Nexa Center for Internet & Society

Politecnico di Torino

Studying the Internet, exploring its potential & experimenting new ideas

How data quality affects reusabilty: a case study on published Italian Open Data

ODQ2015

Open Data Quality: from Theory to Practice

Munich, 30th March 2015

Open Data quality

- **Low data quality** ⇨ **low reuse potential** and **high costs** for reusing data (sometimes too high!)
  - Example - App for free parking spots in Munich: data is reusable only if up-to-date, standardized and complete
- **Causes** of low data quality:

  Data that are **high quality inside the organization** (saved in system that originally weren't made for data to be opened) are opened **without following a formalized procedure**
  - **Consequences**
    - Missing metadata, low understandability;
    - Static DB visualization causes issues regarding: coherency, accuracy (and duplications), timeliness.
    - Data is not granular enough

Open Data quality

- **Available tools** for opening data:
  - CKAN: has integrated Open Refine for checking data quality
  - SOCRATA: gives warnings on data with metadata issues

Case study on Italian government's transparency OD

- **117 analyzed Municipalities** (province capitals);
- **5 different dataset** categories for each municipality:
  - Active rentals of public buildings
  - Passive rentals
  - Real estate register
  - Beneficiaries' register
  - Public concession acts (**more regulated** than the others);
- Total of **585 analyzed dataset**

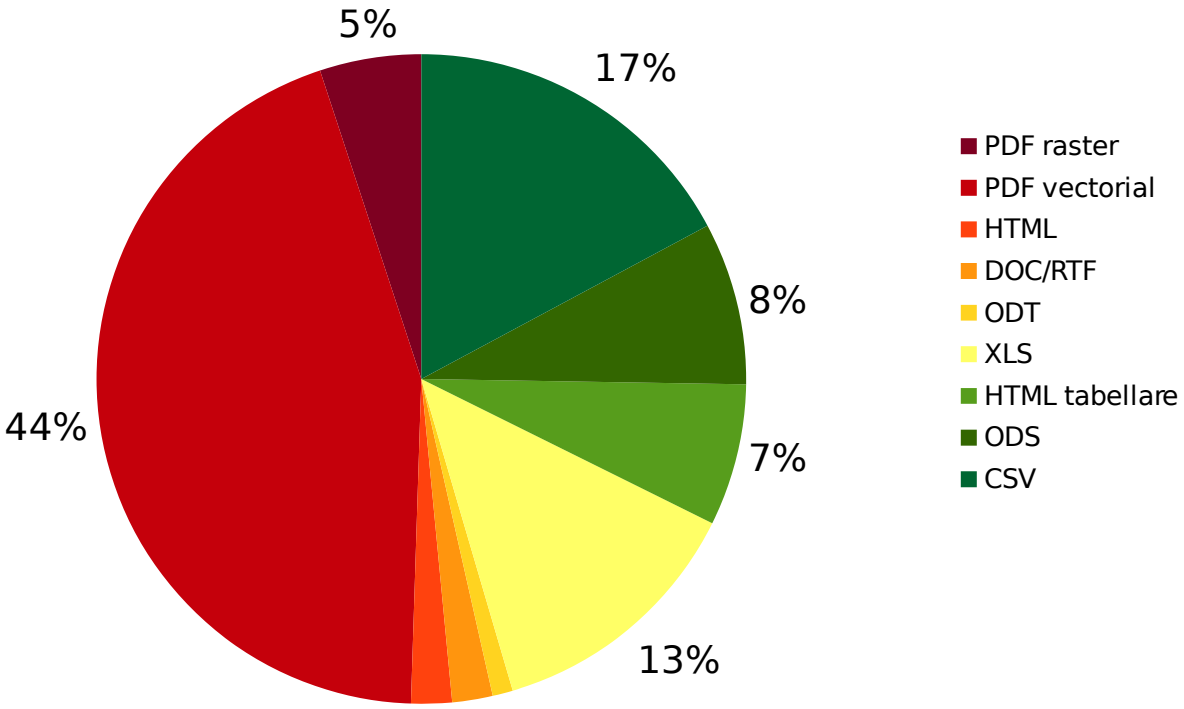
A basic Open Data Quality analysis

- has the dataset been **published**? - publication errors
- Is it **machine processable**? - file format
- Does it contain **enough information**? - Number and usefulness of published attributes – harder to measure

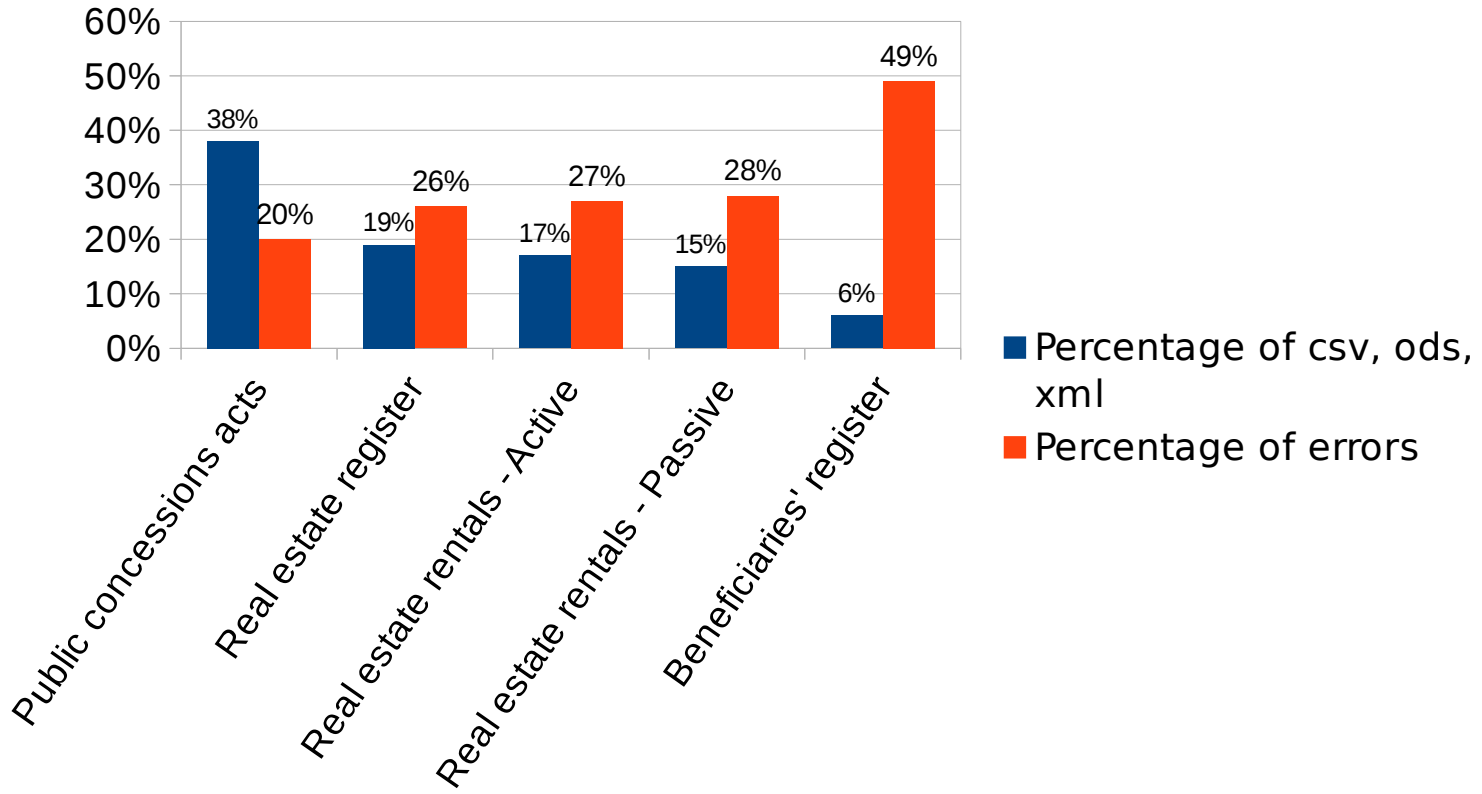
Pubblication errors

- Not published datasets = 143 = 24%
- Datasets with publication errors (es: not-tabular, too aggregate, not comprehensible) = 34
- **Total number of technically not reusable datasets = 177 = 30%**

Dataset formats

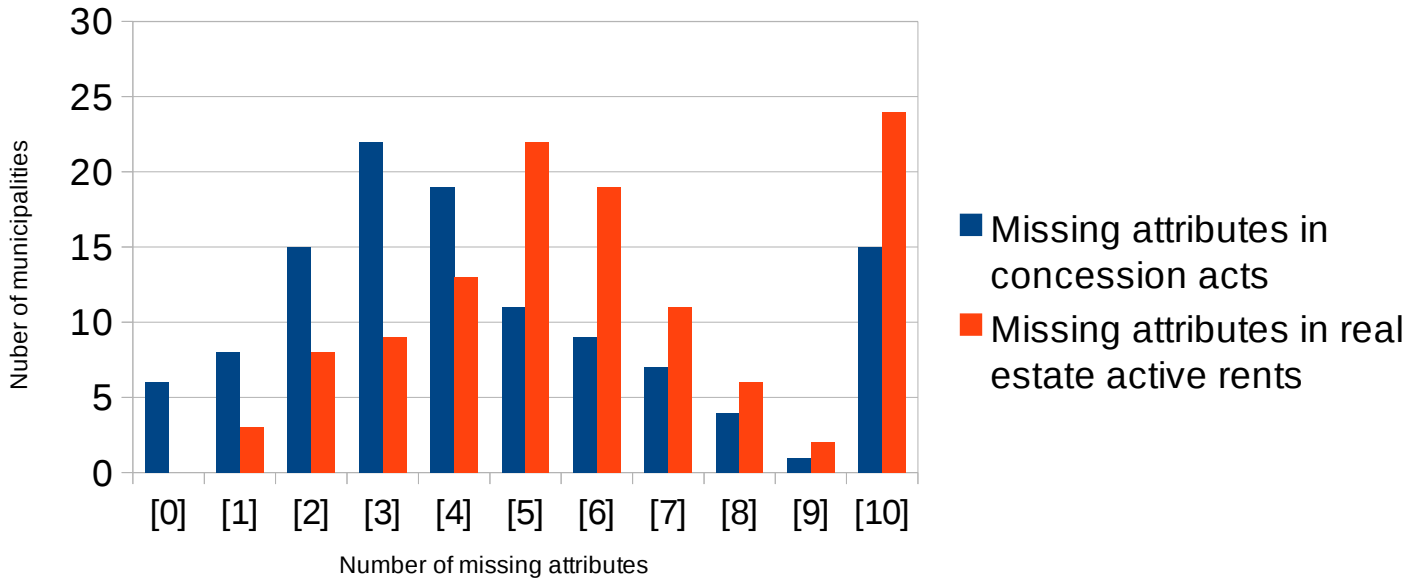


Pubblication errors and formats - Categories



Information quantity completeness

- Example: real estate active rents vs. concession acts (“guideline effect”?)



A simple reuse example

- **Objective:** calculate the mean €/mq requested by for renting a public real estate in each city
- Potentially available dataset number: 117
- Easily processable CSV files: 10
- Dataset with monthly/yearly rent and surface: 4
- **Metadata** issues – e.g.: is the represented amount annual or monthly? Does it include VAT? Are there represented volumes or surfaces?
- **Result: It was not possible to calculate** the mean €/mq requested in each city
- And we didn't mention: **timeliness, completeness, accuracy...**

Conclusions

- The public datasets analyzed are **low quality** and **hard to reuse** even for simple analyses
  - More standardization is needed in: formats, metadata, attributes
- In this case a more **prescriptive law** implied a better quality dataset
  - “guideline effect”
  - “penalty effect”

Open Issues

- Opening dataset in a **centralized fashion** for interoperability?
- Would **specific guidelines** be beneficial?
- Would it be beneficial (from the publisher side) **defining possible reuses** of datasets and **afterwards the quality standards** for publishing different datasets?

Nexa Center for Internet & Society

Politecnico di Torino

Studying the Internet, exploring its potential & experimenting new ideas

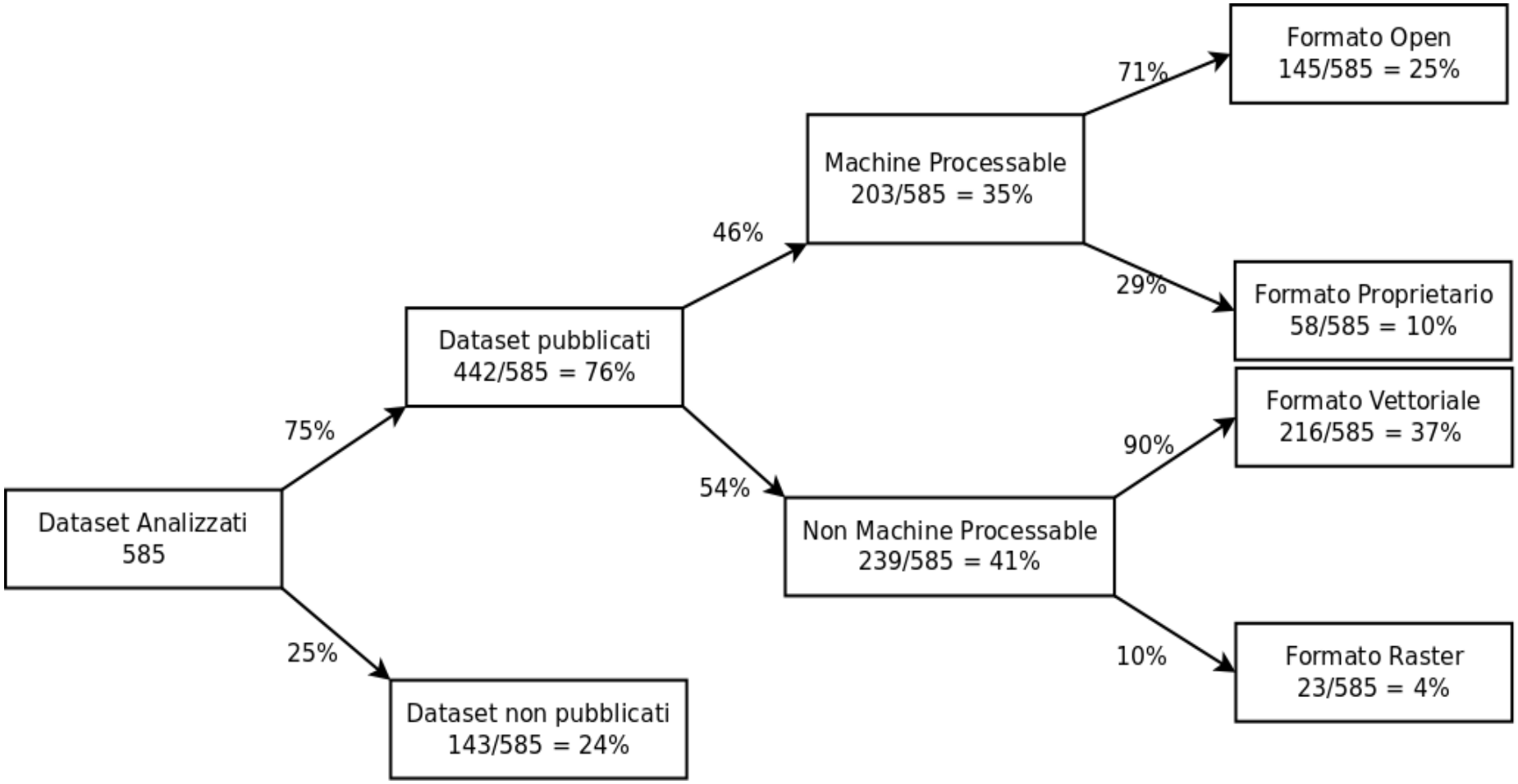
Thank you!

ODQ2015

Open Data Quality: from Theory to Practice

Munich, 30th March 2015

Formats


Dataset Analizzati
585
75%
25%
Dataset pubblicati
442/585 = 76%
Dataset non pubblicati
143/585 = 24%
46%
54%
Machine Processable
203/585 = 35%
Non Machine Processable
239/585 = 41%
71%
29%
90%
10%
Formato Open
145/585 = 25%
Formato Proprietario
58/585 = 10%
Formato Vettoriale
216/585 = 37%
Formato Raster
23/585 = 4%