg
1 1/4 C. hormone-free milk
1 tsp. organic vegetable oil

Preparation Instructions
Pre-heat griddle over medium heat. Combine dry ingredients in a mixing bowl. Stir in egg, milk and oil. Use a large spoon or gravy ladle to transfer pancake batter to the griddle. Pancakes are ready to flip when large bubbles can be seen on top.

Serving Instructions
Top with fruit and berries or serve with traditional maple syrup.

Less processability
**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>
<thead>
<tr>
<th>Car Type</th>
<th>Engine (cm³)</th>
<th>Fuel Type</th>
<th>CO2 (g/km)</th>
<th>Tax (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>900</td>
<td>Gasoline</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>Type 2</td>
<td>1.100</td>
<td>Gasoline</td>
<td>120</td>
<td>5</td>
</tr>
<tr>
<td>Type 3</td>
<td>1.300</td>
<td>Gasoline</td>
<td>125</td>
<td>5</td>
</tr>
<tr>
<td>Type 4</td>
<td>1.400</td>
<td>Gasoline</td>
<td>150</td>
<td>5</td>
</tr>
<tr>
<td>Type 5</td>
<td>1.800</td>
<td>Diesel</td>
<td>180</td>
<td>10</td>
</tr>
<tr>
<td>Type 6</td>
<td>2.200</td>
<td>Diesel</td>
<td>190</td>
<td>10</td>
</tr>
<tr>
<td>Type 7</td>
<td>2.500</td>
<td>Gasoline</td>
<td>210</td>
<td>15</td>
</tr>
</tbody>
</table>

Less relevance

<table>
<thead>
<tr>
<th>Car Type</th>
<th>Engine (cm³)</th>
<th>Fuel Type</th>
<th>CO2 (g/km)</th>
<th>Color</th>
<th>Tax (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>900</td>
<td>Gasoline</td>
<td>90</td>
<td>Red</td>
<td>0</td>
</tr>
<tr>
<td>Type 2</td>
<td>1.100</td>
<td>Gasoline</td>
<td>120</td>
<td>Silver</td>
<td>5</td>
</tr>
<tr>
<td>Type 3</td>
<td>1.300</td>
<td>Gasoline</td>
<td>125</td>
<td>Black</td>
<td>5</td>
</tr>
<tr>
<td>Type 4</td>
<td>1.400</td>
<td>Gasoline</td>
<td>150</td>
<td>White</td>
<td>5</td>
</tr>
<tr>
<td>Type 5</td>
<td>1.800</td>
<td>Diesel</td>
<td>180</td>
<td>Silver</td>
<td>10</td>
</tr>
<tr>
<td>Type 6</td>
<td>2.200</td>
<td>Diesel</td>
<td>190</td>
<td>Blue</td>
<td>10</td>
</tr>
<tr>
<td>Type 7</td>
<td>2.500</td>
<td>Gasoline</td>
<td>210</td>
<td>Black</td>
<td>15</td>
</tr>
</tbody>
</table>
**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

Less timeliness
Best practices

Best practices for publishing high-quality data and metadata.
### W3C: Best practices for publishing Linked Open Data

<table>
<thead>
<tr>
<th><strong>IDENTIFY</strong></th>
<th>Indentify data sets that other people may wish to re-use.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MODEL</strong></td>
<td>Model the data in an application-independent, objective way in terms of representation. Denormalize the data as necessary.</td>
</tr>
<tr>
<td><strong>METADATA</strong></td>
<td>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).</td>
</tr>
<tr>
<td><strong>PI</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>NAME</strong></td>
<td>Use HTTP URLs 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.</td>
</tr>
<tr>
<td><strong>STANDARD VOCABULARIES</strong></td>
<td>Describe objects with standard vocabularies whenever possible.</td>
</tr>
<tr>
<td><strong>VOCABULARY USE</strong></td>
<td>Use vocabularies as loosely coupled modular components.</td>
</tr>
<tr>
<td><strong>REPRESENTATION</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>HUMAN READABLE</strong></td>
<td>Provide human readable descriptions with your Linked Data.</td>
</tr>
<tr>
<td><strong>MACHINE ACCESSIBLE</strong></td>
<td>Provide access to the data representation via RESTful API, SPARQL endpoint(s) and RDF download.</td>
</tr>
<tr>
<td><strong>SPECIFY LICENSE</strong></td>
<td>Specify an appropriate license.</td>
</tr>
<tr>
<td><strong>HOST</strong></td>
<td>Deliver open government data on authoritative domain to increase perceived trust.</td>
</tr>
<tr>
<td><strong>ANNOUNCE</strong></td>
<td>Announce open government data, have a feedback mechanism and be prepared to be responsive to feedback.</td>
</tr>
<tr>
<td><strong>SOCIAL CONTRACT</strong></td>
<td>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 undesirable. URI strategy and implementation are critical.</td>
</tr>
</tbody>
</table>

**See also:** [http://www.slideshare.net/OpenDataSupport/the-linked-open-government-data-lifecycle](http://www.slideshare.net/OpenDataSupport/the-linked-open-government-data-lifecycle)
### 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    | 2 | 49 | Usage rights are provided for an unlimited period |
| License    | 3 | 51 | The datasets are accompanied by a summary and a link to the full version of the license |
| Linked data| 2 | 52 | Any vocabularies used within the dataset are identified and documented |
| Linked data| 3 | 53 | Data adheres to the defined syntax of any specified vocabularies |
| Linked data| 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](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](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

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

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

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

Take also the online test [here](http://www.visualpharm.com)!
Thank you!

...and now YOUR questions?
References

Slide 5:

Slide 6:

Slide 8-26:

Slide 13:

Slide 14:

Slide 28:

Slide 29:
Further reading


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

Eurostat. European Statistical System
Be part of our team...

Find us on

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

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

Follow us

@OpenDataSupport

Join us on

http://www.opendatasupport.eu

Contact us

contact@opendatasupport.eu