How can you use Open Data? ... And why you should!

Axel Polleres

web: http://polleres.net  twitter: @AxelPolleres
What is Open Data?

**Availability and Access**: the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.

**Reuse and Redistribution**: the data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets. The data must be machine-readable.

**Universal Participation**: everyone must be able to use, reuse and redistribute – there should be no discrimination against fields of endeavour or against persons or groups. For example, ‘non-commercial’ restrictions that would prevent ‘commercial’ use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

See more at: [http://opendefinition.org/okd/](http://opendefinition.org/okd/)
Open Data vs. Big Data


BIG DATA

1. Non-public data for marketing, business analysis, national security
3. Large datasets from scientific research, social media, or other non-govt. sources
6. Large public government datasets (eg weather, GPS, Census, SEC, health care)

OPEN GOVT

2. Citizen engagement programs not based on data (eg petition websites)
4. Public data from state, local, federal gov't. (eg budget data)
5. Business reporting (eg ESG data); other business data (eg consumer complaints)

OPEN DATA
Open Data is a global trend:

- Cities, International Organizations, National and European Portals, Int'l. Conferences:
In today's talk:

- 2 projects on recent projects at WU:
- **What is the status of Open Data and what are the challenges using Open Data?**
  - OpenData PortalWatch – a project at WU
- **How can Open Data be used?**
  - Open City Data Pipeline – a joint project with Siemens on using Open Data

- **What's next?**
  - Improving Open Data Quality: ADEQUATE (FFG - project)

- **Why should you care?**
  - WU can help you in your Open Data Strategy!
Challenges:
Open Data also has the "Vs"

- **Volume:**
  - It's growing! (we currently monitor 90 CKAN portals, 512543 resources/ 160069 datasets, at the moment (statically) ~1TB only CSV files...)

- **Variety:**
  - different datasets (from different cities, countries, etc.), only partially comparable, partially not.
  - Different metadata
  - Different data formats

- **Velocity:**
  - Open Data changes regularly (fast and slow)
  - New datasets appear, old ones disappear
Periodically monitoring a list of Open Data Portals
- 90 CKAN powered Open Data Portals
Quality assessment
Evolution tracking
- Meta data
- Data

http://data.wu.ac.at/portalwatch/
Open Data Portals

CKAN ... http://ckan.org/

• almost „de facto“ standard for Open Data Portals
• facilitates search, metadata (publisher, format, publication date, license, etc.) for datasets

• http://datahub.io/
• http://data.gv.at/

• machine-processable? ...
... partially
## Open Data Portal Watch

**Brief overview of 89 Open Data CKAN portals**

<table>
<thead>
<tr>
<th>Portal</th>
<th>Country</th>
<th>Data Sets</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>annuario.comune.fi.it</td>
<td>Italy</td>
<td>358</td>
<td>1363</td>
</tr>
<tr>
<td>catalogue.datalocale.fr</td>
<td>France</td>
<td>303</td>
<td>751</td>
</tr>
<tr>
<td>dados.gov.br</td>
<td>Brazil</td>
<td>501</td>
<td>4344</td>
</tr>
<tr>
<td>data.buenosaires.gob.ar</td>
<td>Argentina</td>
<td>123</td>
<td>628</td>
</tr>
<tr>
<td>data.edostate.gov.ng</td>
<td>Nigeria</td>
<td>164</td>
<td>207</td>
</tr>
<tr>
<td>data.glasgow.gov.uk</td>
<td>United Kingdom</td>
<td>384</td>
<td>1943</td>
</tr>
<tr>
<td>datagm.org.uk</td>
<td>United Kingdom</td>
<td>360</td>
<td>506</td>
</tr>
<tr>
<td>data.gov.sk</td>
<td>Slovakia</td>
<td>216</td>
<td>556</td>
</tr>
<tr>
<td>ckan.data.graz.gv.at</td>
<td>Austria</td>
<td>151</td>
<td>341</td>
</tr>
<tr>
<td>data.kk.dk</td>
<td>Denmark</td>
<td>102</td>
<td>346</td>
</tr>
<tr>
<td>data.lexingtonky.gov</td>
<td>government</td>
<td>93</td>
<td>186</td>
</tr>
<tr>
<td>data.nsw.gov.au</td>
<td>Australia</td>
<td>311</td>
<td>456</td>
</tr>
<tr>
<td>data.ohouston.org</td>
<td>non-commercial</td>
<td>227</td>
<td>361</td>
</tr>
<tr>
<td>data.ottawa.ca</td>
<td>Canada</td>
<td>119</td>
<td>493</td>
</tr>
<tr>
<td>data.cityofsantacruz.com</td>
<td>commercial</td>
<td>52</td>
<td>72</td>
</tr>
<tr>
<td>dados.recife.pe.gov.br</td>
<td>Brazil</td>
<td>43</td>
<td>318</td>
</tr>
</tbody>
</table>
# Quality Dimensions

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrievability</td>
<td>The extent to which meta data and resources can be retrieved.</td>
</tr>
<tr>
<td>Usage</td>
<td>The extent to which available meta data keys are used to describe a dataset.</td>
</tr>
<tr>
<td>Completeness</td>
<td>The extent to which the used meta data keys are non empty.</td>
</tr>
<tr>
<td>Accuracy</td>
<td>The extent to which certain meta data values accurately describe the resources.</td>
</tr>
<tr>
<td>Openness</td>
<td>The extent to which licenses and file formats conform to the open definition.</td>
</tr>
<tr>
<td>Contactability</td>
<td>The extent to which the data publisher provide contact information.</td>
</tr>
</tbody>
</table>

Objective measures which can be automatically computed in a scalable way.
Portal Overview

Open Data Portal Watch

Portal: GovData | Datenportal für Deutschland - GovData

Available Snapshots
- WS 2014
- WS 2015
- W4 2016
- W4 2015
- W5 2015
- W6 2015
- W7 2015
- W8 2015

Snapshot: Sun Feb 22 2015 23:52:47 GMT+0100 (CET)

Quality
- Q1
- Q2
- Q3
- Q4
- Q5

Size
- Datasets: 13195
- Resources: 37256

Openness
- License: AVG. 0.17
- Format: AVG. 0.94

Retrievability
- Datasets: AVG. 1.00
- Resources: AVG. 0.79

Contactability
- Email: AVG. 0.92
- URL: AVG. 0.00
ODP Evolution

Open Data Portal Watch

Portal: GovData | Datenportal für Deutschland - GovData

Available snapshots

Fri Oct 10 2014 14:51:16 GMT+0200 (CEST)

Evolution of quality measures

Evolution of quality measures
## Changes between the first and last snapshots

### Dataset Changes

**70 Portals with Dataset Changes**

- Avg. increase by 87.05% for 60 portals
- Avg. decrease by -64.16% for 10 portals

<table>
<thead>
<tr>
<th>Portal</th>
<th>From</th>
<th>To</th>
<th>Change</th>
<th>Change Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>data.sa.gov.au</td>
<td>484</td>
<td>5721</td>
<td>5237</td>
<td>1082.02%</td>
</tr>
<tr>
<td>datos.codeandomexico.org</td>
<td>94</td>
<td>715</td>
<td>621</td>
<td>660.64%</td>
</tr>
<tr>
<td>data.opendataportal.at</td>
<td>46</td>
<td>323</td>
<td>277</td>
<td>602.17%</td>
</tr>
<tr>
<td>annuario.comune.fi.it</td>
<td>50</td>
<td>351</td>
<td>301</td>
<td>602.00%</td>
</tr>
<tr>
<td>(2014-08-07) – (2015-03-15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u dct-data.aigid.jp</td>
<td>431</td>
<td>2110</td>
<td>1679</td>
<td>389.56%</td>
</tr>
<tr>
<td>(2014-08-07) – (2015-03-16)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>catalogo.datos.gob.mx</td>
<td>111</td>
<td>360</td>
<td>249</td>
<td>224.32%</td>
</tr>
<tr>
<td>(2014-08-08) – (2015-03-15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Data Dumps**

- OPEN DATA PORTAL WATCH provides an archive of Open Data portal crawls (weekly snapshots/dynamic crawling framework):

### Open Data Portal Watch Dumps

<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>africaopendata.org/</td>
<td>16-Mar-2015 13:03</td>
<td>-</td>
</tr>
<tr>
<td>annuario.comune.fi.it/</td>
<td>16-Mar-2015 13:03</td>
<td>-</td>
</tr>
<tr>
<td>bermuda.io/</td>
<td>16-Mar-2015 13:14</td>
<td>-</td>
</tr>
<tr>
<td>catalog.data.gov/</td>
<td>05-Feb-2015 15:28</td>
<td>-</td>
</tr>
<tr>
<td>catalog.data.ug/</td>
<td>16-Mar-2015 13:07</td>
<td>-</td>
</tr>
<tr>
<td>catalogo.datos.gob.mx/</td>
<td>16-Mar-2015 13:08</td>
<td>-</td>
</tr>
<tr>
<td>catalogodatos.gub.uy/</td>
<td>16-Mar-2015 13:15</td>
<td>-</td>
</tr>
</tbody>
</table>

### Open Data Portal Watch Dumps

<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>2014-07-17.gz</td>
<td>05-Feb-2015 15:13</td>
<td>2.2M</td>
</tr>
<tr>
<td>2014-07-25.gz</td>
<td>05-Feb-2015 15:13</td>
<td>2.2M</td>
</tr>
<tr>
<td>2014-08-05.gz</td>
<td>05-Feb-2015 15:13</td>
<td>2.2M</td>
</tr>
<tr>
<td>2014-08-12.gz</td>
<td>05-Feb-2015 15:13</td>
<td>2.2M</td>
</tr>
<tr>
<td>2014-08-27.gz</td>
<td>05-Feb-2015 15:13</td>
<td>2.2M</td>
</tr>
<tr>
<td>2014-09-01.gz</td>
<td>05-Feb-2015 15:14</td>
<td>2.2M</td>
</tr>
<tr>
<td>2014-09-07.gz</td>
<td>05-Feb-2015 15:14</td>
<td>2.2M</td>
</tr>
<tr>
<td>2014-09-14.gz</td>
<td>05-Feb-2015 15:14</td>
<td>2.2M</td>
</tr>
</tbody>
</table>
Open Data Portal Watch

Towards assessing the quality evolution of Open Data portals

Jürgen Umbrich, Sebastian Neumaier, Axel Polleres
Vienna University of Economics and Business, Vienna, Austria

In this work, we present the Open Data Portal Watch project, a public framework to continuously monitor and assess the (meta-)data quality in Open Data portals. We critically discuss the objectiveness of various quality metrics. Further, we report on early findings based on 22 weekly snapshots of 90 CKAN portals and highlight interesting observations and challenges.

http://data.wu.ac.at/portalwatch/

- Key findings:
  - Varying quality across portals
  - Rapid growth for some portals
  - Huge variety and range of datasets
But: How to use all that Open Data?

- More challenges:
  - How to find the right datasets?
  - How to integrate related datasets?
  - How to deal with heterogeneous/missing data
Use Case: City Data – Important for Infrastructure Providers & for City Decision Makers

• City Assessment and Sustainability reports

• Tailored offerings by Infrastructure Providers

→ Needs up-to-date City Data and calculates City KPIs in a way that allows to display the current state and run scenarios of different product applications.

e.g. towards a “Dynamic” Green City Index:

... however, these are often outdated before even published!

Goal (short term):
- Leverage Open Data for calculating a city’s performance from public sources on the Web automatically

Goal (long term):
- Define and Refine KPI models to assess specific impact of infrastructural investments and gather/check input automatically
Open City Data Pipeline

We present the City Data Pipeline – a system for gathering city performance indicators published as Open Data in order to ease the compilation of studies and reports used within Siemens. Under the assumption that Open Data provides means to automatise tedious data research tasks, we have built a system that integrates basic indicators for cities from various Open Data sources. The architecture is flexible, extensible, and natively based on RDF & SPARQL.
City Data Pipeline: Architecture

Crawl and integrate Open data sources with city data: Wikipedia, Eurostat, UNData, USCencus, Etc.

Use statistical methods to approximate missing values

Re-publish as Linked Data (SPARQL endpoint, provenance information, etc.)
Found a large amount of missing values

Two Reasons:
- Incomplete data published by providers (Tables 1+2)
- The combination of different data sets with disjoint cities and indicators (later)

Table 1: Urban Audit Data Set

<table>
<thead>
<tr>
<th>Year(s)</th>
<th>Cities</th>
<th>Indicators</th>
<th>Filled</th>
<th>Missing</th>
<th>% of Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>177</td>
<td>121</td>
<td>2 480</td>
<td>18 937</td>
<td>88.4</td>
</tr>
<tr>
<td>2000</td>
<td>477</td>
<td>156</td>
<td>10 347</td>
<td>64 065</td>
<td>85.0</td>
</tr>
<tr>
<td>2005</td>
<td>651</td>
<td>167</td>
<td>23 494</td>
<td>85 223</td>
<td>78.4</td>
</tr>
<tr>
<td>2010</td>
<td>905</td>
<td>202</td>
<td>90 490</td>
<td>92 320</td>
<td>50.5</td>
</tr>
<tr>
<td>2004 - 2012</td>
<td>943</td>
<td>215</td>
<td>531 146</td>
<td>1 293 559</td>
<td>70.9</td>
</tr>
<tr>
<td>All (1990 - 2012)</td>
<td>943</td>
<td>215</td>
<td>638 934</td>
<td>4 024 201</td>
<td>86.3</td>
</tr>
</tbody>
</table>

Table 2: United Nations Data Set

<table>
<thead>
<tr>
<th>Year(s)</th>
<th>Cities</th>
<th>Indicators</th>
<th>Filled</th>
<th>Missing</th>
<th>% of Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>11</td>
<td>52.4</td>
</tr>
<tr>
<td>2000</td>
<td>1 391</td>
<td>147</td>
<td>7 492</td>
<td>196 985</td>
<td>96.3</td>
</tr>
<tr>
<td>2005</td>
<td>1 048</td>
<td>142</td>
<td>3 654</td>
<td>145 162</td>
<td>97.5</td>
</tr>
<tr>
<td>2010</td>
<td>2 008</td>
<td>151</td>
<td>10 681</td>
<td>292 527</td>
<td>96.5</td>
</tr>
<tr>
<td>2004 - 2012</td>
<td>2 733</td>
<td>154</td>
<td>44 944</td>
<td>3 322 112</td>
<td>98.7</td>
</tr>
<tr>
<td>All (1990 - 2012)</td>
<td>4 319</td>
<td>154</td>
<td>69 772</td>
<td>14 563 000</td>
<td>99.5</td>
</tr>
</tbody>
</table>
Challenges – Missing values

- Individual datasets (e.g. from Eurostat) have missing values
- **Merging together datasets** with different indicators/cities adds sparsity

### Data from Source 1

<table>
<thead>
<tr>
<th></th>
<th>Vienna</th>
<th>Augsburg</th>
<th>Valletta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cars</td>
<td>655806</td>
<td>111561</td>
<td>95858</td>
</tr>
<tr>
<td>Nationals</td>
<td>1342704</td>
<td>216289</td>
<td>203657</td>
</tr>
<tr>
<td>Women per 1000 Men</td>
<td>109.8</td>
<td>108.7</td>
<td>101.9</td>
</tr>
</tbody>
</table>

### Data from Source 2

<table>
<thead>
<tr>
<th></th>
<th>Marbella</th>
<th>Stockholm</th>
<th>Funchal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available Beds per 1000</td>
<td>138.3</td>
<td>14969</td>
<td>166.1</td>
</tr>
<tr>
<td>Average area of living</td>
<td>36.42</td>
<td>37.24</td>
<td>38.16</td>
</tr>
<tr>
<td>Cinema Seats</td>
<td>4691</td>
<td>12751</td>
<td>2676</td>
</tr>
</tbody>
</table>

### Combined data from Source 1 and Source 2
Missing Values – Hybrid approach
choose best prediction method per indicator:

- Our assumption: every indicator has its own distribution and relationship to others.
- Basket of "standard" regression methods:
  - K-Nearest Neighbour Regression (KNN)
  - Multiple Linear Regression (MLR)
  - Random Forest Decision Trees (RFD)
Instead of using indicators directly we use PCs, built from the indicators.

For building the PCs, fill in missing data points with neutral values → predict all rows.
City Data Pipeline

citydata.wu.ac.at

- Search for indicators & cities
- Obtain results incl. sources
- Integrated data served as Linked Data
- Predicted values for missing data...

Vienna

Municipal waste (1000 t)

- 2004: 778,905392176222 1000 t (from http://citydata.wu.ac.at/ns#Prediction, predicted by an estimated error of %RMSE)
- 2005: 813,77643147163 1000 t (from http://citydata.wu.ac.at/ns#Prediction, predicted by an estimated error of %RMSE)
- 2006: 813,88824195497 1000 t (from http://citydata.wu.ac.at/ns#Prediction, predicted by an estimated error of %RMSE)
- 2007: 811,538914636665 1000 t (from http://citydata.wu.ac.at/ns#Prediction, predicted by an estimated error of %RMSE)
- 2008: 811,010344391444 1000 t (from http://citydata.wu.ac.at/ns#Prediction, predicted by an estimated error of %RMSE)

...assumption: Predictions get better, the more Open data we integrate...

Open Data: The more, the merrier!
Stefan Bischof, Christoph Martin, Axel Polleres, and Patrik Schneider. Open City Data Pipeline: Collecting, Integrating, and Predicting Open City Data. In 4th Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD), co-located with ESWC2015, Portoroz, Slovenia, May 2015.

**Open City Data Pipeline**  
**Collecting, Integrating, and Predicting Open City Data**

Stefan Bischof¹,², Christoph Martin², Axel Polleres², and Patrik Schneider²,³  
¹ Siemens AG Österreich, Vienna, Austria  
² Vienna University of Economics and Business, Vienna, Austria  
³ Vienna University of Technology, Vienna, Austria

**Abstract.** Having access to high quality and recent data is crucial both for decision makers in cities as well as for informing the public, likewise, infrastructure providers could offer more tailored solutions to cities based on such data. However, even though there are many data sets containing relevant indicators about cities available as open data, it is cumbersome to integrate and analyze them, since the collection is still a manual process and the sources are not connected to each other upfront. Further, disjoint indicators and cities across the available data sources lead to a large proportion of missing values when integrating these sources. In this paper we present a platform for collecting, integrating, and enriching open data about cities in a re-usable and comparable manner: we have integrated various open data sources and present approaches for predicting missing values, where we use standard regression methods in combination with principal component analysis to improve quality and amount of predicted values. Further, we re-publish the integrated and predicted values as linked open data.
What's next? Collaborations to make Open Data usage more effective:

- Improving Open Data Quality

Upcoming:

**ADEQUATe: Analytics & Data Enrichment to improve the QUALiTy of Open Data**
Project Start: Fall 2015
Open your data & include Open Data in your Data Strategy!

- A "sister" portal for [http://data.gv.at](http://data.gv.at) for non-governmental open data launched in 2014
  [http://www.opendataportal.at/](http://www.opendataportal.at/)

- We can help you to use and publish Open Data!
  WU, TU, SWC, DUK have **just founded** a network node of the [Open Data Institute](http://www.opendatainstitute.org/)

Official Launch:
4. OGD D-A-CH-LI - Konferenz - Open X
24. Juni 2015, Wiener Rathaus
Want to learn more?

- → Talk to me about **Your Open Data Strategy**!
  
  **Axel.Polleres@wu.ac.at**  **Twitter:**  **@AxelPolleres**

![InfoBiz Logo]

http://wu.ac.at/infobiz/

- Maybe see you at one of the following events:

  - [European Data Forum 2015](http://2015.data-forum.eu/)
  - [SEMANTiCS](http://semantics.cc/)

![European Data Forum 2015](image1)

*European Data Forum 2015*

*November 16-17, 2015, Luxembourg*

![SEMANTiCS](image2)

*SEMANTiCS*

*Vienna 2015*

*VIENNA, SEPTEMBER 15-17, 2015*