Data Integration for (Linked?) Open Data on the Web

Axel Polleres twitter: @AxelPolleres web: polleres.net

My background

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- Logic Programming, Artificial Intelligence
- Semantic Web, Knowledge Representation & Reasoning
- Web Standard, Query Language, Linked Data
- Data Integration, Open Data

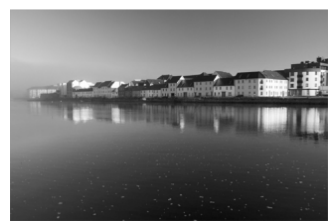
Let's start here

2003 2005 2007 2011 2013

TU Vienna

Univ. Innsbruck

Univ. Rey Juan Carlos Madrid



DERI, NUI Galway, Ireland

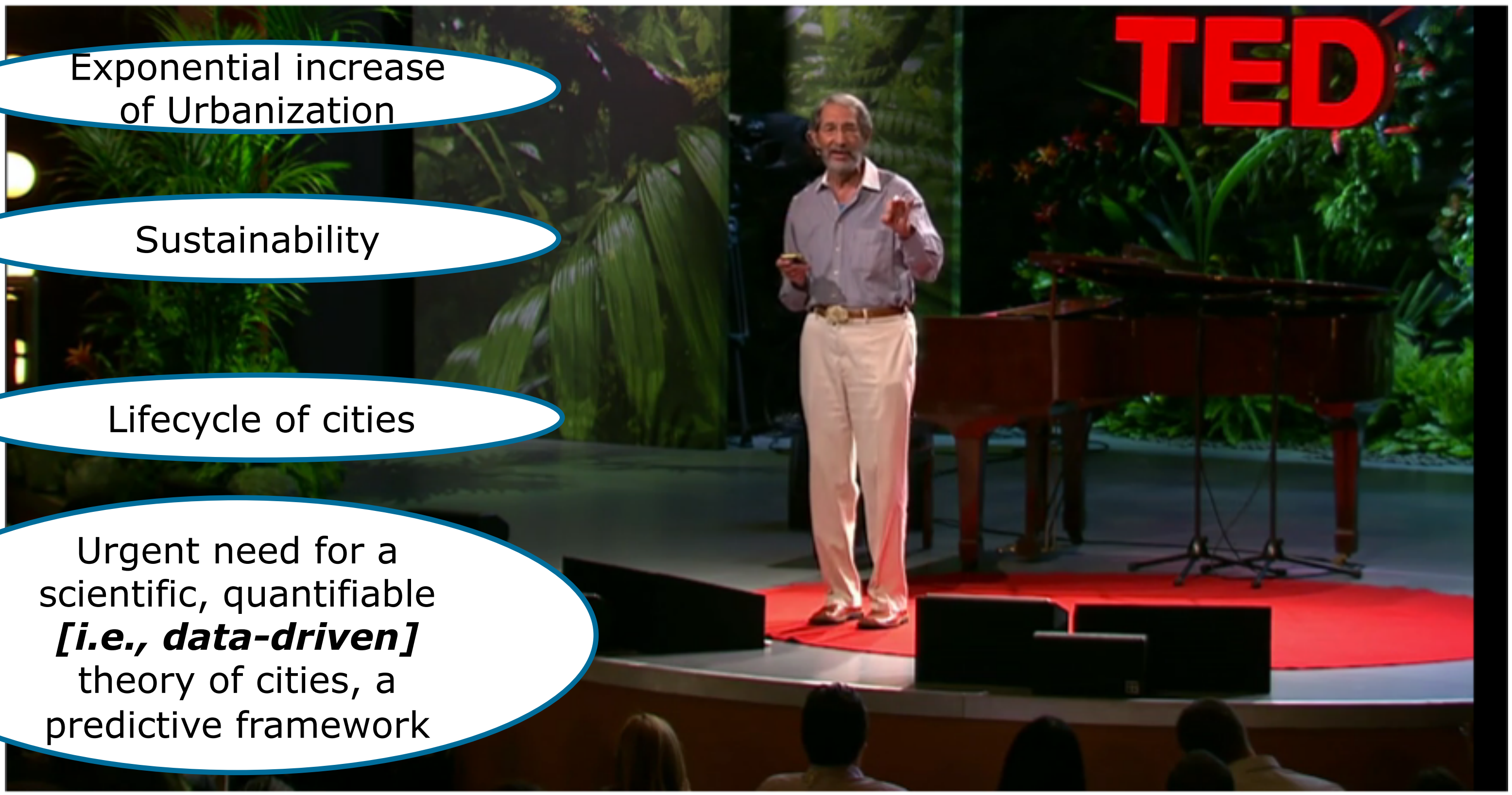
Siemens AG Österreich

WU Vienna

Disclaimer: this talk is meant as a "teaser" …
(technical details in my class in spring term: BIOMEDIN 274)

A motivating use case: Geoffrey West
(former director of the Santa Fe Institute) 2011

Conjecture: the functioning of cities can be explained by data



https://www.ted.com/talks/geoffrey_west_the_surprising_math_of_cities_and_corporations/

Back at around that time… City Data – Important for Infrastructure Providers & for City Decision Makers

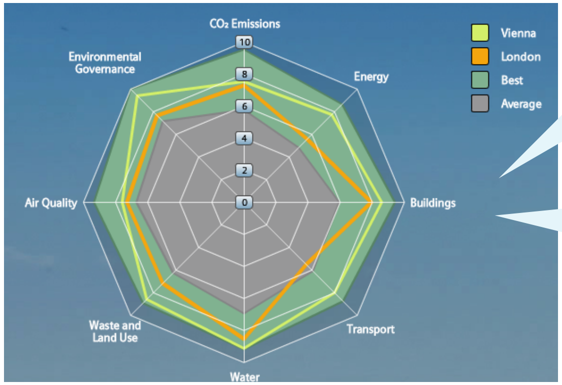
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- 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… however, these are often **outdated** before even published!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:



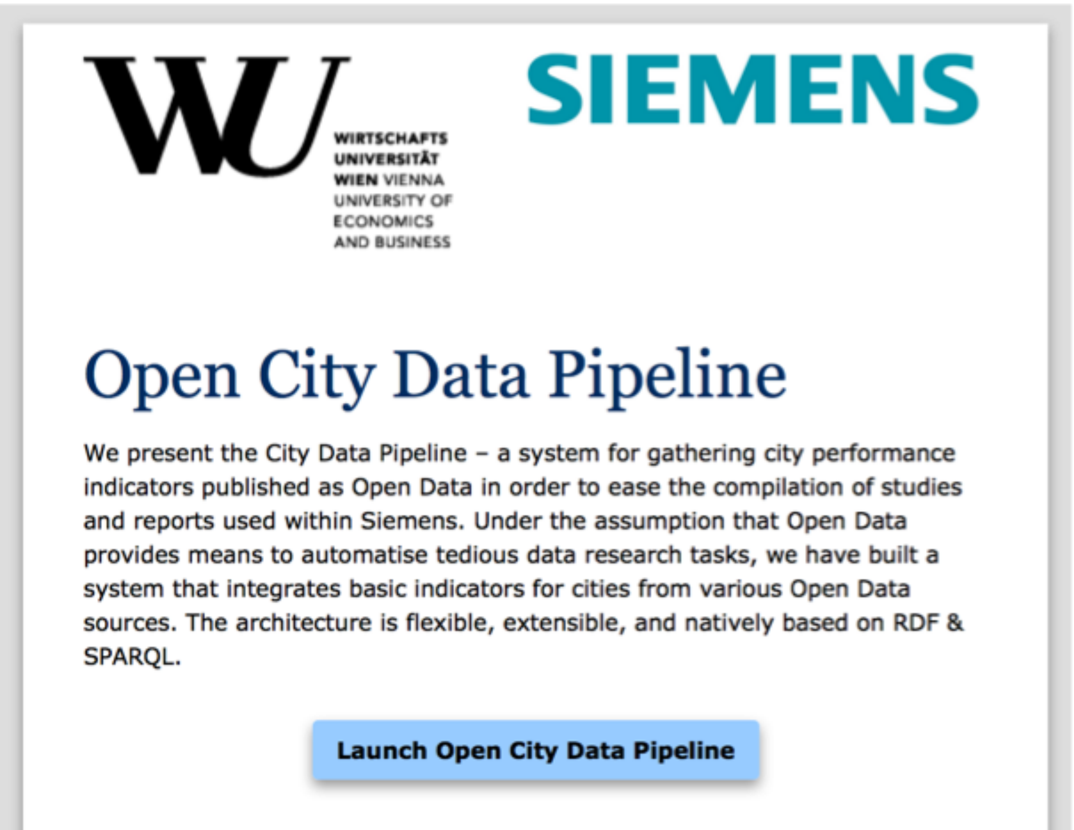
Goal (short term):
▪Leverage Open Data for calculating a city' performance from public sources on the Web **automatically**

Goal (long term):
▪Define and Refine KPI models to assess specific impact of infrastructural investments and gather/check input **automatically**

EQUIS

City Data Pipeline (started 2012)

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

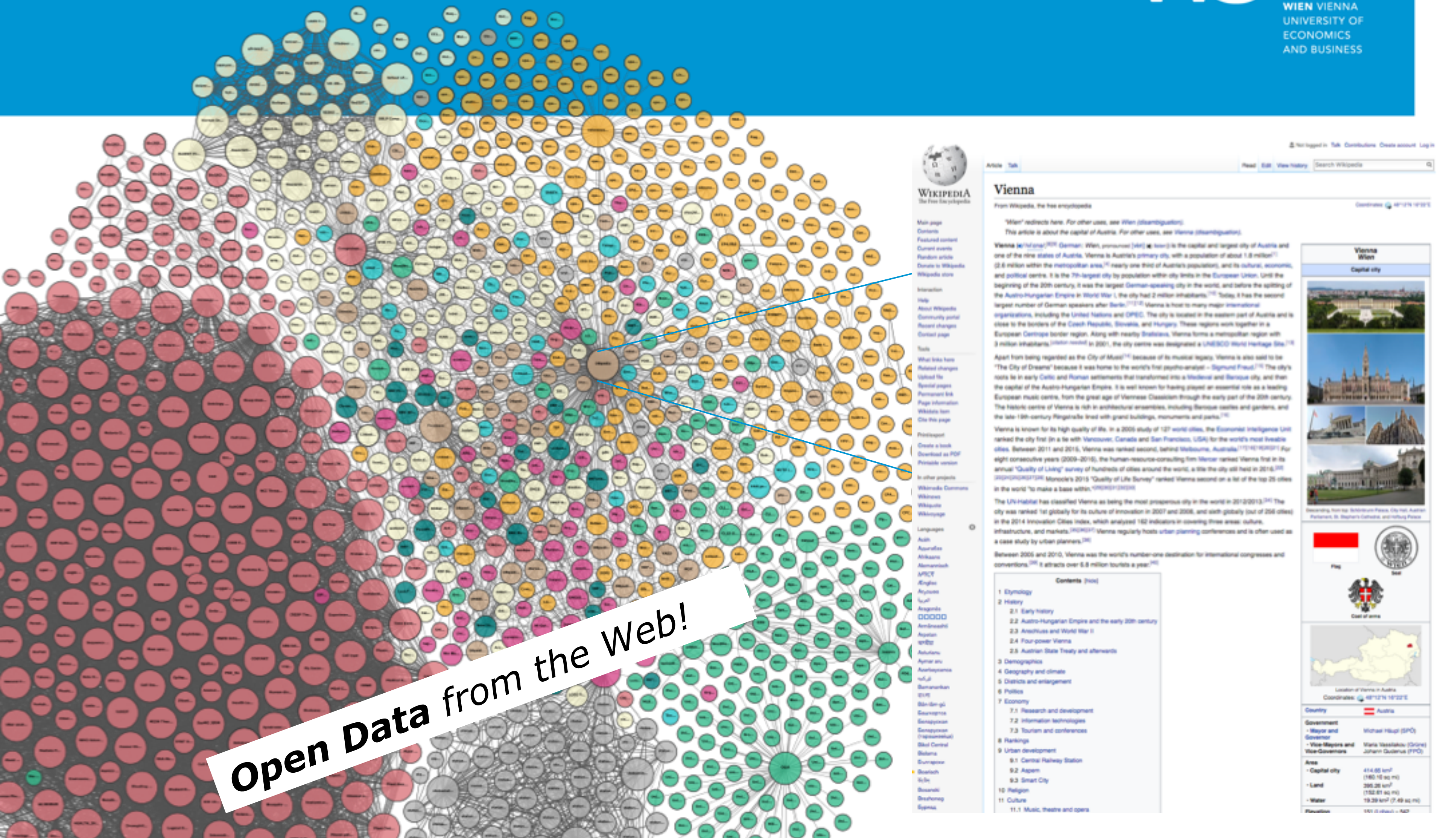


Which assets can we* draw from?

*the Semantic Web community

- Where do we find Data?
  - Semantic Search
  - Linked Data
- How do we combine Data?
  - RDFS and OWL inference
- Is that enough?
  - Probably not…

This is what Linked Data offers us:



Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

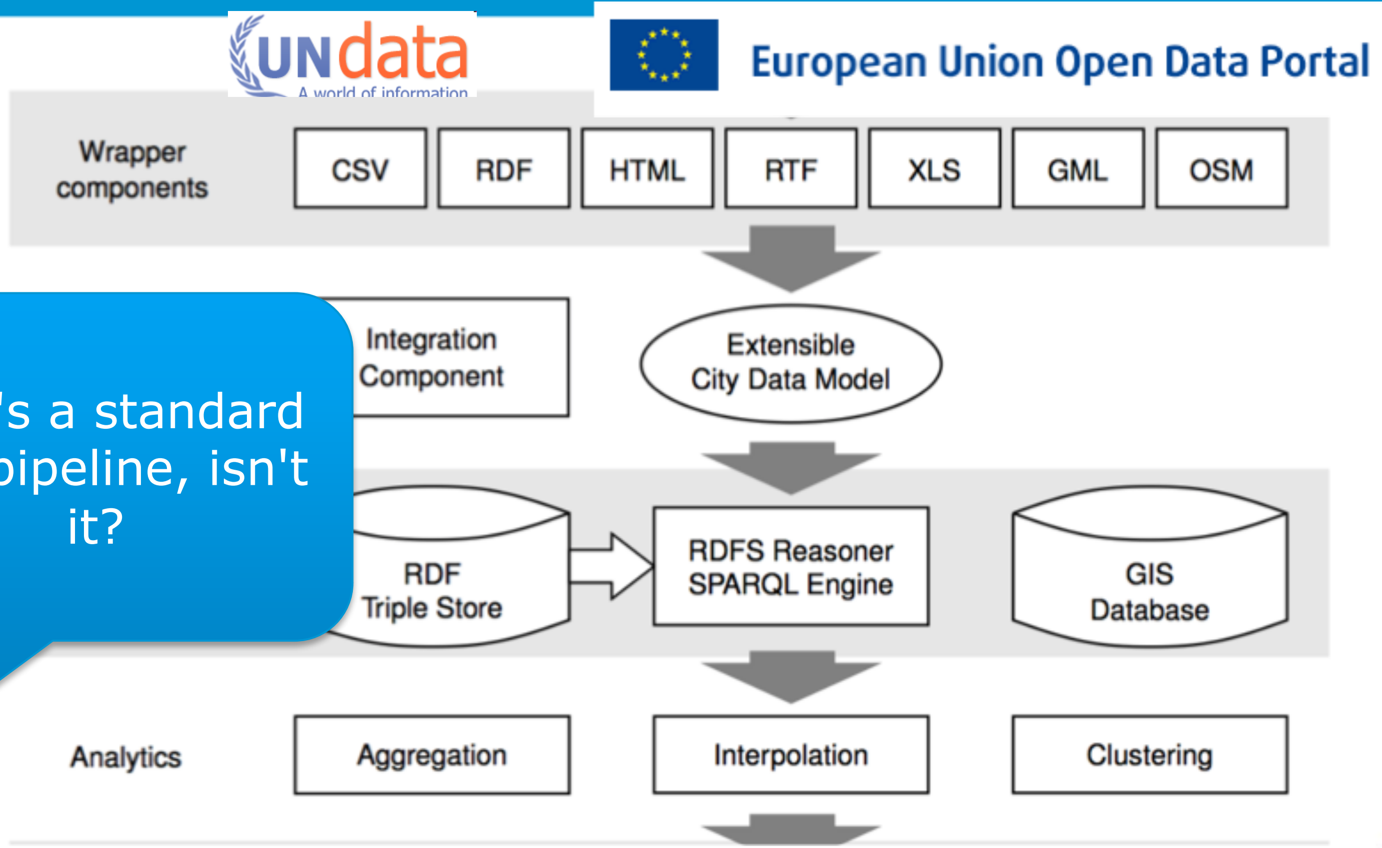
But: there's a lot of Open Data missing (apart from Linked Data):

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

Ok, now... how can I use it?

Attempt 1: use OWL&RDFS

A concrete use case:
The "City Data Pipeline"



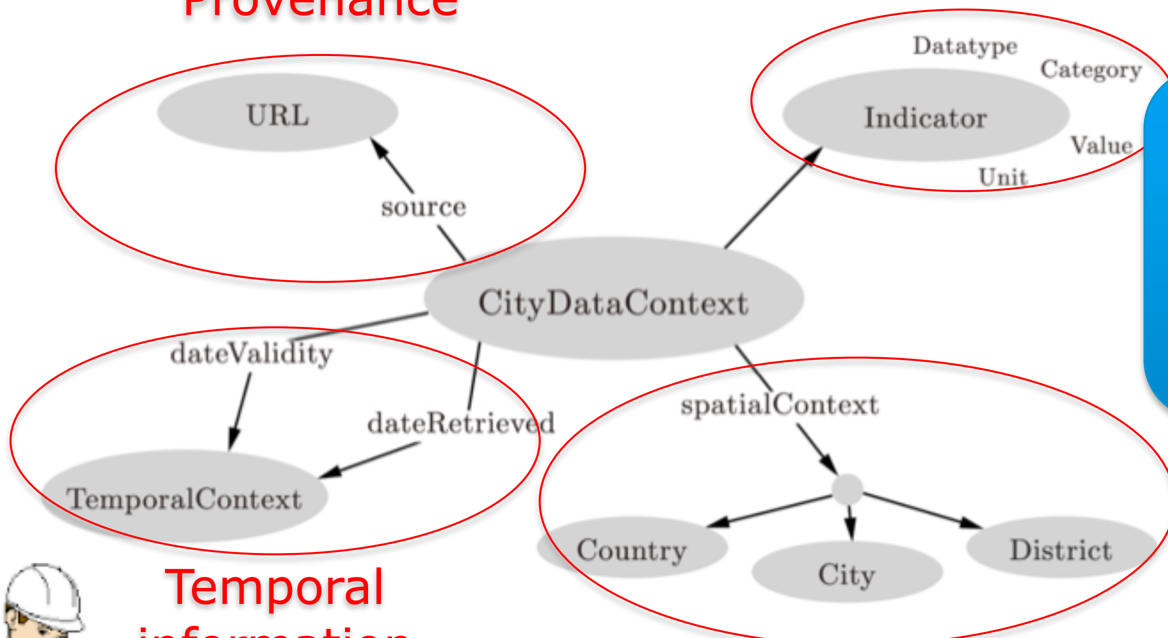
That's a standard ETL pipeline, isn't it?

A concrete use case: The "City Data Pipeline"

City Data Model: extensible $\mathcal{ALH}(\mathbf{D})$ ontology:

Indicators,
e.g. area in km2,
tons CO2/capita

Provenance



Temporal information

Spatial context

But we use and flexible Semantic integration using **ontologies** and **reasoning**!

A concrete use case: The "City Data Pipeline"

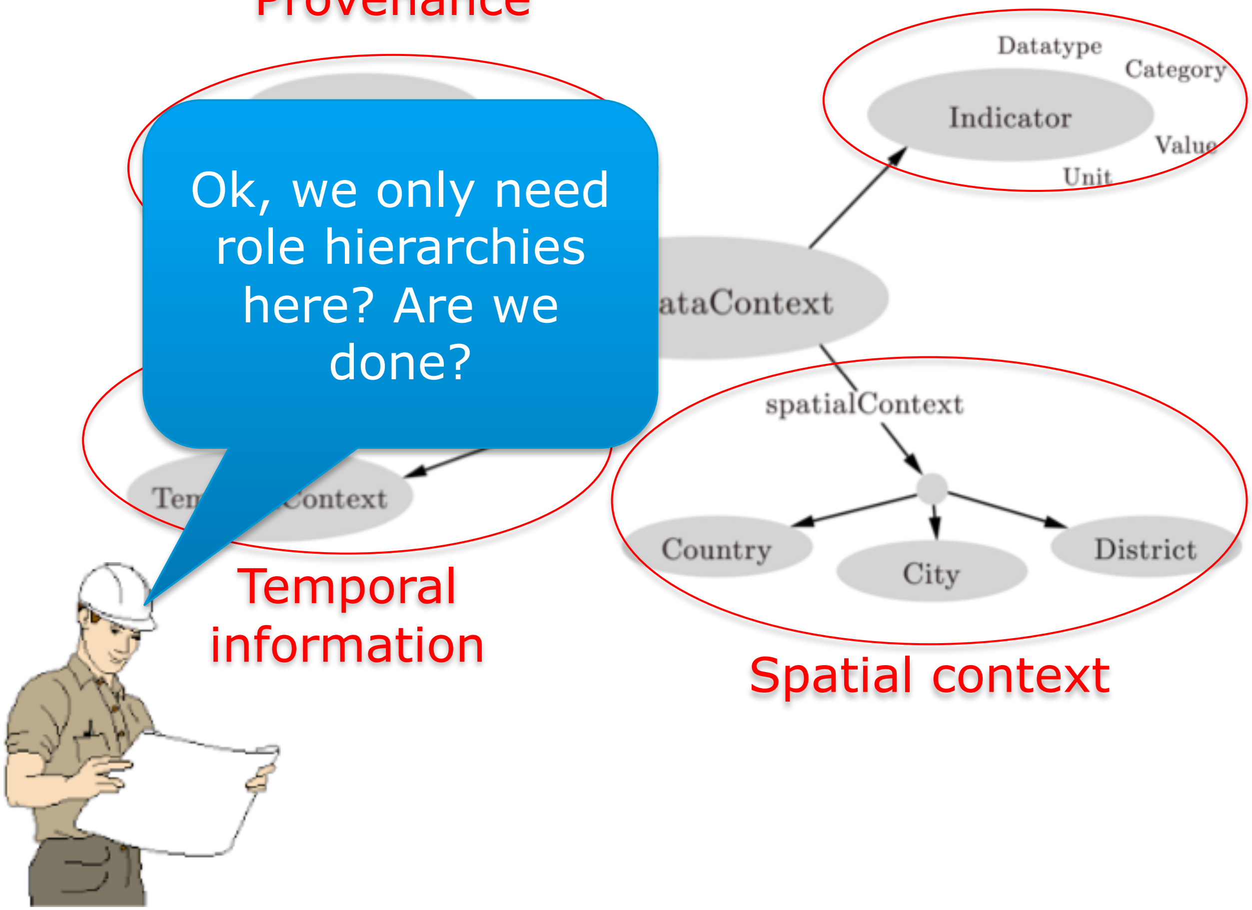
City Data Model: extensible $\mathcal{ALH}(\mathbf{D})$ ontology:

dbpedia:areakm $\sqsubseteq$:area

eurostat:area $\sqsubseteq$:area

Provenance

Indicators,
e.g. area in km2, tons CO2/capita



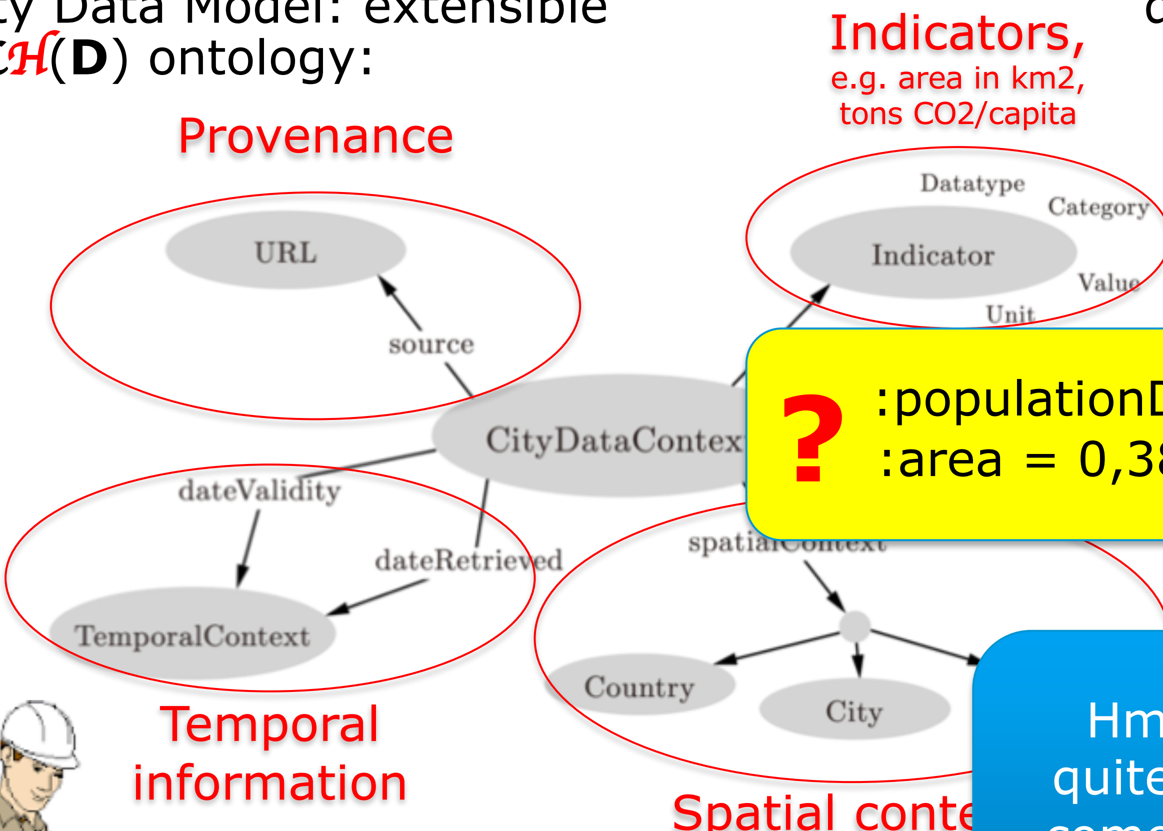
Spatial context

A concrete use case:
The "City Data Pipeline"

City Data Model: extensible $\mathcal{ALH}(\mathbf{D})$ ontology:

dbpedia:areakm2 $\sqsubseteq$:area

eurostat:area $\sqsubseteq$:area



? :populationDensity = :population/:area
:area = 0,386102 * dbpedia:areaMi2

Hmmm, not quite... Let me come up with a solution...

Can equational knowledge co-exist with OWL?

- *Can equational knowledge co-exist with OWL?*
  - *We need a syntax & define a formal semantics*

- *Syntax:* :populationDensity = :population/:area
  :area = 0,386102 * dbpedia:areaMi2

:populationDensity **:defineByEquation** “population/:area” .
:area **:defineByEquation** “areaMi2 * 0,386102 ” .
dbPedia:populationTotal **:rdfs:subPropertyOf** :population.

- Semantics:
  - Requirements:
    - “Fit” with common model-theoretic semantics for OWL and RDFS
    - Treat equivalent equations equivalently, combine with **query rewriting** and **rule-based reasoning** techniques:

:area = 0,386102 * dbpedia:areaMi2

:areaMi2 = 2,589988 * :area

Can equational knowledge co-exist with OWL?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

:Vienna dbPedia:populationTotal 1852997.

:Vienna :area 414.65.

dbPedia:populationTotal **:rdfs:subPropertyOf** :population.

:populationDensity **:defineByEquation** “population/:area” .
:area **:defineByEquation** “areaMi2 * 0,386102 ” .
dbPedia:populationTotal **:rdfs:subPropertyOf** :population.

:Vienna :populationDensity 4 467.

- Semantics:
  - Requirements:
    - “Fit” with common model-theoretic semantics for OWL and RDFS
    - Treat equivalent equations equivalently, combine with **query rewriting** and **rule-based reasoning** techniques:

:area = 0,386102 * dbpedia:areaMi2

:areaMi2 = 2,589988 * :area

Can equational knowledge co-exist with OWL?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

RDFS with Attribute Equations via SPARQL Rewriting

So:
- RDFS and OWL inference
- & Equational Knowlege

Is that enough?
Probably not…

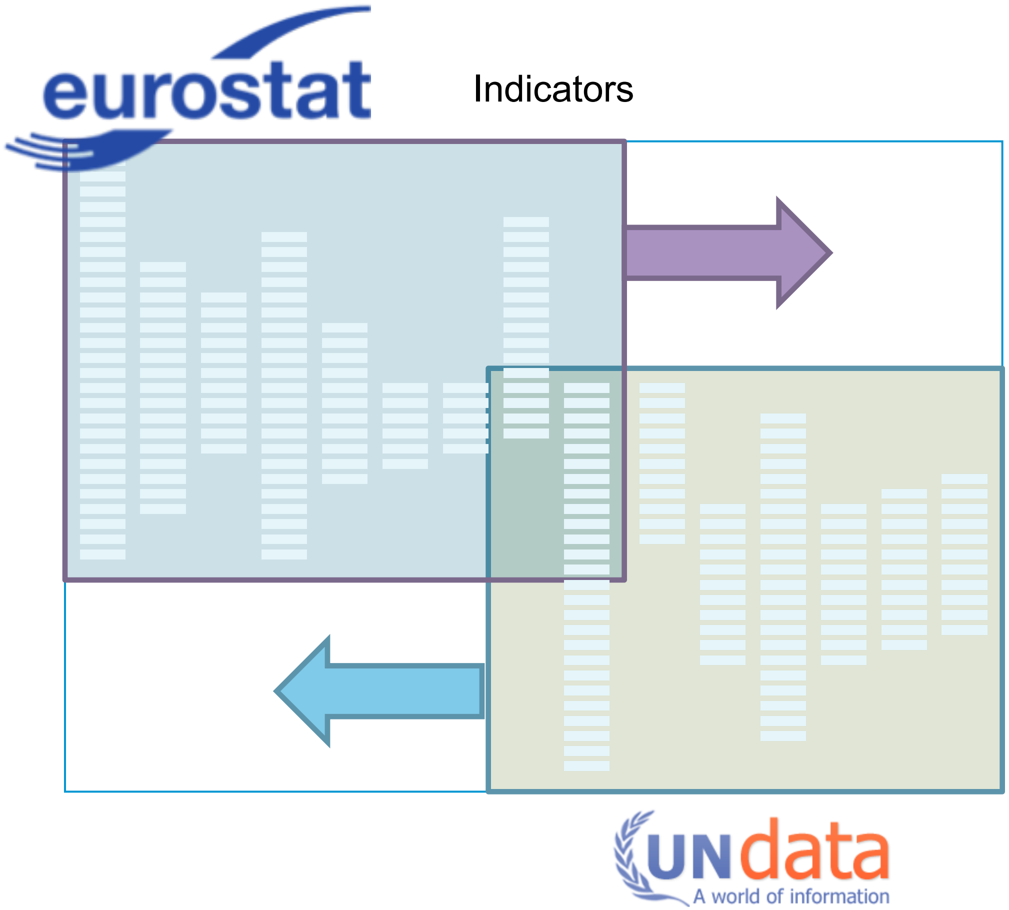
Stefan Bischof, Axel Polleres. ESWC2013

EFMD EQUIS ACCREDITED

Challenges – Too many Missing values

Problem: Equational knowledge is not enough to deal with many missing values…

Idea: using both first-wave and second wave AI (ML&statistics) methods



Challenges – Too many 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

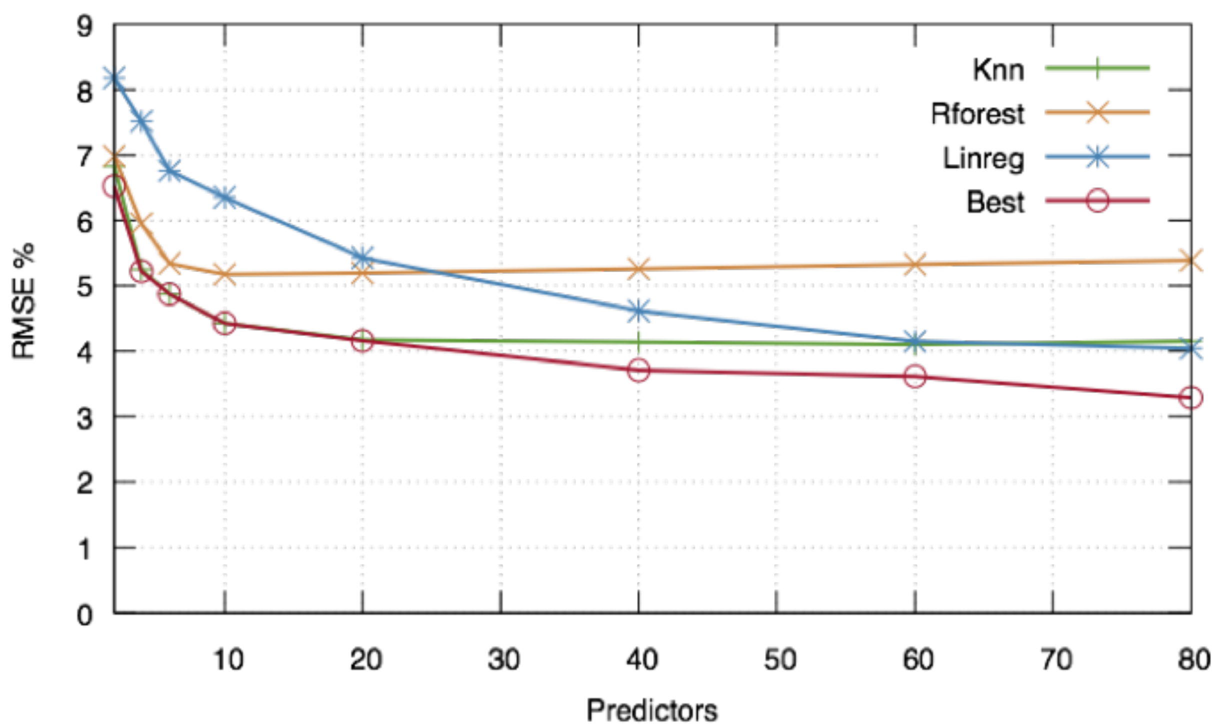
	Vienna	Augsburg	Valletta	Marbella	Stockholm	Funchal
Cars	655806	111561	95858			
Nationals	1342704	216289	203657			
Women per 1000 Men	109.8	108.7	101.9			
Available Beds per 1000				138.3	14969	166.1
Average area of living				36.42	37.24	38.16
Cinema Seats				4691	12751	2676

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 Principle Components (PCA), built from the indicators
- For buidling the PCs, fill in missing data points with neutral values → predict all rows
-
-



More Details:

Stefan Bischof, Christoph Martin, Axel Polleres, and Patrik Schneider.Collecting, integrating, enriching and republishing open city data as linked data. *In Proceedings of the 14th International Semantic Web Conference (ISWC 2015)*

Collecting, Integrating, Enriching and Republishing Open City Data as Linked Data*

Stefan Bischof[1,2], Christoph Martin[2], Axel Polleres[2], and Patrik Schneider[2,3]

[1] Siemens AG Österreich, Vienna, Austria
[2] Vienna University of Economics and Business, Vienna, Austria
[3] Vienna University of Technology, Vienna, Austria

Abstract. Access to high quality and recent data is crucial both for decision makers in cities as well as for 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 reusable 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 (PCA) to improve quality and amount of predicted values. Since indicators and cities only have partial overlaps across data sets, we particularly focus on predicting indicator values across data sets, where we extend, adapt, and evaluate our prediction model for this particular purpose: as a "side product" we learn ontology mappings (simple equations and sub-properties) for pairs of indicators from different data sets. Finally, we republish the integrated and predicted values as linked open data.

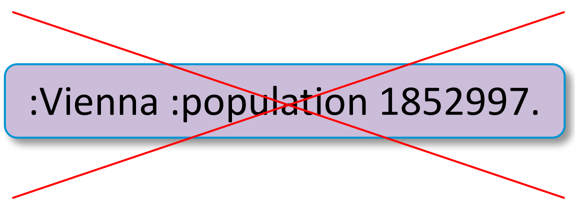
Next step:

Combine ML and equations "iteratively" (under submission)

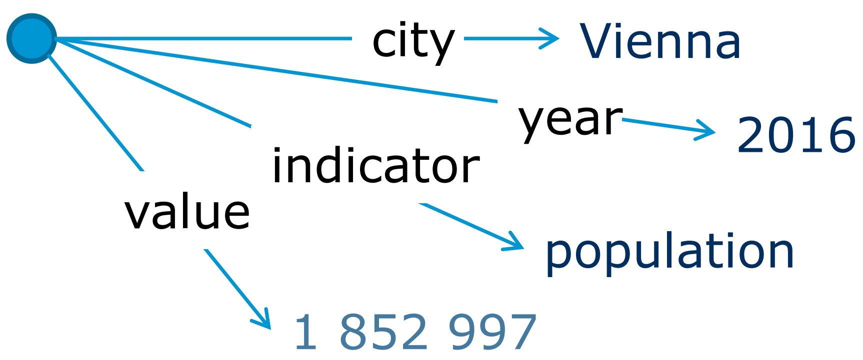
http://epub.wu.ac.at/5438/

QB equations: Combine ML and equations "iteratively" … How?

- First of all, RDF data about cities doesn't look like this:

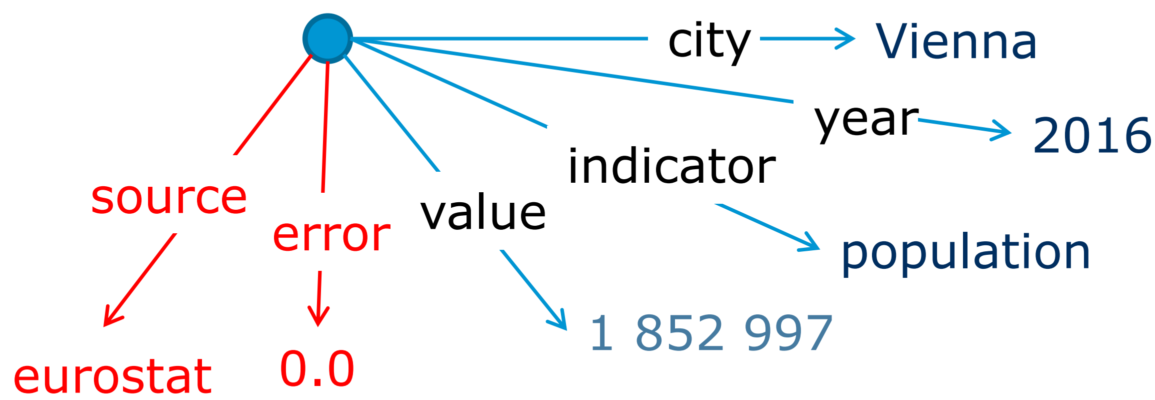


- But like this:

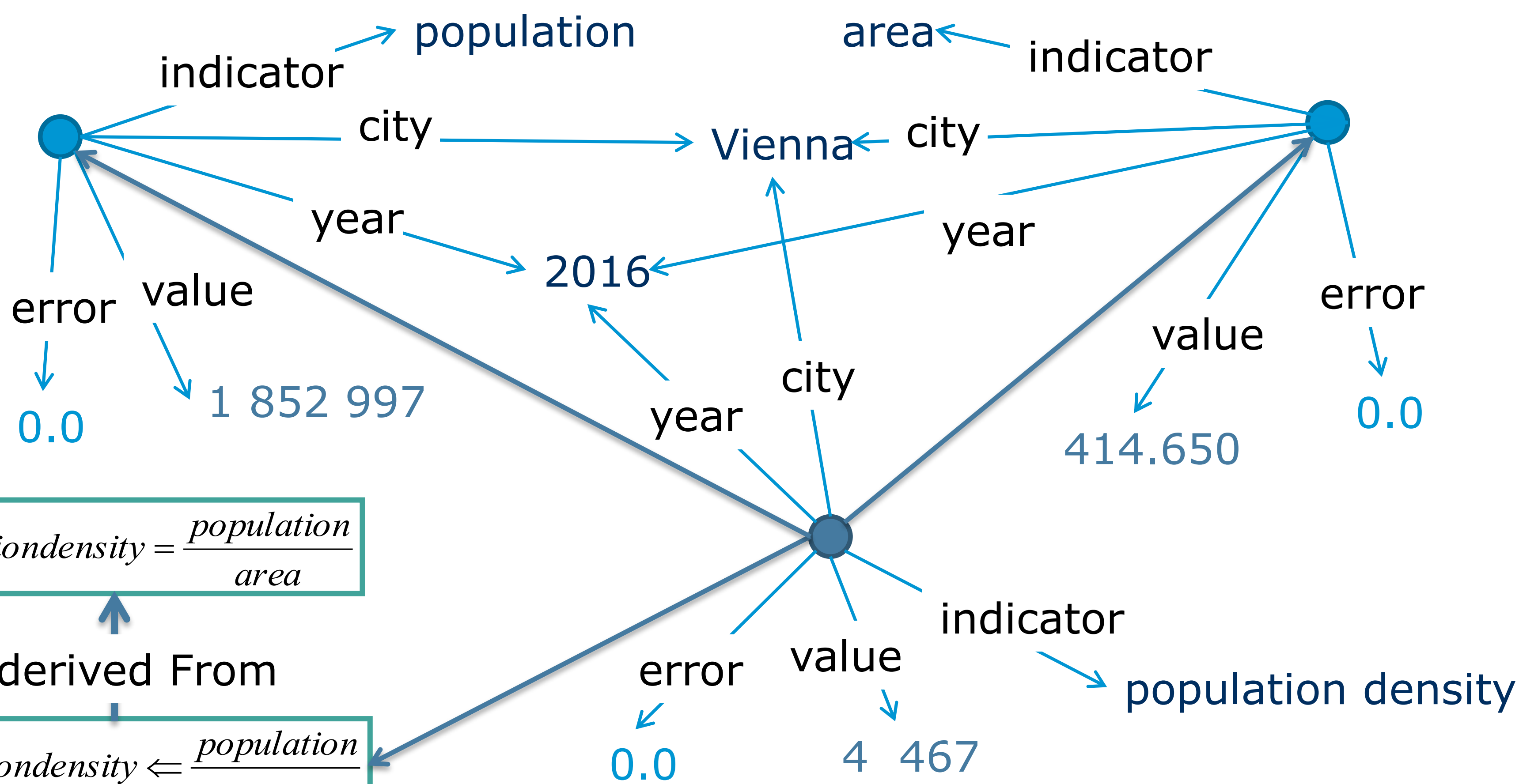


RDF Attribute Equations are not enough

- Data from some sources like eurostat come as multidimensional data - Data Cube vocabulary (**QB**):
  - Temporal (December)
  - Unit of measurement (degrees Celsius)
  - Aggregation (mean, min, max, …)
  - *Indicator (temperature, population density)*



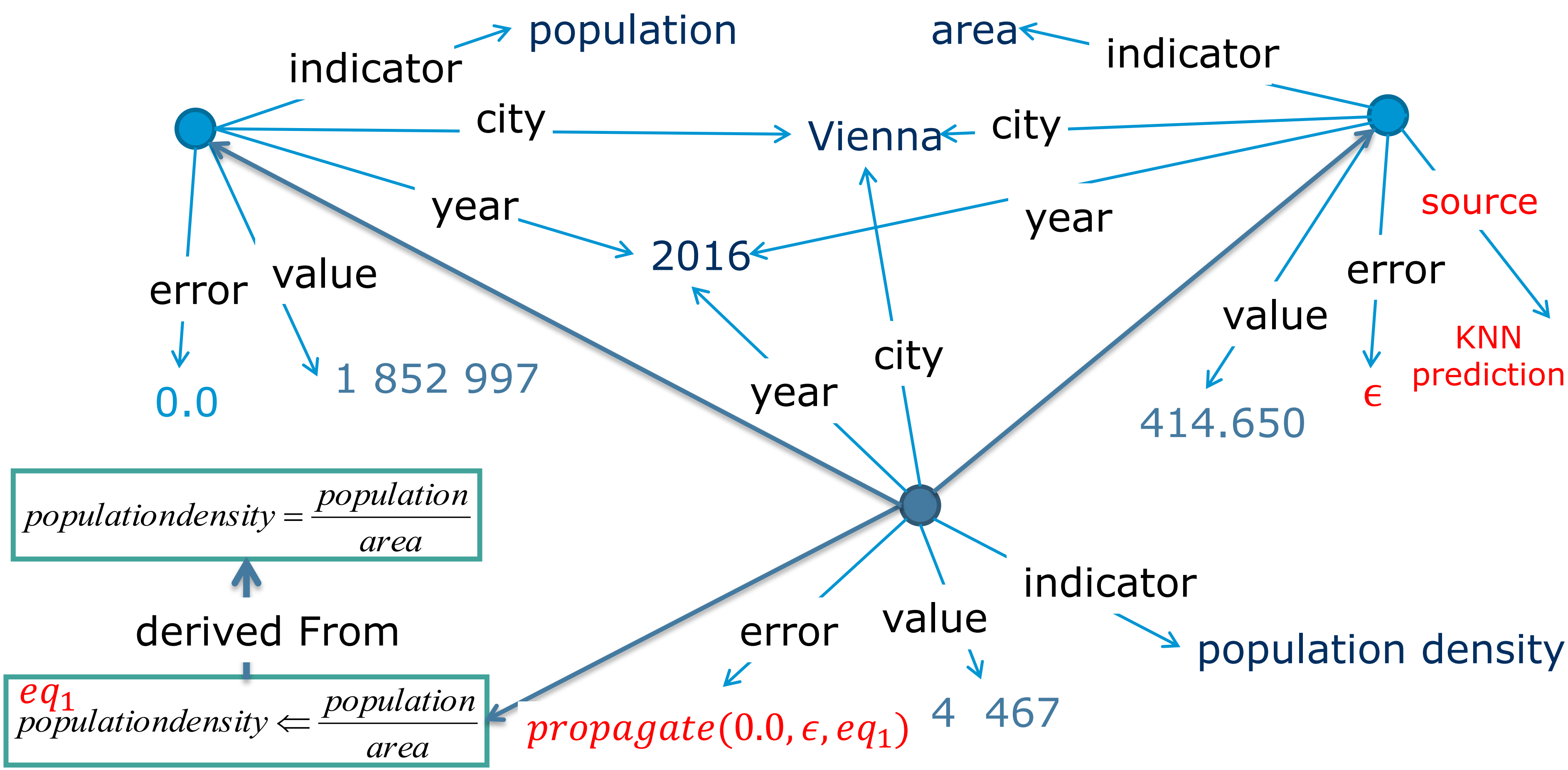
Example QB equations: compute population density



$$populationdensity = \frac{population}{area}$$

$$populationdensity \Leftarrow \frac{population}{area}$$

QB equations: deal with approximated values by using error propagation



More Details:

Stefan Bischof, Andreas Harth, Benedikt Kämpgen, Axel Polleres, and PatrikSchneider.Enriching integrated statistical open city data by combining equational knowledge and missing value imputation. *Journal of Web Semantics (JWS)*, October 2017.In press.

Open City Data Pipeline

Collecting, Integrating, and Predicting Open City Data

Stefan Bischof[1,2], Christoph Martin[2], Axel Polleres[2], and Patrik Schneider[2,3]

[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.

Main idea:
Combine
(1) ontological reasoning,
(2)ML, **(3) equations**
"iteratively" with
QB equations

Evaluation Combination
PCA Regression + QB Equations

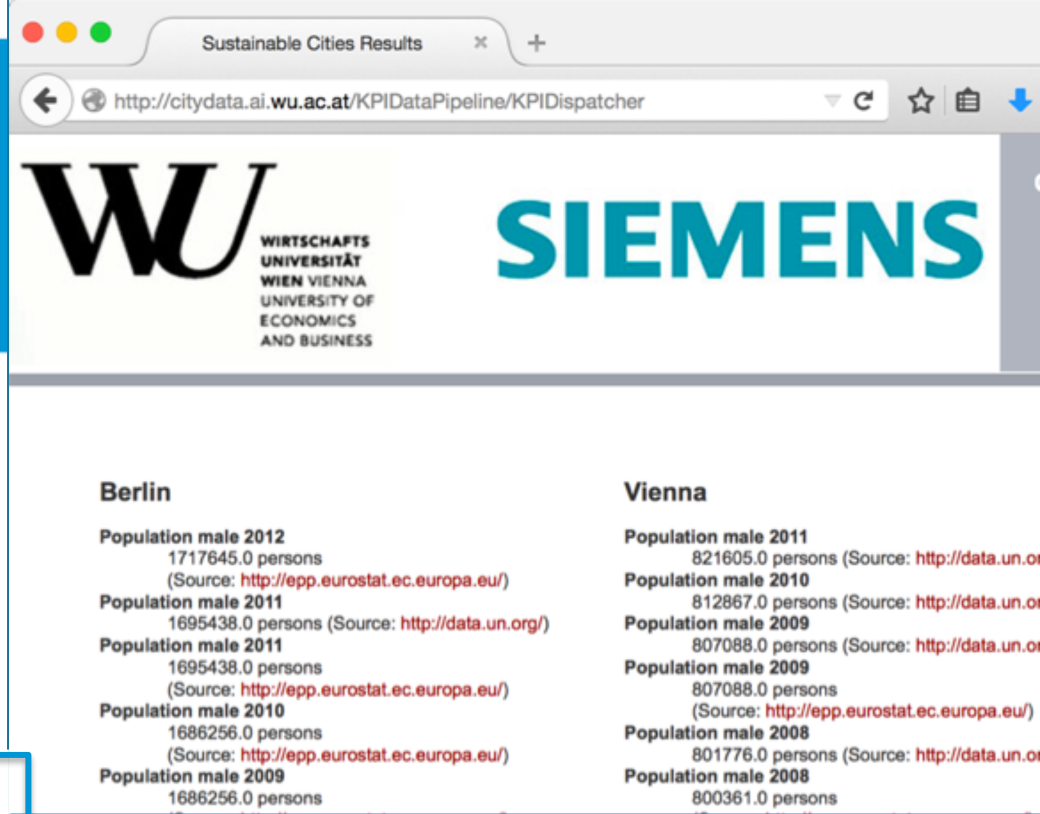
- Statistics one iteration (PCA regression + QB Equations)
  - 991k observations from crawled data
  - 522k new or better observations from PCA regression
  - 230k better observations from QB Equations
  - 232k new observations from QB Equations
- Same or better values (improved RMSE) for 80 of 82 indicators
  - QB Equations are sensitive to correct error estimates

  - More details: http://www.stefanbischof.at/slides/Rigorosum_Bischof.pdf

City Data Pipeline Prototype

citydata.wu.ac.at

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


Sustainable Cities Results
http://citydata.ai.wu.ac.at/KPIDataPipeline/KPIDispatcher

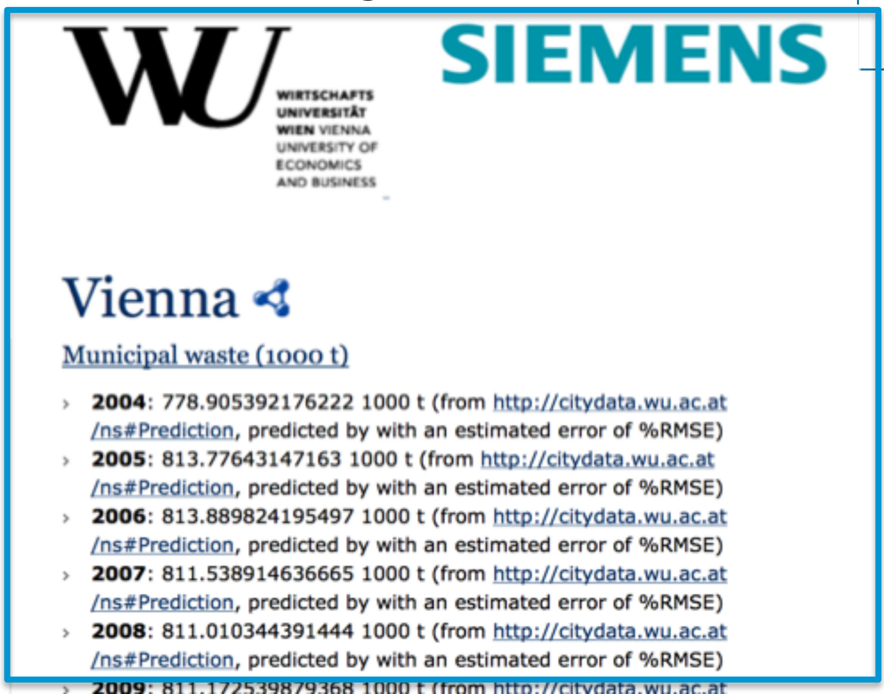
WU WIRTSCHAFTSUNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS
SIEMENS

Berlin

Population male 2012
1717645.0 persons
(Source: http://epp.eurostat.ec.europa.eu/)
Population male 2011
1695438.0 persons (Source: http://data.un.org/)
Population male 2011
1695438.0 persons
(Source: http://epp.eurostat.ec.europa.eu/)
Population male 2010
1686256.0 persons
(Source: http://epp.eurostat.ec.europa.eu/)
Population male 2009
1686256.0 persons

Vienna

Population male 2011
821605.0 persons (Source: http://data.un.o
Population male 2010
812867.0 persons (Source: http://data.un.o
Population male 2009
807088.0 persons (Source: http://data.un.o
Population male 2009
807088.0 persons
(Source: http://epp.eurostat.ec.europa.eu/)
Population male 2008
801776.0 persons (Source: http://data.un.o
Population male 2008
800361.0 persons


WU WIRTSCHAFTSUNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS
SIEMENS

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**: 811.53891463665 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)
- **2009**: 811.172539879368 1000 t (from http://citydata.wu.ac.at

...it's not finished, but:
assumption: Predictions get better, the more Open data we integrate...

However:

(Strong)Limitations:

- We combined 3-4 specific OD sources (there are 100s of Open Data Portals out there)
- We manually created an ontology for mapping those sources and set of equations from Eurostat?

Open Questions:

- How can I build a scalable repository of Open Data?
- How can I automate finding relevant data?
- How can I automatize building an Open Data Knowledge graph?

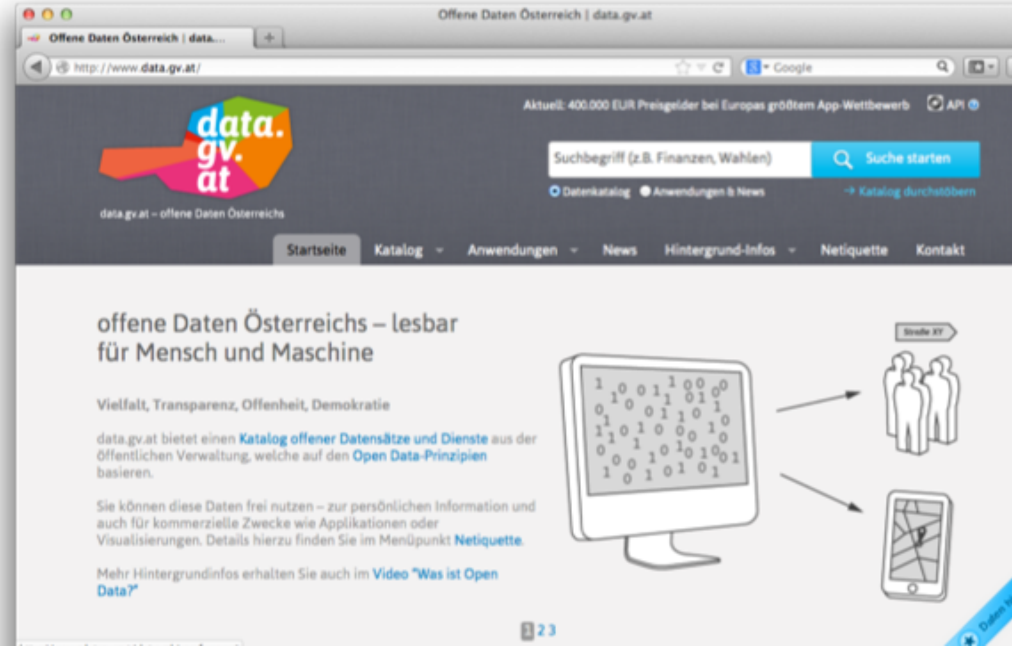
Open Data Portals

CKAN … http://ckan.org/

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

- http://opendataportal.at/
- http://data.gv.at/

- machine-processable? …
  … **partially**



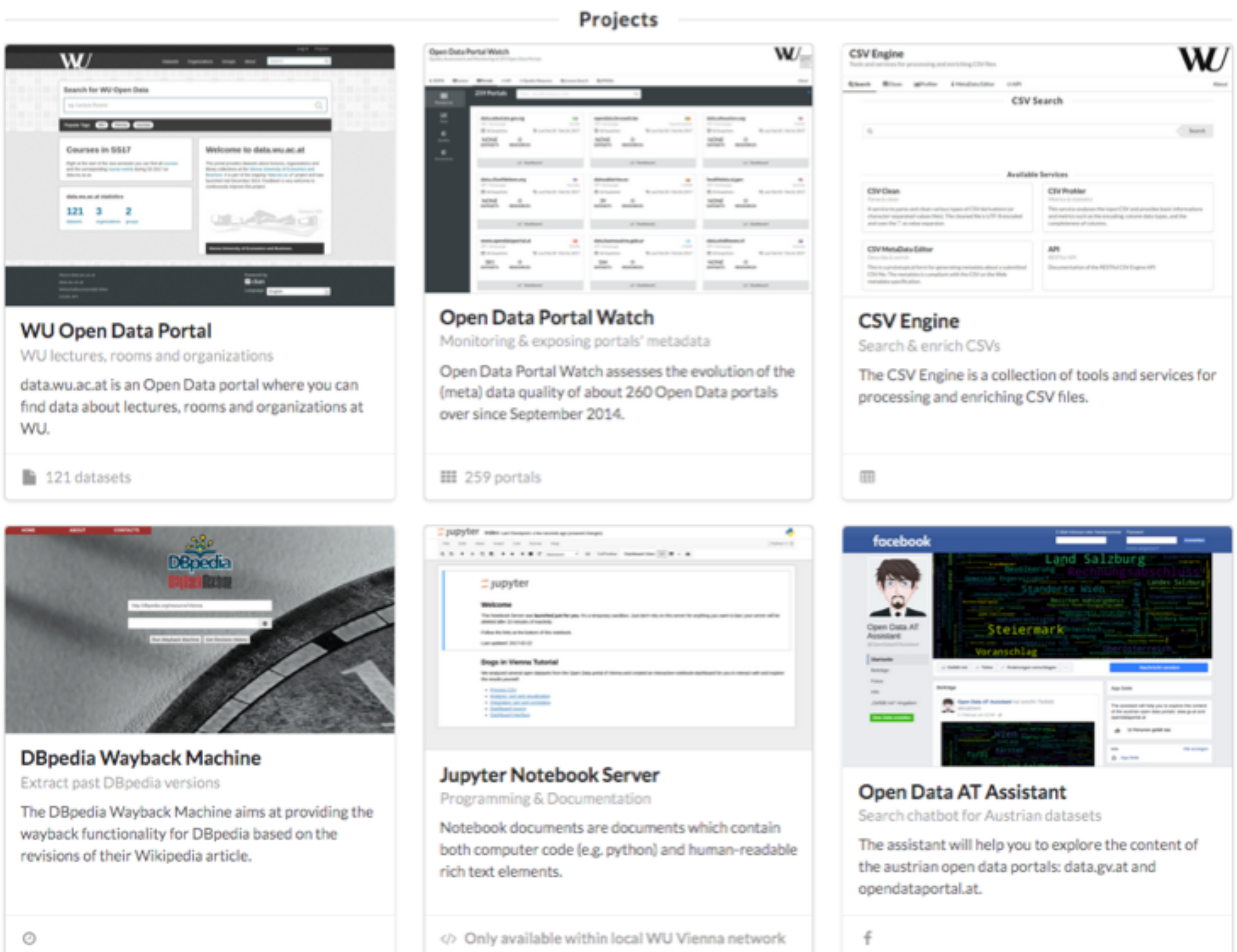
Our ongoing research:
data.wu.ac.at

- ***What is the status of Open Data and what are the challenges using Open Data?***
  - OpenData PortalWatch – a project at WU
  - Improving and assessing Open **Data Quality** : ADEQUATE (FFG)

- ***What's next?***
  - Making Open Data Searchable
  - Building an Open Data **Knowledge Graph**!

Ongoing Projects (data.wu.ac.at)

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS



EFMD EQUIS ACCREDITED

Open Data Portal Watch

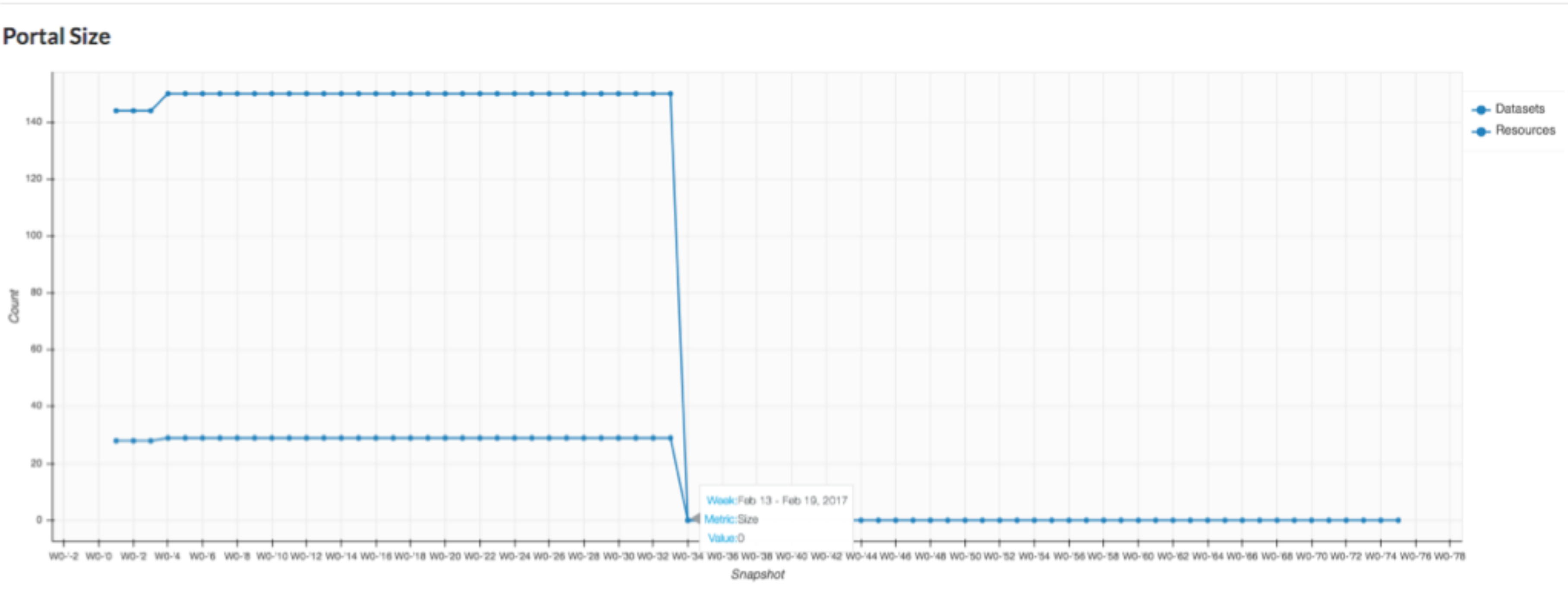
http://data.wu.ac.at/portalwatch/

- Periodically monitoring a list of Open Data Portals
  - 260 CKAN powered Open Data Portals worldwide
- Quality assessment
- Evolution tracking
  - Meta data
  - Data
  - Formats, growth

Portalwatch Example:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

http://data.wu.ac.at/portalwatch/portal/open_whitehouse_gov/1804/



EFMD EQUIS ACCREDITED

Portalwatch Example:

http://data.wu.ac.at/portalwatch/portal/open_whitehouse_gov/1804/

A

Automated Quality Assessment of Metadata across Open Data Portals

SEBASTIAN NEUMAIER, Vienna University of Economics and Business
JÜRGEN UMBRICH, Vienna University of Economics and Business
AXEL POLLERES, Vienna University of Economics and Business

The Open Data movement has become a driver for publicly available data on the Web. More and more data – from governments, public institutions but also from the private sector – is made available online and is mainly published in so called Open Data portals. However, with the increasing number of published resources, there are a number of concerns with regards to the quality of the data sources and the corresponding metadata, which compromise the searchability, discoverability and usability of resources.

In order to get a more complete picture of the severity of these issues, the present work aims at developing a generic metadata quality assessment framework for various Open Data portals: we treat data portals independently from the portal software frameworks by mapping the specific metadata of three widely used portal software frameworks (CKAN, Socrata, OpenDataSoft) to the standardized DCAT metadata schema. We subsequently define several quality metrics, which can be evaluated automatically and in a efficient manner. Finally, we report findings based on monitoring a set of over 260 Open Data portals with 1.1M datasets. This includes the discussion of general quality issues, e.g. the retrievability of data, and the analysis of our specific quality metrics.

CCS Concepts: •**General and reference** → **Measurement; Metrics;** •**Information systems** → *Web searching and information discovery;* Digital libraries and archives;

Additional Key Words and Phrases: Open Data, quality assessment, data quality, data portal

ACM Reference Format:
Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres, 2015. Automated Quality Assessment of Metadata across Open Data Portals. *ACM J. Data Inform. Quality* V, N, Article A (January YYYY), 29 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

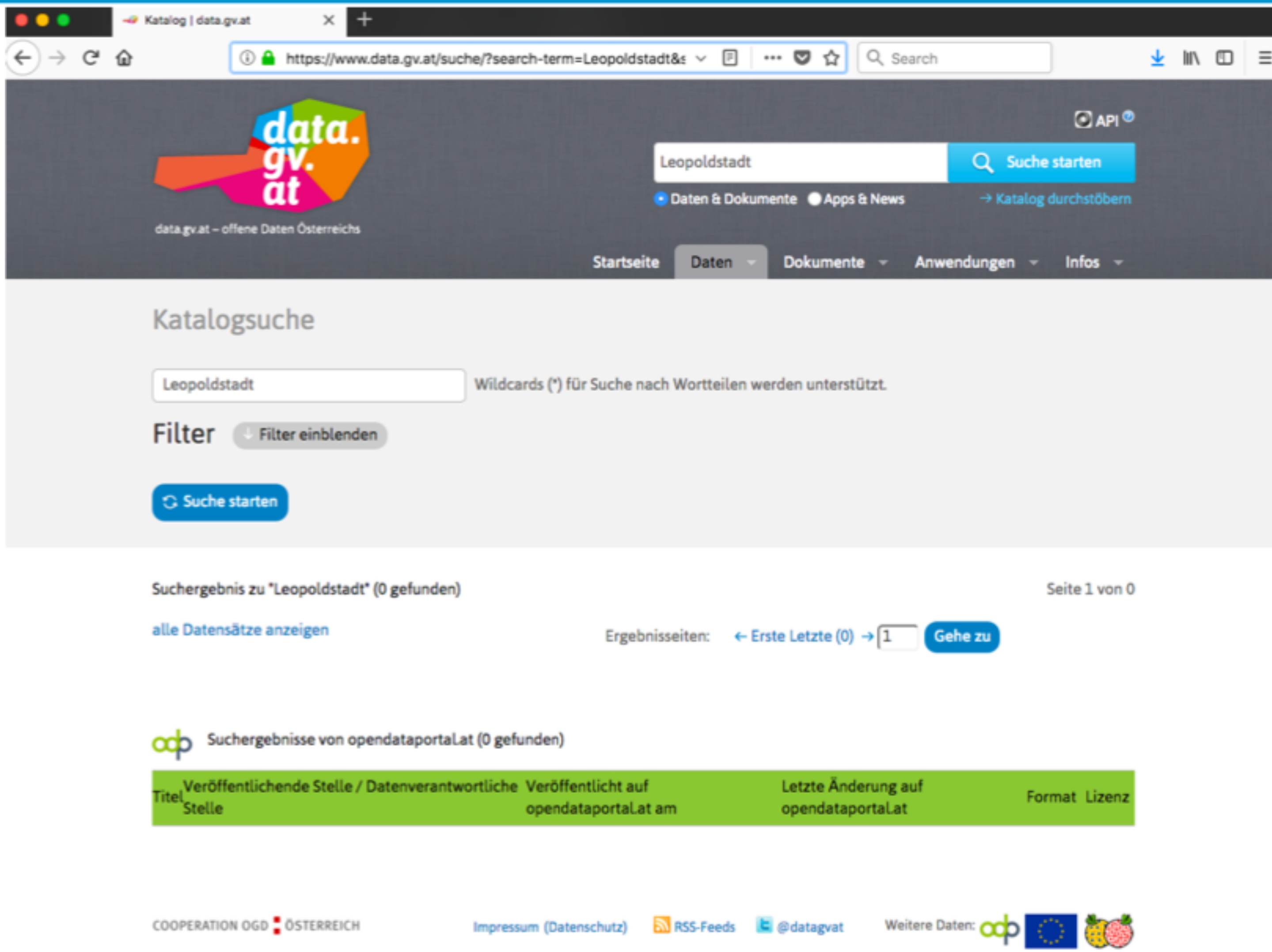
Our research:
data.wu.ac.at

- ***What is the status of Open Data and what are the challenges using Open Data?***
  - OpenData PortalWatch – a project at WU
  - Improving and assessing Open Data Quality: ADEQUATE (FFG)

- ***What's next?***
  - Making Open Data Searchable
  - Building an Open Data **Knowledge Graph**!

EFMD EQUIS ACCREDITED

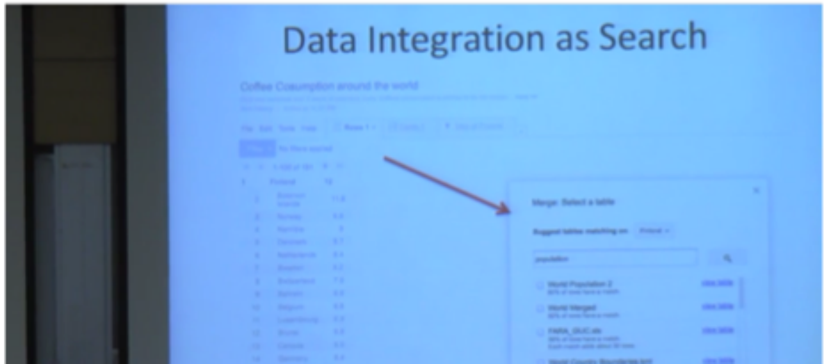
Why is Search in Open Data a problem?



Why is Search in Open Data a problem?

https://www.youtube.com/watch?v=kCAymmbYIvc

Structured Data in Web Search by Alon Halevy



VS.

B	C	D	E	F	G	H	I
NUTS2	NUTS3	DISTRICT_CODE	SUB_DISTRICT_CODE	POP_TOTAL	POP_MEN	POP_WOMEN	REF_DATE
AT13	AT130	90101	0	16131	7726	8405	01.01.2014
AT13	AT130	90201	0	99597	48650	50947	01.01.2014
AT13	AT130	90301	0	86454	41085	45369	01.01.2014
AT13	AT130	90401	0	31452	14903	16549	01.01.2014
AT13	AT130	90501	0	53610	26299	27311	01.01.2014
AT13	AT130	90601	0	30613	14833	15780	01.01.2014
AT13	AT130	90701	0	30792	14703	16089	01.01.2014
AT13	AT130	90801	0	24279	11855	12424	01.01.2014
AT13	AT130	90901	0	40528	19286	21242	01.01.2014
AT13	AT130	91001	0	186450	91638	94812	01.01.2014
AT13	AT130	91101	0	93440	45541	47899	01.01.2014
AT13	AT130	91201	0	90874	43752	47122	01.01.2014

Open Data Search is hard…

a) No natural language „cues" like in Web tables…

b) Existing knowledge graphs don't cover the domain of "Open Data"

c) Open Data is not properly geo-referenced

Some starting points:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- First baby steps on building an Open Data Knowledge Graph:
- Ongoing work to make Open Data **geo-searchable** e.g. in our project communidata.at (just submitted to ESWC)

International Semantic Web conference 2016:

Multi-level semantic labelling of numerical values

Sebastian Neumaier[1], Jürgen Umbrich[1], Josiane Xavier Parreira[2], and Axel Polleres[1]

[1] Vienna University of Economics and Business, Vienna, Austria
[2] Siemens AG Österreich, Vienna, Austria

Abstract. With the success of Open Data a huge amount of tabular data sources became available that could potentially be mapped and linked into the Web of (Linked) Data. Most existing approaches to "semantically label" such tabular data rely on mappings of textual information to classes, properties, or instances in RDF knowledge bases in order to link – and eventually transform – tabular data into RDF. However, as we will illustrate, Open Data tables typically contain a large portion of numerical columns and/or non-textual headers; therefore solutions that solely focus on textual "cues" are only partially applicable for mapping such data sources. We propose an approach to find and rank candidates of semantic labels and context descriptions for a given bag of numerical values. To this end, we apply a hierarchical clustering over information taken from DBpedia to build a background knowledge graph of possible "semantic contexts" for bags of numerical values, over which we perform a nearest neighbour search to rank the most likely candidates. Our evaluation shows that our approach can assign fine-grained semantic labels, when there is enough supporting evidence in the background knowledge graph. In other cases, our approach can nevertheless assign high level contexts to the data, which could potentially be used in combination with other approaches to narrow down the search space of possible labels.

EFMD EQUIS ACCREDITED

Towards linking Open Data to a Knowledge Graph

- Attempt to link numeric Open data to the dbpedia knowledge graph…

International Semantic Web conference 2016:

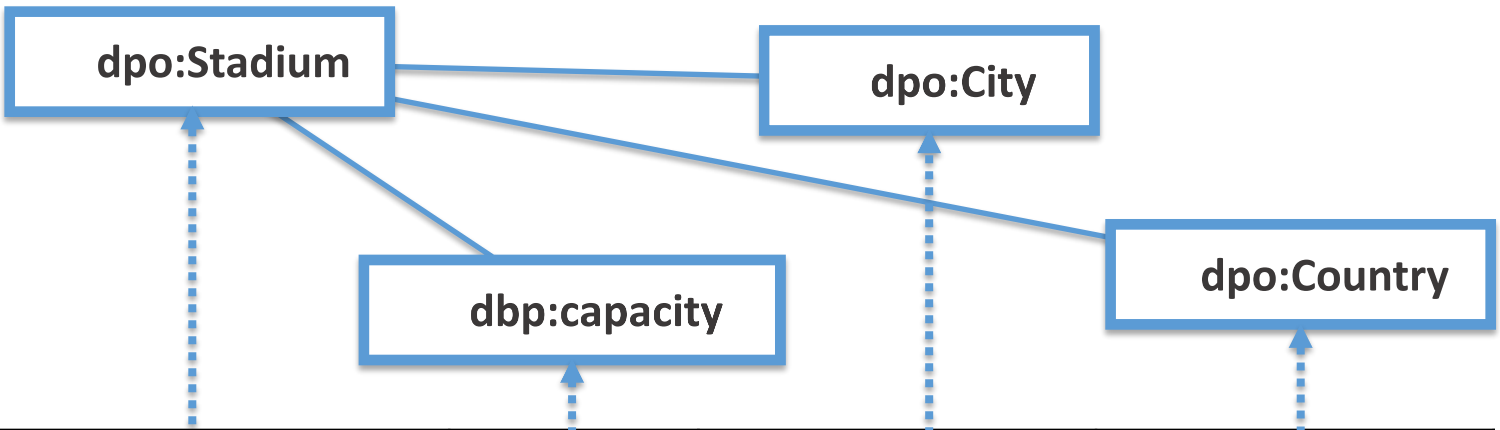
Multi-level semantic labelling of numerical values

Sebastian Neumaier[1], Jürgen Umbrich[1], Josiane Xavier Parreira[2], and Axel Polleres[1]

[1] Vienna University of Economics and Business, Vienna, Austria
[2] Siemens AG Österreich, Vienna, Austria

Abstract. With the success of Open Data a huge amount of tabular data sources became available that could potentially be mapped and linked into the Web of (Linked) Data. Most existing approaches to "semantically label" such tabular data rely on mappings of textual information to classes, properties, or instances in RDF knowledge bases in order to link – and eventually transform – tabular data into RDF. However, as we will illustrate, Open Data tables typically contain a large portion of numerical columns and/or non-textual headers; therefore solutions that solely focus on textual "cues" are only partially applicable for mapping such data sources. We propose an approach to find and rank candidates of semantic labels and context descriptions for a given bag of numerical values. To this end, we apply a hierarchical clustering over information taken from DBpedia to build a background knowledge graph of possible "semantic contexts" for bags of numerical values, over which we perform a nearest neighbour search to rank the most likely candidates. Our evaluation shows that our approach can assign fine-grained semantic labels, when there is enough supporting evidence in the background knowledge graph. In other cases, our approach can nevertheless assign high level contexts to the data, which could potentially be used in combination with other approaches to narrow down the search space of possible labels.

Example



stadium name	***capacity***	***city***	***country***
Emirates Stadium	60361	London	England
Villa Park	42785	Birmingham	England
Ewood Park	31154	Blackburn	England
…	…	…	…

But:

Web/HTML tables differ from typical Open Data tables:

- **Domain**: e.g., public administration data, statistical data, weather data, elections, …
- **Structure**: OD tables contain large amount of numerical columns

Wohnungen in den 250 Zaehlbezirken in Wien - Registerzaehlung 2011 | Housing units in 250 su					
NUTS1	NUTS2	NUTS3	DISTRICT_CODE	SUB_DISTRICT_CODE	WHG_TOTAL
AT1	AT13	AT130	90100	90101	3004
AT1	AT13	AT130	90100	90102	1049
AT1	AT13	AT130	90100	90103	1389
AT1	AT13	AT130	90100	90104	1014
AT1	AT13	AT130	90100	90105	1337
AT1	AT13	AT130	90100	90106	1915
AT1	AT13	AT130	90100	90107	2032
AT1	AT13	AT130	90200	90201	5178
AT1	AT13	AT130	90200	90202	6345
AT1	AT13	AT130	90200	90203	7549
AT1	AT13	AT130	90200	90204	8388
AT1	AT13	AT130	90200	90205	5358
AT1	AT13	AT130	90200	90206	4237
AT1	AT13	AT130	90200	90207	7812
AT1	AT13	AT130	90200	90208	1478
AT1	AT13	AT130	90200	90209	7547

Example (Cont'd)

stadium	***capacity***	***city***	***country***
Emirates Stadium	60361	London	England
Villa Park	42785	Birmingham	England
Ewood Park	31154	Blackburn	England
…	…	…	…

Example (Cont’d)

	TOTAL	*DISTRICT_CODE*	*ISO_2*
Emirates Stadium	60361	SW1A 0AA	GB
Villa Park	42785	B23 7QG	GB
Ewood Park	31154	B26 6QA	GB
…	…	…	…

Why not use numeric values?

- Identifying the most likely semantic label for a bag of numerical values
- Deliberately ignore surroundings

	TOTAL	*DISTRICT_CODE*	*ISO_2*
Emirates Stadium	60361	SW1A 0AA	GB
Villa Park	42785	B23 7QG	GB
Ewood Park	31154	B26 6QA	GB
...	...	...	...

Why not use numeric values?

- Identifying the most likely semantic label for a bag of numerical values
- Deliberately ignore surroundings

Emirates Stadium	60361	SW1A 0AA	GB
Villa Park	42785	B23 7QG	GB
Ewood Park	31154	B26 6QA	GB
…	…	…	…

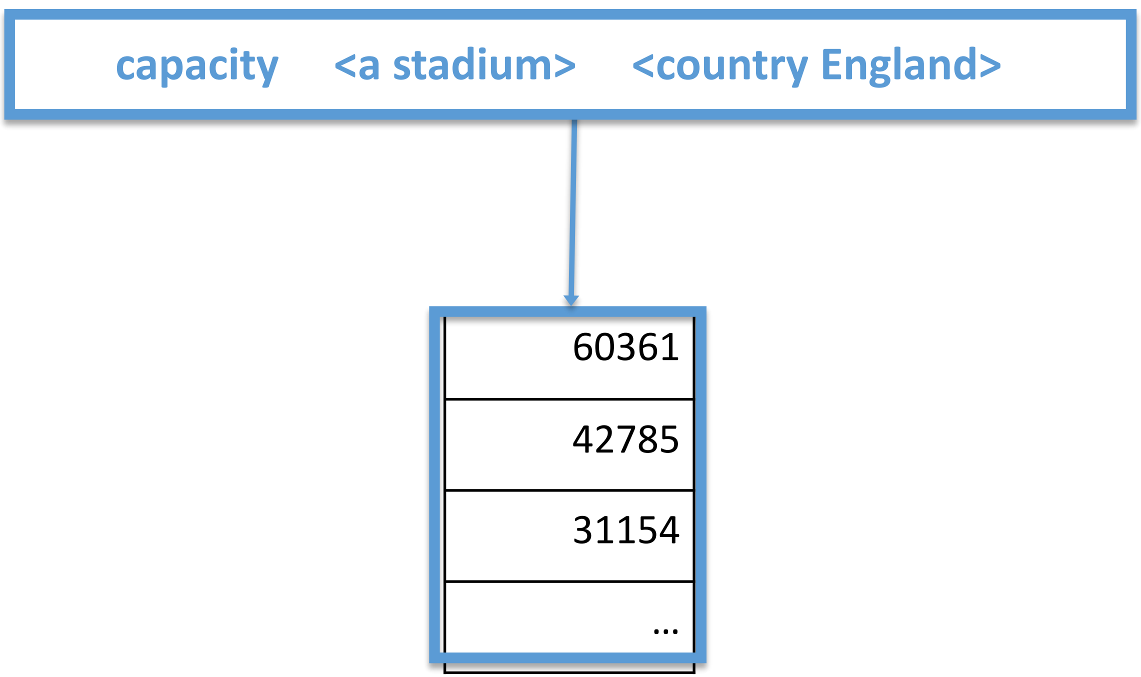
Why not use numeric values?

- Identifying the most likely semantic label for a bag of numerical values
- Deliberately ignore surroundings

60361
42785
31154
…

Why not use numeric values?

- Identifying the most likely semantic label for a bag of numerical values
- Deliberately ignore surroundings

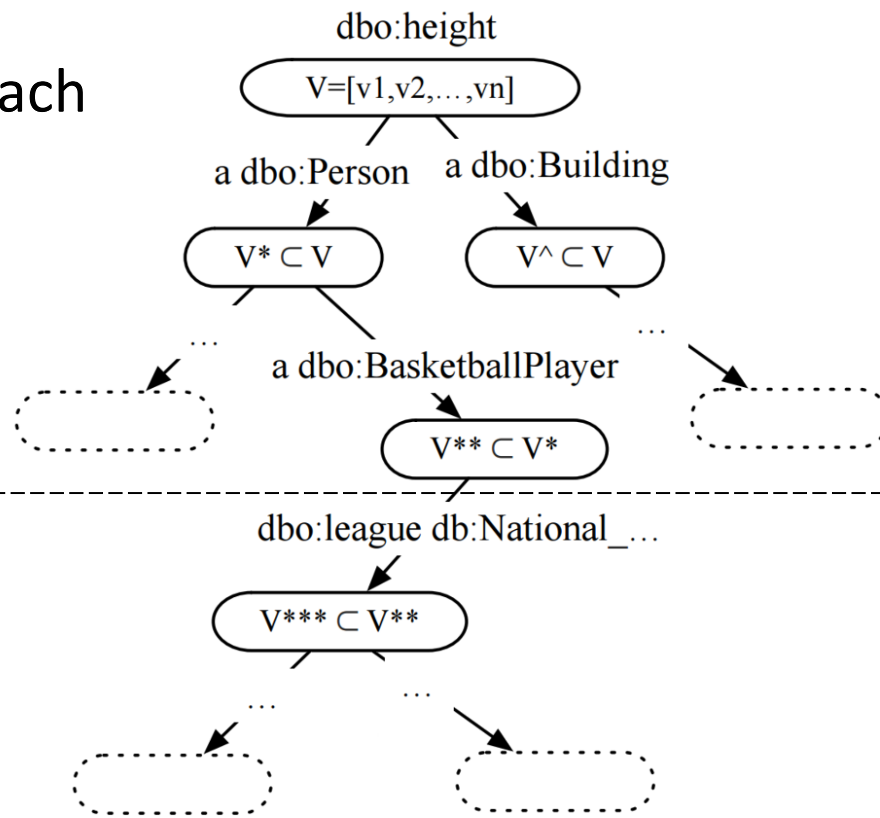


Our Approach

1. **Hierarchical clustering** over an RDF knowledge base
    - to build background knowledge graph (**BKG**)
    - nodes consist of **typical numerical values**, annotated with context information, i.e.:
      grouped by **properties** and their **shared domain (subject) pairs**
2. k-nearest neighbors search
3. **Aggregation of the results** at different levels to find the most likely context:
    - property
    - type
    - context

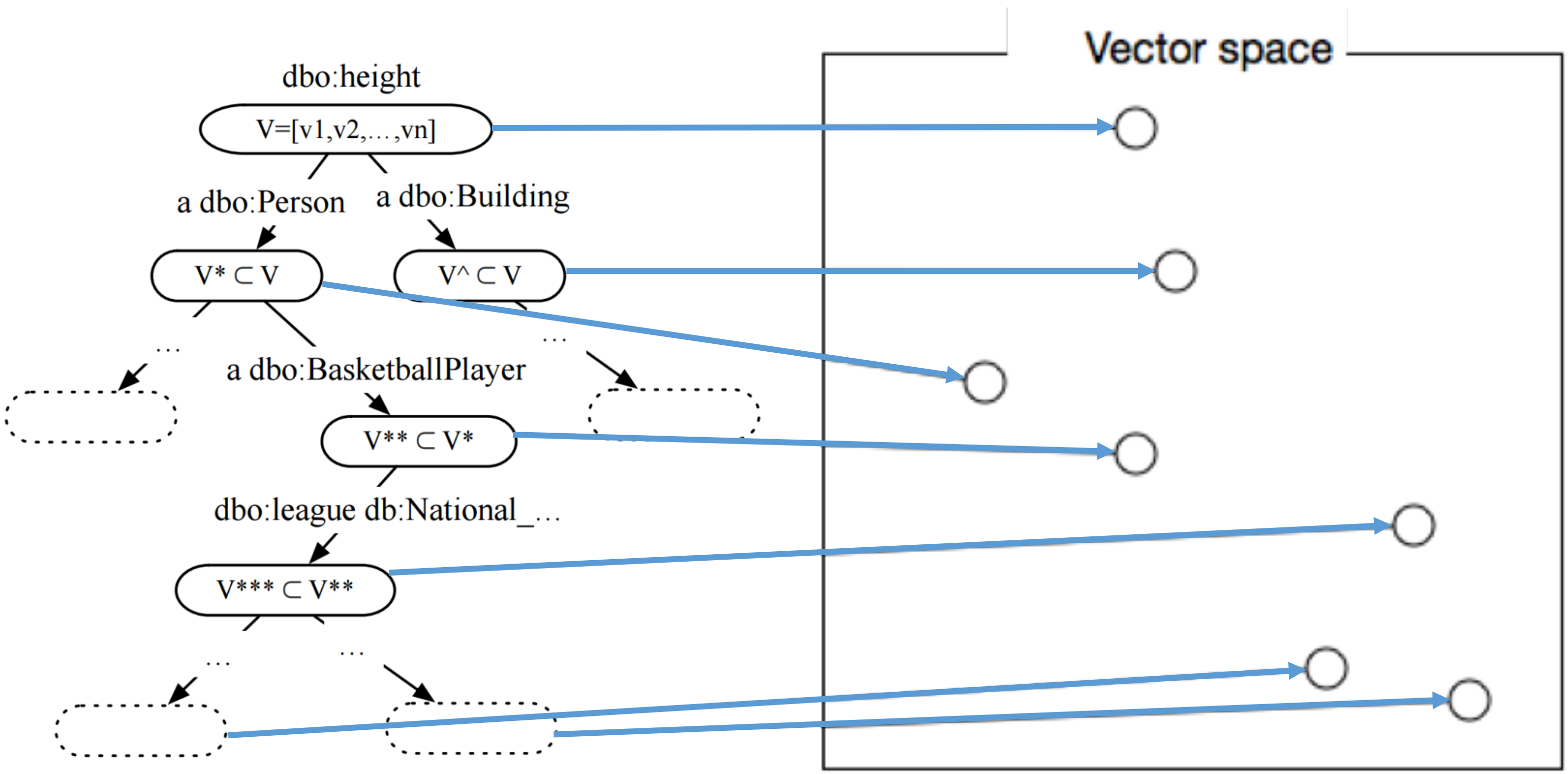
1. Background Knowledge Graph

- Find properties with **numerical range**
- Hierarchical clustering approach
- Two hierarchical layers:
    - **Type** hierarchy (using OWL classes)
    - **Property-object** hierarchy (shared property-object pairs)



2. *k*-Nearest neighbor search

Mapping bags of numerical value to vector space (feature vector)



Towards linking Open Data to a Knowledge Graph

- Attempt to link numeric Open data to the dbpedia knowledge graph…
    - Some Caveats:
        - Method works well if you have a suitable knowledge graph, but:
        - ***Open Data has a lot of attributes that do not match current knowledge graphs … like these:***

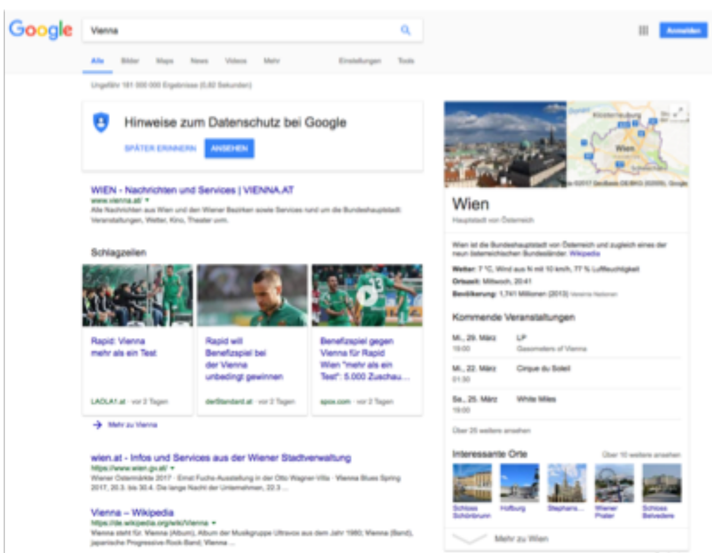
International Semantic Web conference 2016:

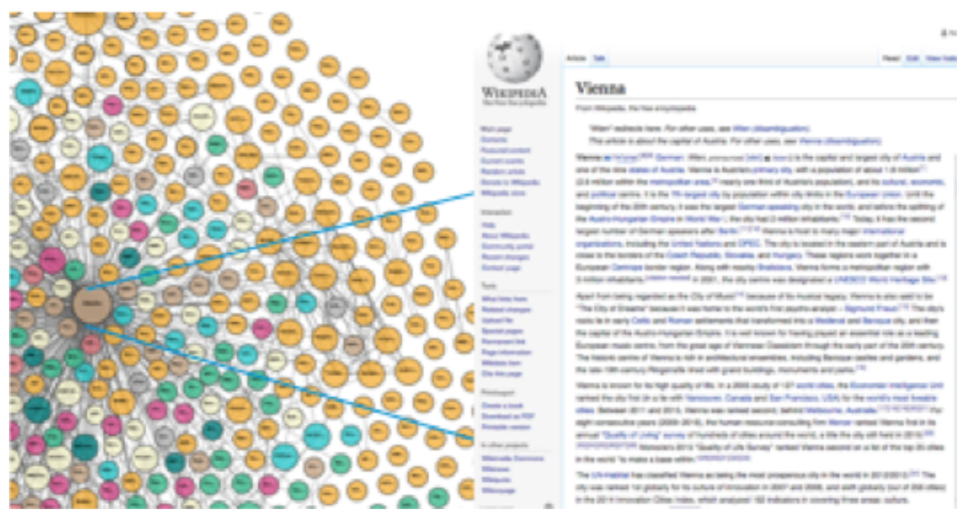
Multi-level semantic labelling of numerical values

Sebastian Neumaier[1], Jürgen Umbrich[1], Josiane Xavier Parreira[2], and Axel Polleres[1]

[1] Vienna University of Economics and Business, Vienna, Austria
[2] Siemens AG Österreich, Vienna, Austria

Abstract. With the success of Open Data a huge amount of tabular data sources became available that could potentially be mapped and linked into the Web of (Linked) Data. Most existing approaches to "semantically label" such tabular data rely on mappings of textual information to classes, properties, or instances in RDF knowledge bases in order to link – and eventually transform – tabular data into RDF. However, as we will illustrate, Open Data tables typically contain a large portion of numerical columns and/or non-textual headers; therefore solutions that solely focus on textual "cues" are only partially applicable for mapping such data sources. We propose an approach to find and rank candidates of semantic labels and context descriptions for a given bag of numerical values. To this end, we apply a hierarchical clustering over information taken from DBpedia to build a background knowledge graph of possible "semantic contexts" for bags of numerical values, over which we perform a nearest neighbour search to rank the most likely candidates. Our evaluation shows that our approach can assign fine-grained semantic labels, when there is enough supporting evidence in the background knowledge graph. In other cases, our approach can nevertheless assign high level contexts to the data, which could potentially be used in combination with other approaches to narrow down the search space of possible labels.





Some starting points:

- First baby steps on building an Open Data Knowledge Graph:

International Semantic Web conference 2016:

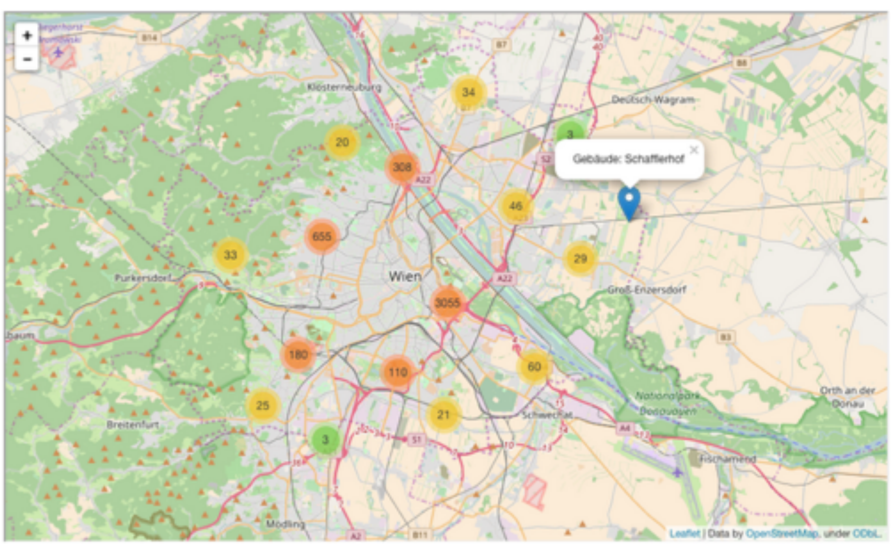
Multi-level semantic labelling of numerical values

Sebastian Neumaier[1], Jürgen Umbrich[1], Josiane Xavier Parreira[2], and Axel Polleres[1]

[1] Vienna University of Economics and Business, Vienna, Austria
[2] Siemens AG Österreich, Vienna, Austria

Abstract. With the success of Open Data a huge amount of tabular data sources became available that could potentially be mapped and linked into the Web of (Linked) Data. Most existing approaches to "semantically label" such tabular data rely on mappings of textual information to classes, properties, or instances in RDF knowledge bases in order to link – and eventually transform – tabular data into RDF. However, as we will illustrate, Open Data tables typically contain a large portion of numerical columns and/or non-textual headers; therefore solutions that solely focus on textual "cues" are only partially applicable for mapping such data sources. We propose an approach to find and rank candidates of semantic labels and context descriptions for a given bag of numerical values. To this end, we apply a hierarchical clustering over information taken from DBpedia to build a background knowledge graph of possible "semantic contexts" for bags of numerical values, over which we perform a nearest neighbour search to rank the most likely candidates. Our evaluation shows that our approach can assign fine-grained semantic labels, when there is enough supporting evidence in the background knowledge graph. In other cases, our approach can nevertheless assign high level contexts to the data, which could potentially be used in combination with other approaches to narrow down the search space of possible labels.

- Ongoing work to make Open Data **geo-searchable** e.g. in our project communidata.at



Sneak preview (just submitted to ESWC):

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

Still Open Questions (with some starting points presented…)

- How can I build a scalable repository of Open Data?
- How can I automate finding relevant data?
- (How) can I automatize
  - cleansing of metadata
  - building an Open Data Knowledge graph?

- What is the **right form of Knowledge Representation** for Knowledge graphs?
  - OWL, Rules, Equations, Property-domain pairs?)
  - How to represent models in an exchangeable manner?

- Eventually: How can I enable fact checking, verify information on the Web, understand cities,… by Open Data?

Collaborators/Current Team:

What I talked about →

Dr. Stefan Bischof
(City Data Pipeline)

Sebastian Neumaier
(**OpenData Quality**,
Knowledge Graphs)

Dr. Jürgen Umbrich
(Search, Crawling,
Knowledge Graphs)

What I tdidn't
yet talk about ↓

Dr. Sabrina Kirrane
(**Policies**, Privacy, Access Control)

Dr. Javier Fernandez
(**Compression, HDT, Archiving Indexing, Query Processing**)

Svitlana Vakulenko
(NLP, event detection, social media analysis)

Erwin Filtz
(Legal Knowledge Graphs, Graph Data Processing)

Dr. Vadim Savenkov
(Database Updates, OBDM, Open Data)

Simon Steyskal
(Policies ODRL, Constraints, SHACL)

Martin Beno
(Open Data, Server Admin)

Giray Havur
(Business Processes, Resource allocation, Constraints/ Logic Programming)

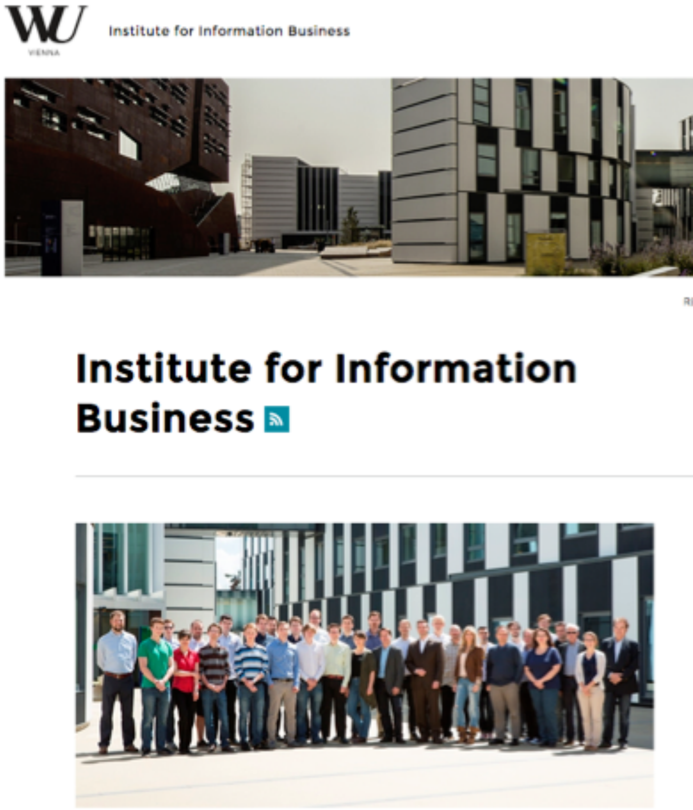
Thanks! Things I did NOT have time to talk about in detail, but would be interested to talk about collaborations:

- Linked/Open Data Monitoring/Archiving, Temporal querying → (Jürgen, Javier)

- RDF Query Processing, Path queries and Updates (Vadim)

- Privacy and data on the Web, Licenses
  - → http://privacylab.at
  - → http://specialprivacy.eu/
  - → https://dalicc.net/

  …



https://www.wu.ac.at/en/infobiz/