

# 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](#).

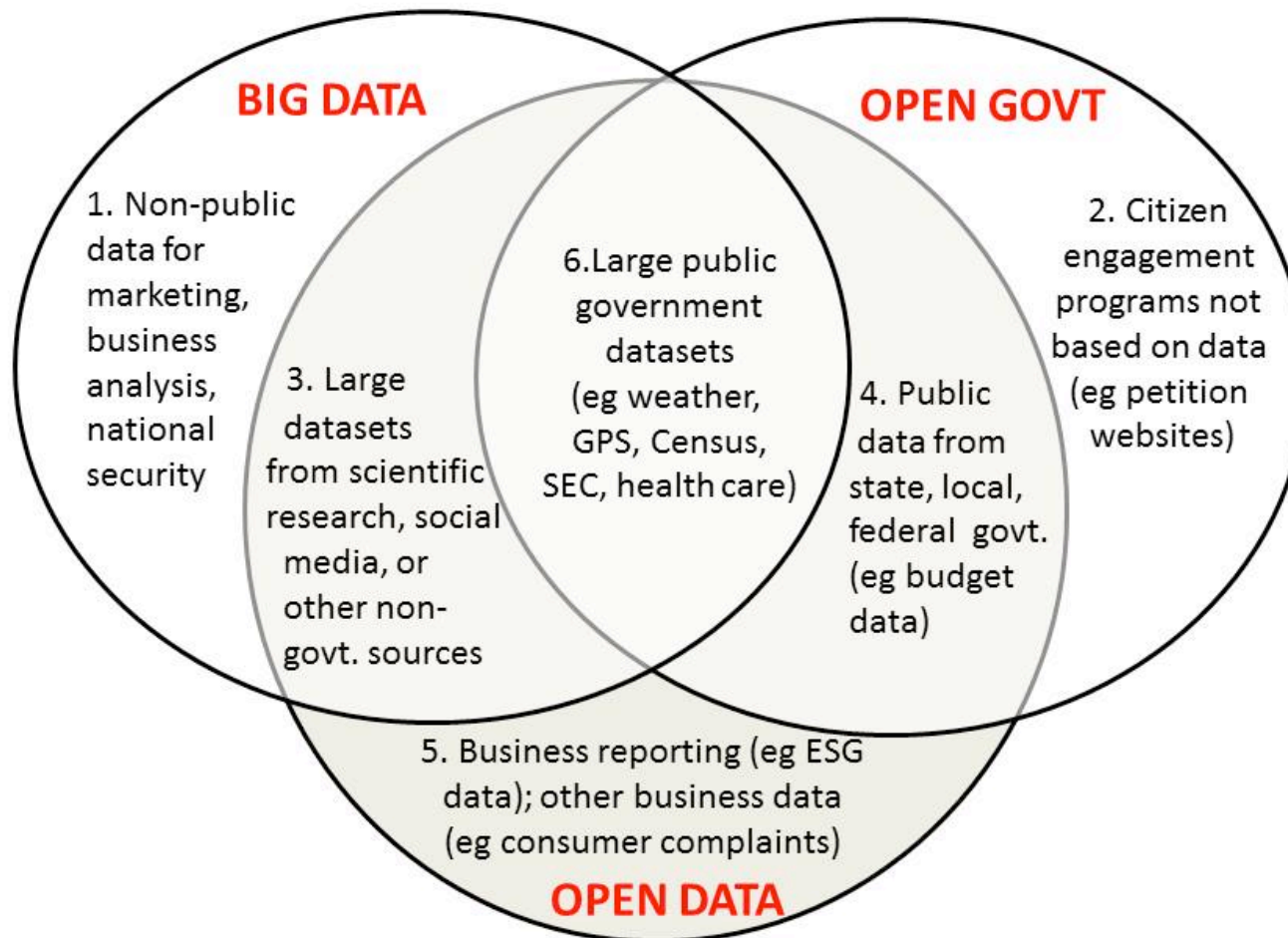
**Universal Participation:** everyone must be able to use, reuse and redistribute – there should be no discrimination against fields of endeavour or against persons or groups. For example, ‘non-commercial’ restrictions that would prevent ‘commercial’ use, or restrictions of use for certain purposes (e.g.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

See more at: <http://opendefinition.org/okd/>



# Open Data vs. Big Data

<http://www.opendatanow.com/2013/11/new-big-data-vs-open-data-mapping-it-out/>



# Open Data is a global trend:

- Cities, International Organizations, National and European Portals, Int'l. Conferences:



**European Data Forum 2015**  
 November 16-17, 2015, Luxembourg



# In today's talk:

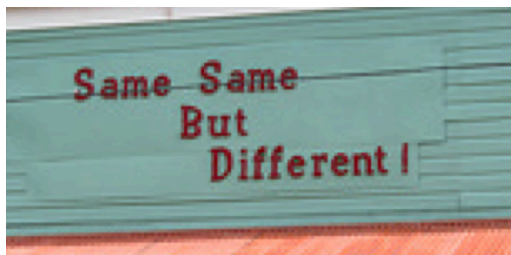
- 2 projects on recent projects at WU:
- ***What is the status of Open Data and what are the challenges using Open Data?***
  - OpenData PortalWatch – a project at WU
- ***How can Open Data be used?***
  - Open City Data Pipeline – a joint project with Siemens on using Open Data
- ***What's next?***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

# OPEN DATA PORTAL WATCH

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

- Periodically monitoring a list of Open Data Portals
  - 90 CKAN powered Open Data Portals
- Quality assessment
- Evolution tracking
  - Meta data
  - Data

The screenshot shows the top section of the 'Open Data Portal Watch' website. At the top right is the WU logo. Below it, the title 'Open Data Portal Watch' is displayed. A navigation bar contains icons for Open Data Commons, a magnifying glass, 'DOWNLOADS', a star, and an information icon. Below the navigation bar is a 'Welcome' section. The main content area features a 'Motivation' heading followed by a paragraph about the Open Data movement. Below that is the 'Open Data Portal Watch' heading, followed by a paragraph describing the project's goal to monitor and assess the quality of Open Data portals. The text mentions monitoring 90 CKAN portals and computing various metrics weekly.

# 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**

The screenshot shows the homepage of data.gv.at, the Austrian Open Data Portal. The browser address bar displays 'http://www.data.gv.at/'. The page features a search bar with the placeholder text 'Suchbegriff (z.B. Finanzen, Wahlen)' and a 'Suche starten' button. Below the search bar, there are navigation links for 'Datenkatalog', 'Anwendungen & News', and 'Katalog durchstöbern'. The main content area includes the text 'offene Daten Österreichs – lesbar für Mensch und Maschine' and 'Vielfalt, Transparenz, Offenheit, Demokratie'. It also mentions that data.gv.at provides a 'Katalog offener Datensätze und Dienste' based on Open Data principles. A diagram on the right shows a computer monitor displaying binary code, with arrows pointing to a group of people and a smartphone, illustrating the accessibility of the data for both humans and machines. The page number '1 2 3' is visible at the bottom right.



# Open Data Portal list

## Open Data Portal Watch



### Brief overview of 89 Open Data CKAN portals

Sort by [Domain](#) [Country](#) [Datasets](#) [Resources](#) Filter:  [Tile view](#) [Table View](#)

<p>annuario.comune.fi.it Italy</p> <p>358 DATASETS 1363 RESOURCES</p>	<p>catalogue.datalocale.fr France</p> <p>303 DATASETS 751 RESOURCES</p>	<p>dados.gov.br Brazil</p> <p>501 DATASETS 4344 RESOURCES</p>	<p>data.buenosaires.gob.ar Argentina</p> <p>123 DATASETS 626 RESOURCES</p>
<p>data.edostate.gov.ng Nigeria</p> <p>164 DATASETS 207 RESOURCES</p>	<p>data.glasgow.gov.uk United Kingdom (common practice)</p> <p>384 DATASETS 1943 RESOURCES</p>	<p>datagm.org.uk United Kingdom (common practice)</p> <p>360 DATASETS 506 RESOURCES</p>	<p>data.gov.sk Slovakia</p> <p>216 DATASETS 556 RESOURCES</p>
<p>ckan.data.graz.gv.at Austria</p> <p>151 DATASETS 341 RESOURCES</p>	<p>data.kk.dk Denmark</p> <p>102 DATASETS 346 RESOURCES</p>	<p>data.lexingtonky.gov government</p> <p>93 DATASETS 186 RESOURCES</p>	<p>data.nsw.gov.au Australia</p> <p>311 DATASETS 458 RESOURCES</p>
<p>data.ohouston.org non-commercial</p> <p>227 DATASETS 361 RESOURCES</p>	<p>data.ottawa.ca Canada</p> <p>119 DATASETS 493 RESOURCES</p>	<p>data.cityofsantacruz.com commercial</p> <p>52 DATASETS 72 RESOURCES</p>	<p>dados.recife.pe.gov.br Brazil</p> <p>43 DATASETS 318 RESOURCES</p>

# QUALITY DIMENSIONS

---

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

---

Objective measures which can be automatically computed in a scalable way

# Portal Overview

## Open Data Portal Watch



Portal: GovData | Datenportal für Deutschland - GovData

OVERVIEW

DETAILS

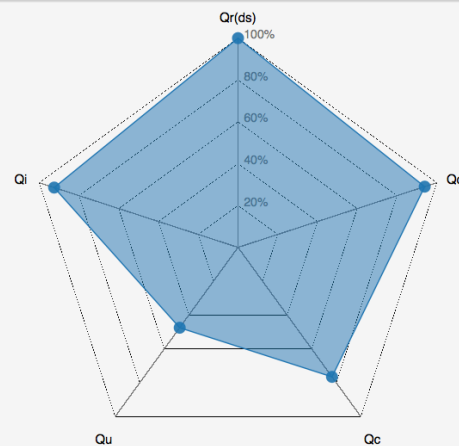
EVOLUTION

### Available Snapshots



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

### QUALITY



### 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

# Portal Details

## Available Snapshots



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

### Age

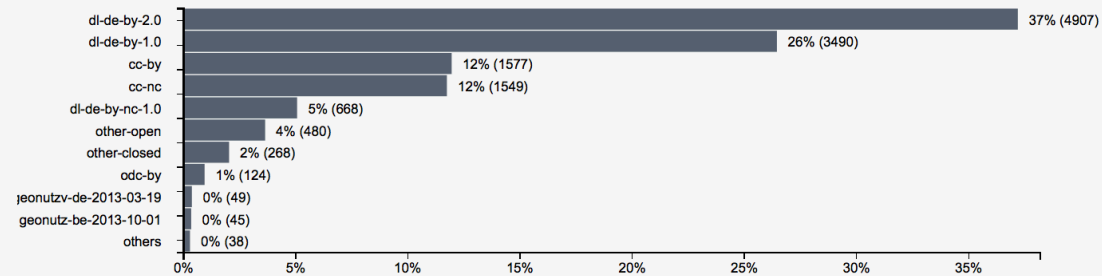
	Datasets			Resources		
	oldest	average	newest	oldest	average	newest
created:	2013-2-17	2014-7-24	2015-2-22	2012-7-9	2014-11-3	2015-2-23
modified:	2013-2-17	2014-11-22	2015-2-23	2012-3-12	2014-2-26	2015-2-23

### Retrievability

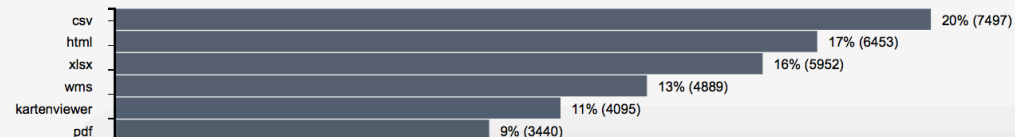
	Total	200 OK	403 Forbidden	404 Not Found	Server Timeout	Others
Datasets	13195	100%	0%	0%	0%	0%
Resources	22734	79%	0%	0%	18%	3%

### Openness

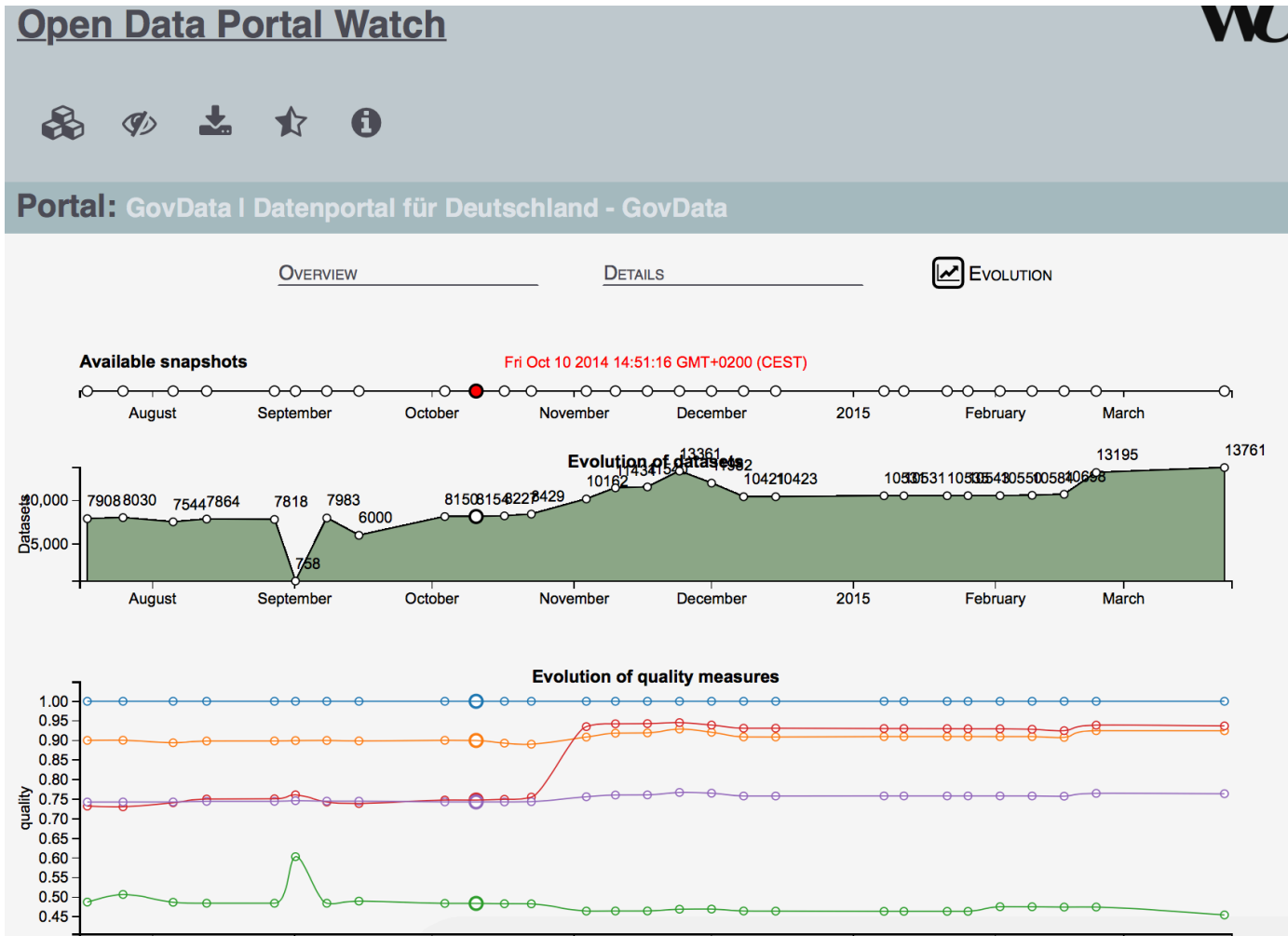
### Available licenses



### Available formats



# ODP Evolution



# ODP CHANGES

## 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

Show  entries









Search:

↑ PORTAL	↑ FROM	↑ TO	↑ CHANGE	↓ CHANGE PERCENTAGE
<b>data.sa.gov.au</b> <i>(2014-07-17) → (2015-03-15)</i>	484	5721	5237	1082.02%
<b>datos.codeandomexico.org</b> <i>(2014-07-17) → (2015-03-15)</i>	94	715	621	660.64%
<b>data.opendataportal.at</b> <i>(2014-07-17) → (2015-03-16)</i>	46	323	277	602.17%
<b>annuario.comune.fi.it</b> <i>(2014-08-07) → (2015-03-15)</i>	50	351	301	602.00%
<b>udct-data.aigid.jp</b> <i>(2014-08-07) → (2015-03-16)</i>	431	2110	1679	389.56%
<b>catalogo.datos.gob.mx</b> <i>(2014-08-08) → (2015-03-15)</i>	111	360	249	224.32%










# Data Dumps

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

## Open Data Portal Watch Dumps

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

## Open Data Portal Watch Dumps

Name	Last modified	Size
 Parent Directory		-
 2014-07-17.gz	05-Feb-2015 15:13	2.2M
 2014-07-25.gz	05-Feb-2015 15:13	2.2M
 2014-08-05.gz	05-Feb-2015 15:13	2.2M
 2014-08-12.gz	05-Feb-2015 15:13	2.2M
 2014-08-27.gz	05-Feb-2015 15:13	2.2M
 2014-09-01.gz	05-Feb-2015 15:14	2.2M
 2014-09-07.gz	05-Feb-2015 15:14	2.2M
 2014-09-14.gz	05-Feb-2015 15:14	2.2M

## Towards assessing the quality evolution of Open Data portals

ODQ  
2015

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

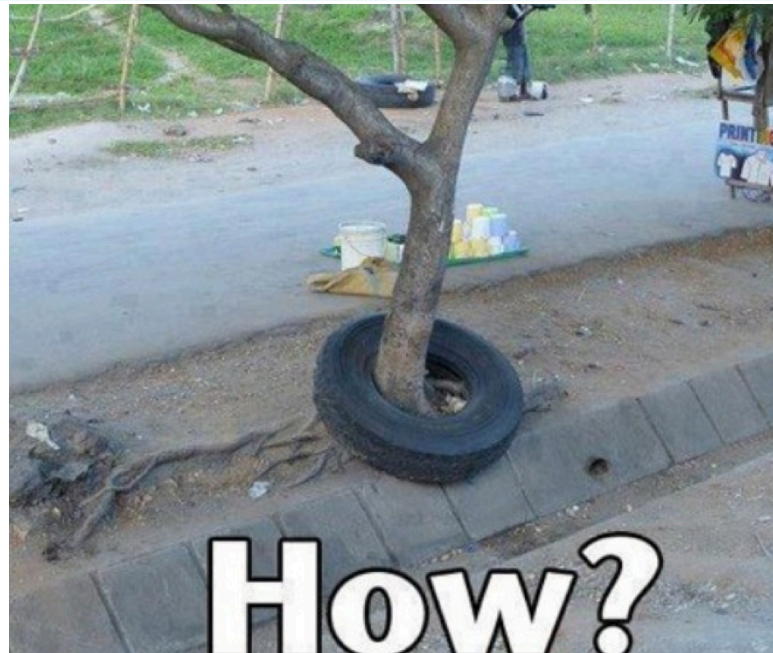
In this work, we present the Open Data Portal Watch project, a public framework to continuously monitor and assess the (meta-)data quality in Open Data portals. We critically discuss the objectiveness of various quality metrics. Further, we report on early findings based on 22 weekly snapshots of 90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

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

- Key findings:
  - Varying quality acrosss 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

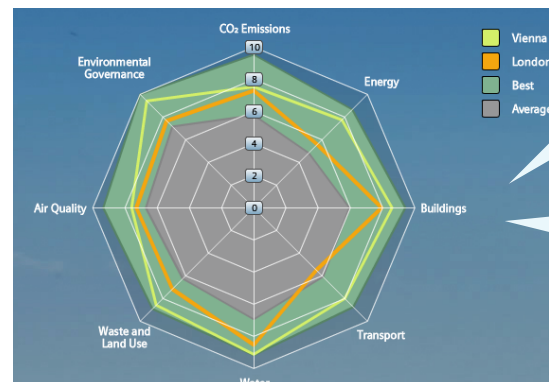
- City Assessment and Sustainability reports
- Tailored offerings by Infrastructure Providers



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

→ 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:



Goal (short term):

- Leverage Open Data for calculating a city’ 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**

# City Data Pipeline

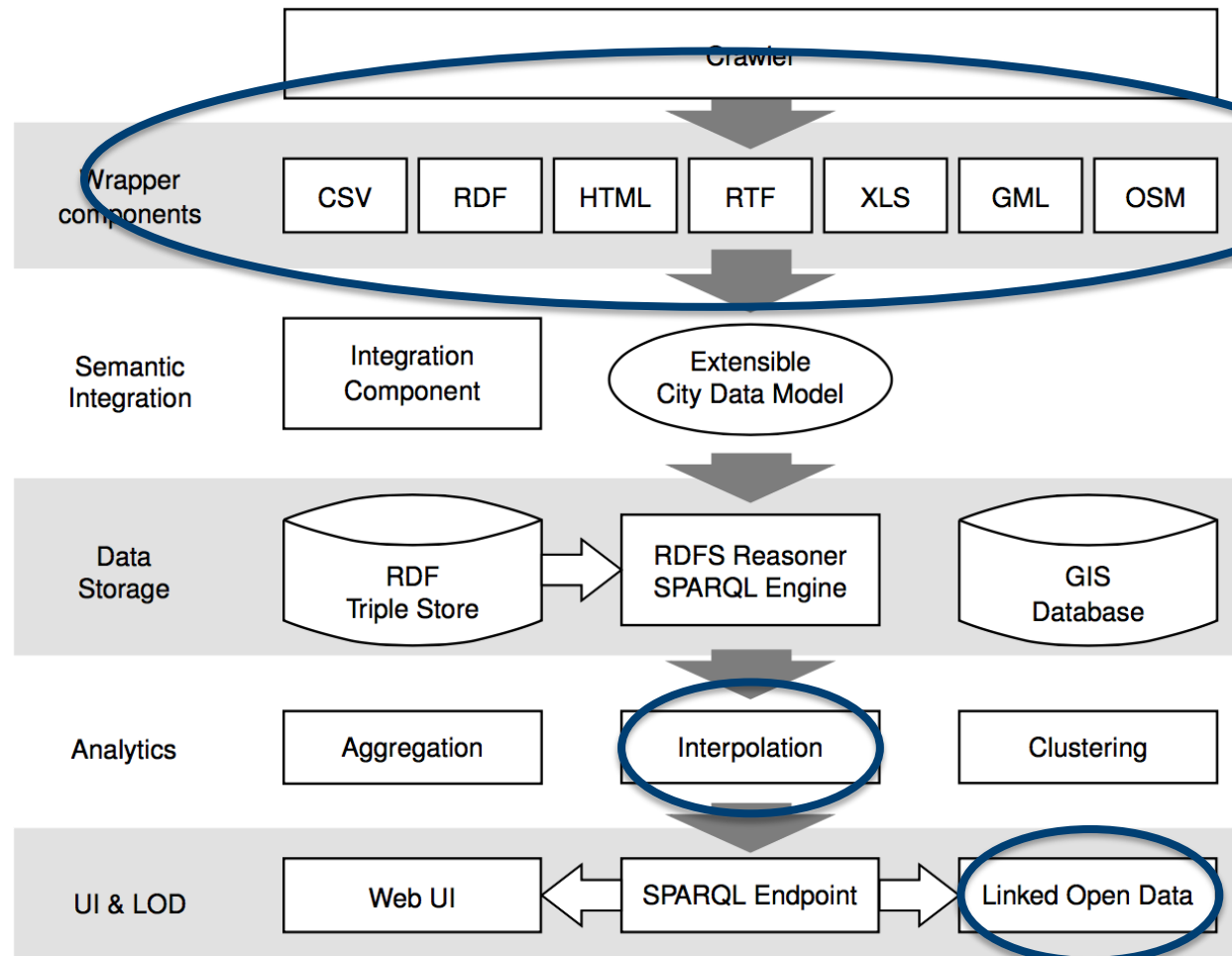
- <http://citydata.wu.ac.at/>

## 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.

[Launch Open City Data Pipeline](#)

# City Data Pipeline: Architecture



Crawl and **integrate Open data** sources with city data: Wikipedia Eurostat UNData USCensus Etc.

Use statistical methods to **approximate missing values**

Re-publish as **Linked Data** (SPARQL endpoint, provenance information, etc.)

# Challenges – Missing values

- 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

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

Table 2: United Nations Data Set

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

# 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

	Vienna	Augsburg	Valletta
Cars	655806	111561	95858
Nationals	1342704	216289	203657
Women per 1000 Men	109.8	108.7	101.9

Data from Source 2

	Marbella	Stockholm	Funchal
Available Beds per 1000	138.3	14969	166.1
Average area of living	36.42	37.24	38.16
Cinema Seats	4691	12751	2676



Combined data from Source 1 and Source 2

	Vienna	Augsburg	Valletta	Marbella	Stockholm	Funchal
Cars	655806	111561	95858			
Nationals	1342704	216289	203657			
Women per 1000 Men	109.8	108.7	101.9			
Available Beds per 1000				138.3	14969	166.1
Average area of living				36.42	37.24	38.16
Cinema Seats				4691	12751	2676

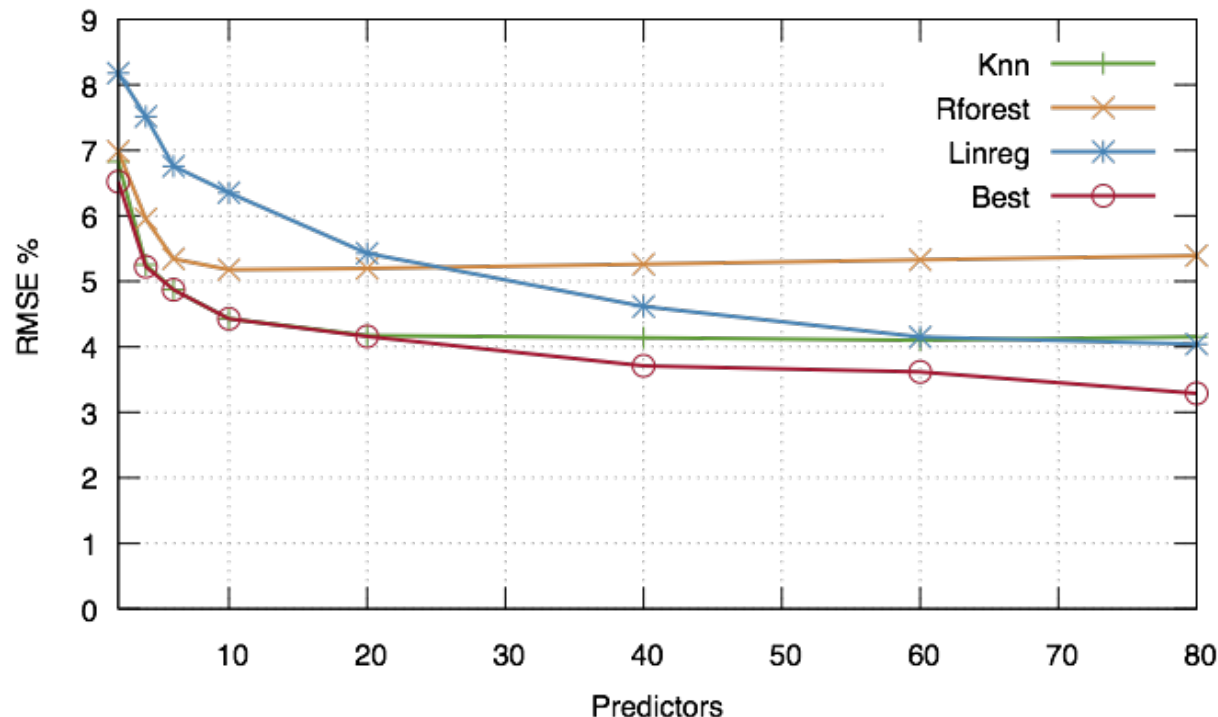
# Missing Values – Hybrid approach choose best prediction method per indicator:

- 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)



# Missing Values – Hybrid approach choose best prediction method per indicator:

- 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](http://citydata.wu.ac.at)

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

The screenshot shows a web browser window with the URL <http://citydata.ai.wu.ac.at/KPIDataPipeline/KPIDispatcher>. The page features the logos for WU (Wirtschaftsuniversität Wien) and Siemens. Below the logos, there are two columns of data for Berlin and Vienna.

Berlin	Vienna
<b>Population male 2012</b> 1717645.0 persons (Source: <a href="http://epp.eurostat.ec.europa.eu/">http://epp.eurostat.ec.europa.eu/</a> )	<b>Population male 2011</b> 821605.0 persons (Source: <a href="http://data.un.org/">http://data.un.org/</a> )
<b>Population male 2011</b> 1695438.0 persons (Source: <a href="http://data.un.org/">http://data.un.org/</a> )	<b>Population male 2010</b> 812867.0 persons (Source: <a href="http://data.un.org/">http://data.un.org/</a> )
<b>Population male 2011</b> 1695438.0 persons (Source: <a href="http://epp.eurostat.ec.europa.eu/">http://epp.eurostat.ec.europa.eu/</a> )	<b>Population male 2009</b> 807088.0 persons (Source: <a href="http://data.un.org/">http://data.un.org/</a> )
<b>Population male 2010</b> 1686256.0 persons (Source: <a href="http://epp.eurostat.ec.europa.eu/">http://epp.eurostat.ec.europa.eu/</a> )	<b>Population male 2009</b> 807088.0 persons (Source: <a href="http://epp.eurostat.ec.europa.eu/">http://epp.eurostat.ec.europa.eu/</a> )
<b>Population male 2009</b> 1686256.0 persons	<b>Population male 2008</b> 801776.0 persons (Source: <a href="http://data.un.org/">http://data.un.org/</a> )
	<b>Population male 2008</b> 800361.0 persons



Vienna 

Municipal waste (1000 t)

- › **2004**: 778.905392176222 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)
- › **2005**: 813.77643147163 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)
- › **2006**: 813.889824195497 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)
- › **2007**: 811.538914636665 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)
- › **2008**: 811.010344391444 1000 t (from <http://citydata.wu.ac.at/ns#Prediction>, predicted by with an estimated error of %RMSE)

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



# More Details:

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 Bischof<sup>1,2</sup>, Christoph Martin<sup>2</sup>, Axel Polleres<sup>2</sup>, and Patrik Schneider<sup>2,3</sup>

<sup>1</sup> Siemens AG Österreich, Vienna, Austria

<sup>2</sup> Vienna University of Economics and Business, Vienna, Austria

<sup>3</sup> 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
- <https://www.data.gv.at/wp-content/uploads/2012/03/Mission-Statement-AG-Qualitaetssicherung-OpenData-Portale.pdf>

COOPERATION OGD  ÖSTERREICH

## Datenqualität und Veröffentlichungsprozesse

Mission Statement Sub-Arbeitsgruppe *Qualitätssicherung auf Open Data-Portalen* der Cooperation Open Government Data Österreich

Version 1.0 - Autoren: Johann Höchtl, Axel Polleres, Jürgen Umbrich, Brigitte Lutz

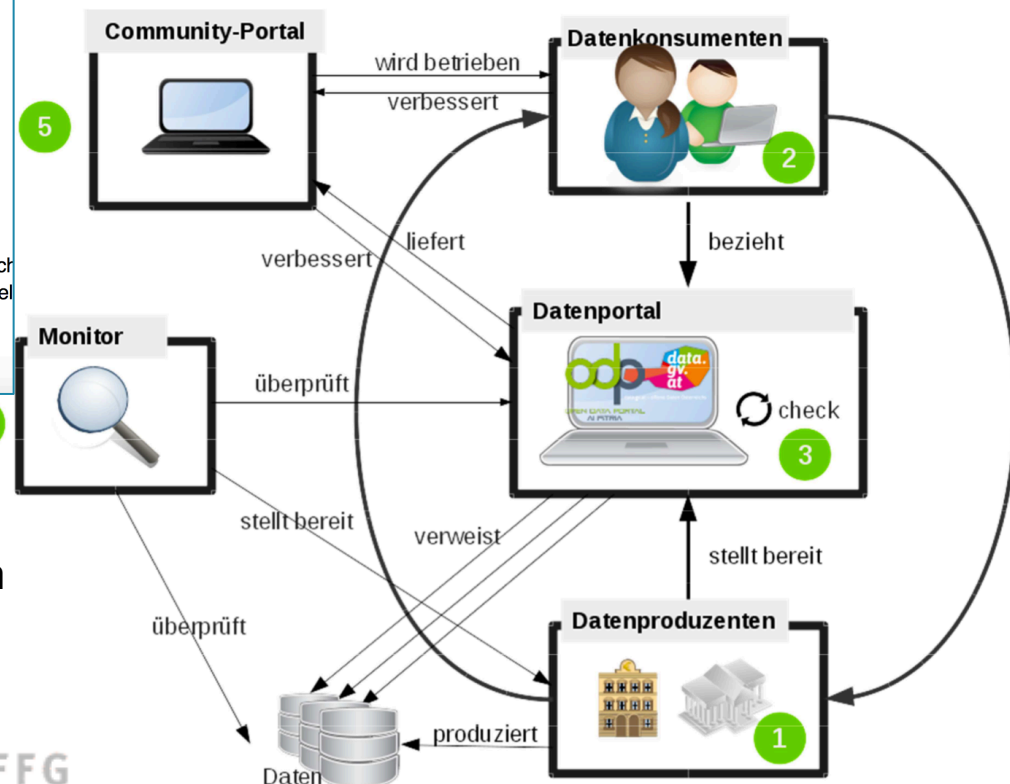
### Mission Statement

Die Sub-Arbeitsgruppe *Qualitätssicherung von Open Data Portalen* verbessert durch technische Maßnahmen und die Erstellung von Leitfäden zur empfohlenen Praxis die Datenqualität aktueller verfügbarer Datensätze und unterstützt durch organisatorische und technische Maßnahmen den Veröffentlichungsprozess, um in Zukunft höhere Qualitätsniveaus, und somit erhöhte Nutzbarkeit und Nachhaltigkeit von offenen Daten zu erreichen.

- 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> for non-governmental open data launched in 2014

<http://www.opendataportal.at/>



*Sie haben mehr Daten als Sie denken!  
Alles was Sie brauchen ist Maschinenlesbarkeit - und  
Ruck Zuck wird ihr Datensatz zum Innovationsschatz.*

- We can help you to use and publish Open Data!  
WU, TU, SWC, DUK have **just founded** a network node of the



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.Pollereres@wu.ac.at](mailto:Axel.Pollereres@wu.ac.at) Twitter: [@AxelPolleres](https://twitter.com/AxelPolleres)



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

- Maybe see you at one of the following events:



<http://2015.data-forum.eu/>

European Data Forum 2015

November 16-17, 2015, Luxembourg

<http://semantics.cc/>

**SEMANTICS**  
Vienna 2015



VIENNA, SEPTEMBER 15-17, 2015