Open Data Quality Measurement Framework: Definition and Application to Open Government Data

Antonio Vetrò – Technische Universität München
Marco Torchiano – Politecnico di Torino
Lorenzo Canova, Raimondo Iemma, Federico Morando
– Nexa Center for Internet and Society at Politecnico di Torino

Motivation

OGD enable active data-driven citizenship and development of services to support it and to ease citizens' life
Given a released OG dataset,

- are DD-citizenship and service development possible?
- To what extent?
- What effort do they demand?

A-priori answers require a domain-independent, automatic, easy-to-use quality assessment framework.

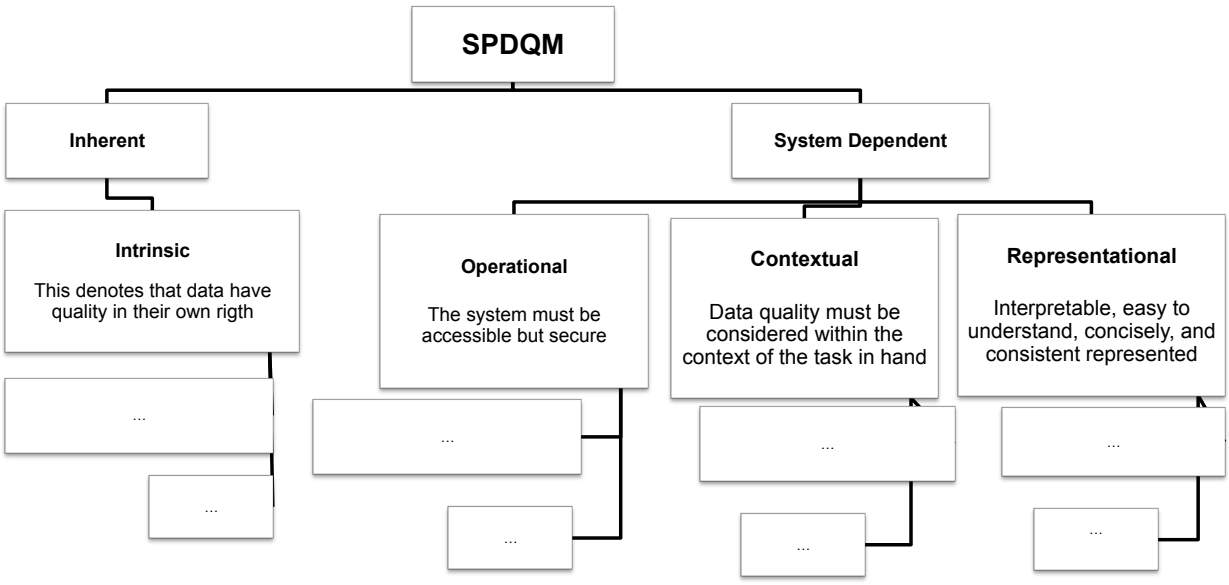
Goal:

Set up and assess a framework of indicators to measure the quality of Open Government Data on a series of quality dimensions

Method:

- Build an evaluation framework derived from a data quality model proposed in the literature,
- Define a set of metrics for a subset of quality characteristics, and
- Apply the framework to a sample of Italian OGD that adopted two distinct disclosure approaches:
  - orchestrated (national-level w/ aggregation of data from regions)
  - decentralized (administrative municipalities level, in different regions)

Quality model used as support: SPDQM



Carmen Moraga, Maria Ángeles Moraga, Coral Calero, Angelica Caro, SQuaRE-Aligned Data Quality Model for Web Portals. QSIC 2009: 117-122

SPDQM

PoV	Category	
Inherent	Intrinsic	Data have quality in their own rigth
System dependent	Operational	The system must be accessible but secure
	Contextual	Data quality must be considered within the context of the task in hand
	Representational	Interpretable, easy to understand, concisely, and consistent represented

Carmen Moraga, Maria Ángeles Moraga, Coral Calero, Angelica Caro, SQuaRE-Aligned Data Quality Model for Web Portals. QSIC 2009: 117-122

Quality characteristics

Characteristic	Metric
Traceability	Track of creation
	Track of updates
Currentness	Percentage of current rows
	Delay in publication
Expiration	Delay after expiration
Completeness	Percentage of complete cells
	Percentage of complete rows
Compliance	Percentage of standardized columns
	eGMS Compliance
	Five star Open Data
Understandability	Percentage of columns with metadata
	Percentage of columns in comprehensible format
Accuracy	Percentage of accurate cells
	Accuracy in aggregation

Quality characteristics selected and metrics

Characteristic	Metric	Level	Description
Traceability	Track of creation	Dataset	Indicates the presence or absence of metadata associated with the process of creation of a dataset.
	Track of updates	Dataset	Indicates the existence or absence of metadata associated with the updates done to a dataset.
Currentness	Percentage of current rows	Cell	Indicates the percentage of rows of a dataset that have current values, it means that they don't have any value that refers to a previous or a following period of time.
	Delay in publication	Dataset	Indicates the ratio between the delay in the publication (number of days passed between the moment in which the information is available and the publication of the dataset) and the period of time referred by the dataset (week, month, year).
Expiration	Delay after expiration	Dataset	Indicates the ratio between the delay in the publication of a dataset after the expiration of its previous version and the period of time referred by the dataset (week, month, year).
Completeness	Percentage of complete cells	Cell	Indicates the percentage of complete cells in a dataset. It means the cells that are not empty and have a meaningful value assigned (i.e. a value coherent with the domain of the column).
	Percentage of complete rows	Cell	Indicates the percentage of complete rows in a dataset. It means the rows that don't have any incomplete cell.
Compliance	Percentage of standardized columns	Cell	Indicates the percentage of standardized columns in a dataset. It just considers the columns that represent some kind of information that has standards associated with it (i.e. Geographic information).
	eGMS Compliance	Dataset	Indicates the degree to which a dataset follows the e-GMS standard (as far as the basic elements are concerned, it essentially boils down to a specification of which Dublin Core metadata should be supplied)
	Five star Open Data	Dataset	Indicates the level of the 5 Star Open Data model in which the dataset is and the advantage offered by this reason.
Understandability	Percentage of columns with metadata	Cell	Indicates the percentage of columns in a dataset that have associated descriptive metadata. This metadata is important because it allows to easily understanding the information of the dataset and the way in which it is represented.
	Percentage of columns in comprehensible format	Cell	Indicates the percentage of columns in a dataset that are represented in a format that can be easily understood by the users and it is also machine-readable.
Accuracy	Percentage of accurate cells	Cell	Indicates the percentage cells in a dataset that have correct values according to the domain and the type of information of the dataset.
	Accuracy in aggregation	Cell	Indicates the ratio between the error in aggregation and the scale of data representation. This metric only applies for the datasets that have aggregation columns or when there are two or more datasets referring to the same information but in a different granularity level.

Example of Metrics

Characteristic	Metric	Variables	Formula	Scale	Normalization
Traceability	Track of creation	s: Source dc: Date of creation	$tc = 2s + dc$	[0, 3]	$tcn = \frac{tc}{3}$
	Track of updates	lu: List of updates du: Dates of updates	$tu = lu + du$	[0, 2]	$tun = \frac{tu}{2}$
Expiration	Delay after expiration	ed: Expiration date cd: Current date sd: Start date of the period of time referred by the dataset ed: End date of the period of time referred by the dataset.	$dae = 1 - \left(\frac{cd - ed}{ed - sd}\right)$	(-∞, +∞)	$if\,(dae \leq 0)$ $daen = 0$ $else\ if\,(dae \leq 1)$ $daen = rs$ $else\ if\,(dae > 1)$ $daen = 1$
Completeness	Percentage of complete cells	nr: Number of rows nc: Number of columns ic: Number of incomplete cells ncl: Number of cells	$ncl = nr * nc$ $pcc = \left(1 - \frac{ic}{ncl}\right) * 100$	[0%, 100%]	$pccn = \frac{pcc}{100}$
	Percentage of complete rows	nr: Number of rows nir: Number of incomplete rows	$pcpr = \left(1 - \frac{nir}{nr}\right) * 100$	[0%, 100%]	$pcprn = \frac{pcpr}{100}$

Datasets analyzed

Orchestrated disclosure

- *Open Coesione*
- portal about the fulfilment of investments using the 2007-2013 European Cohesion funds
- 85 billion Euros are being tracked, 850K projects

Decentralized disclosure

Datasets	Torino	Roma	Milano	Firenze	Bologna
Residents	X	X	X	X	X
Marriages	X		X	X	
Business	X	X	X		

Datasets analyzed

Orchestrated disclosure

- *Open Coesione*
- portal about the fulfilment of investments using the 2007-2013 European Cohesion funds
- 85 billion Euros are being tracked, 850K projects

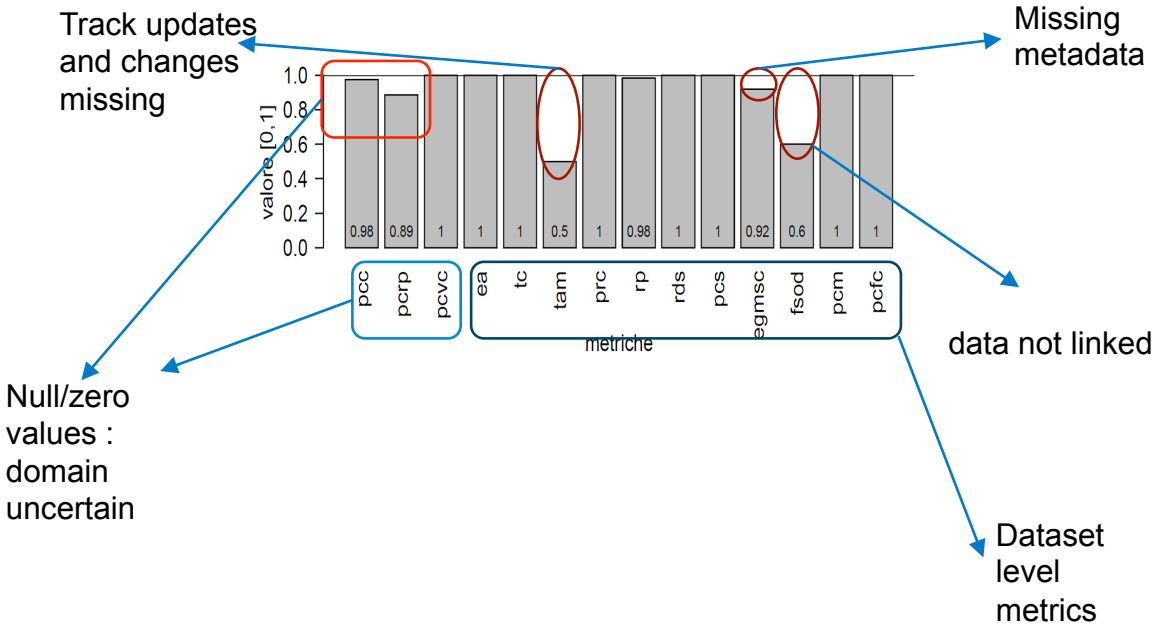
Decentralized disclosure

Dataset	Torino	Roma	Milano	Firenze	Bologna
Residents	X	X	X	X	X
Marriages	X		X	X	
Business	X	X	X		

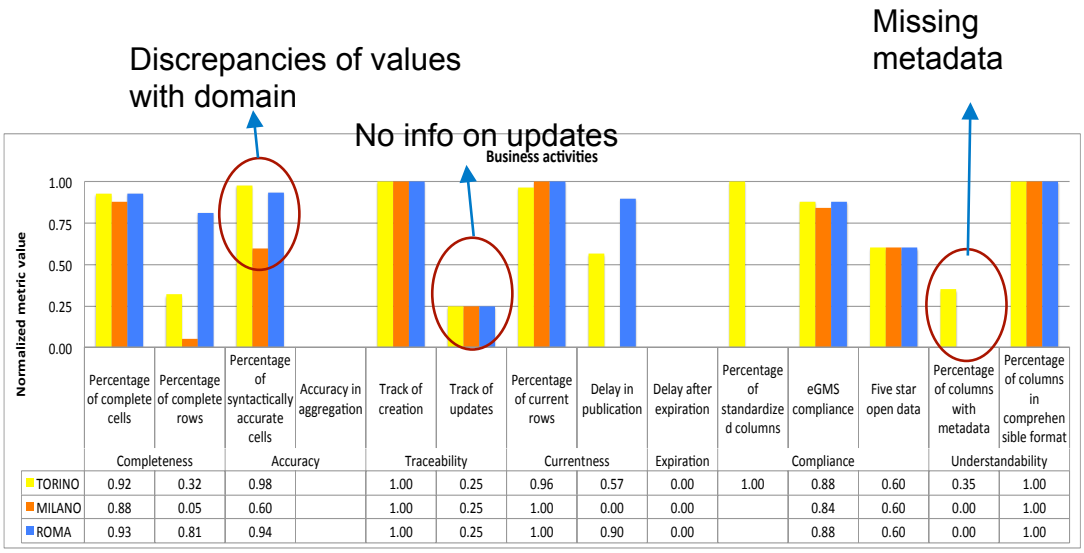
Results

Dimension	Metric	Open Coesione	Municipalities	P value	Conf. interval
Completeness	Percentage of complete cells	+	+	0.55	{-0.04; 0.04}
	Percentage of complete rows	-	+/-	**< 0.05**	{-1.00;-0.81}
Accuracy	Percentage of syntactically accurate cells	+	+/-	**< 0.05**	{-0.001; 0.04}
	Accuracy in aggregation	+	+	NaN	NaN
Traceability	Track of creation	+	+	NaN	NaN
	Track of updates	-	-	**< 0.05**	{0.25; 0.25}
Currentness	Percentage of current rows	+	+	0.20	{0 ; 0}
	Delay in publication	+	+/-	**< 0.05**	{0.08; 0.43}
Expiration	Delay after expiration	+	+/-	< 0.05	{0.99; 0.99}
Compliance	Percentage of standardized columns	+	+	NaN	NaN
	eGMS compliance	+	+	**< 0.05**	{0.04; 0.04}
	Five star Open Data	+	+	NaN	NaN
Understandability	Percentage of columns with metadata	+	-	**< 0.05**	{1.00; 1.00}
	Percentage of columns in comprehensible format	+	+/-	0.06	{0 ; 0}

Some explanations (Open Coesione)



Some explanations (municipalities data)



	Percentage of complete cells	Percentage of complete rows	Percentage of syntactically accurate cells	Accuracy in aggregation	Track of creation	Track of updates	Percentage of current rows	Delay in publication	Delay after expiration	Percentage of standardized columns	eGMS compliance	Five star open data	Percentage of columns with metadata	Percentage of columns in comprehensible format
	Completeness		Accuracy		Traceability		Currentness		Expiration	Compliance			Understandability	
TORINO	0.92	0.32	0.98		1.00	0.25	0.96	0.57	0.00	1.00	0.88	0.60	0.35	1.00
MILANO	0.88	0.05	0.60		1.00	0.25	1.00	0.00	0.00		0.84	0.60	0.00	1.00
ROMA	0.93	0.81	0.94		1.00	0.25	1.00	0.90	0.00		0.88	0.60	0.00	1.00

Open issues and Discussion points

- Are the metrics capturing the quality problems that most affect data reuse ?
- Which other quality dimensions are important?
- How to better automate metrics ?

THANK YOU !