Integrating Open Data: (How) Can Description Logics Help me?

Axel Polleres

web: http://polleres.net

twitter: @AxelPolleres

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

What is Open Data?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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, […] 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

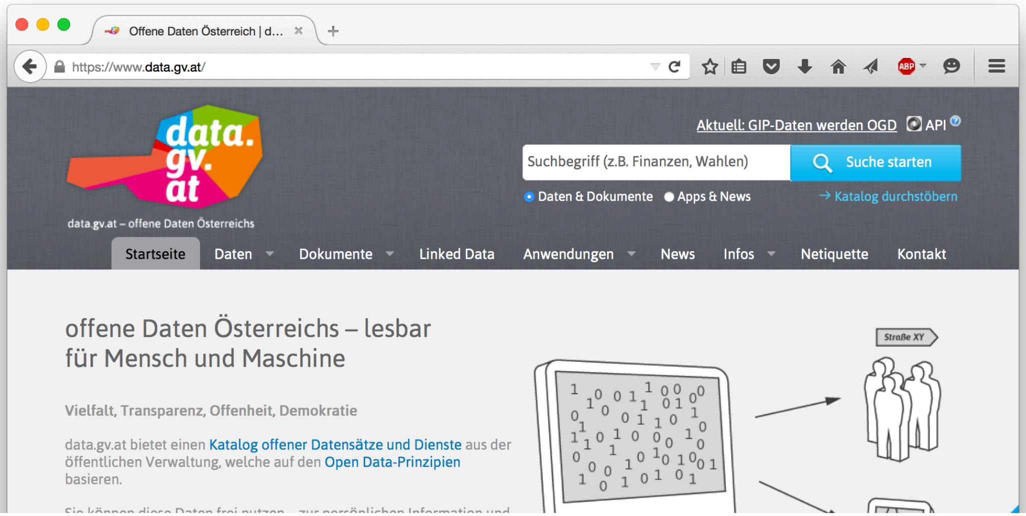
Universal Participation: everyone must be able to use, reuse and redistribute – […] no discrimination against fields of endeavour, persons or groups. For example, no 'non-commercial' […]restrictions.

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

Open Knowledge Foundation

Open Data is a global trend:

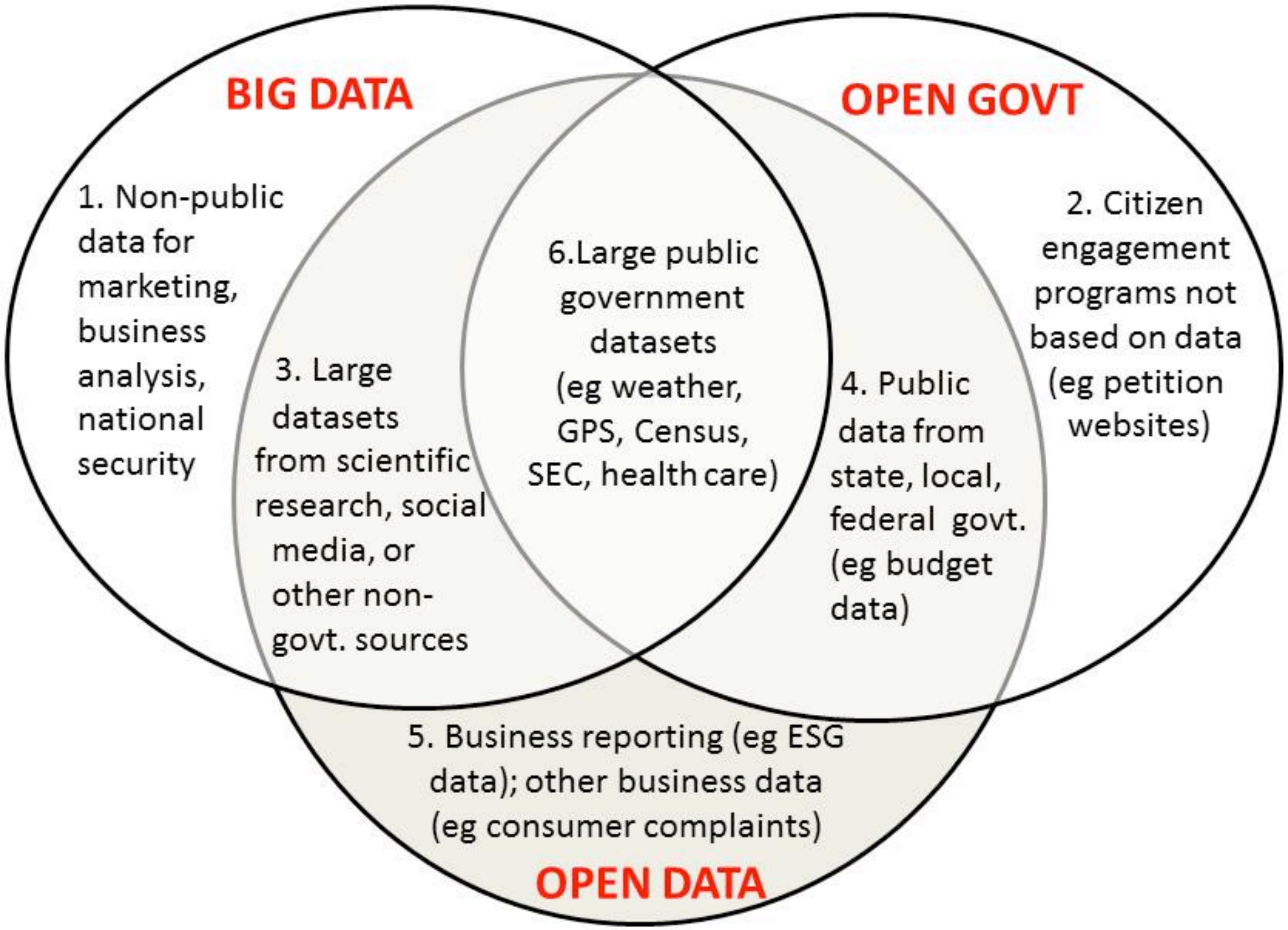
- Cities, International Organizations, National and European **portals**, etc.:



Buzzword Bingo 1/3: Open Data vs. Big Data

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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



Buzzword Bingo 2/3: Open Data vs. Big Data

- **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 to describe datasets
  - Different data formats

- **Velocity:**
  - Open Data changes regularly (fast and slow)
  - New datasets appear, old ones disappear

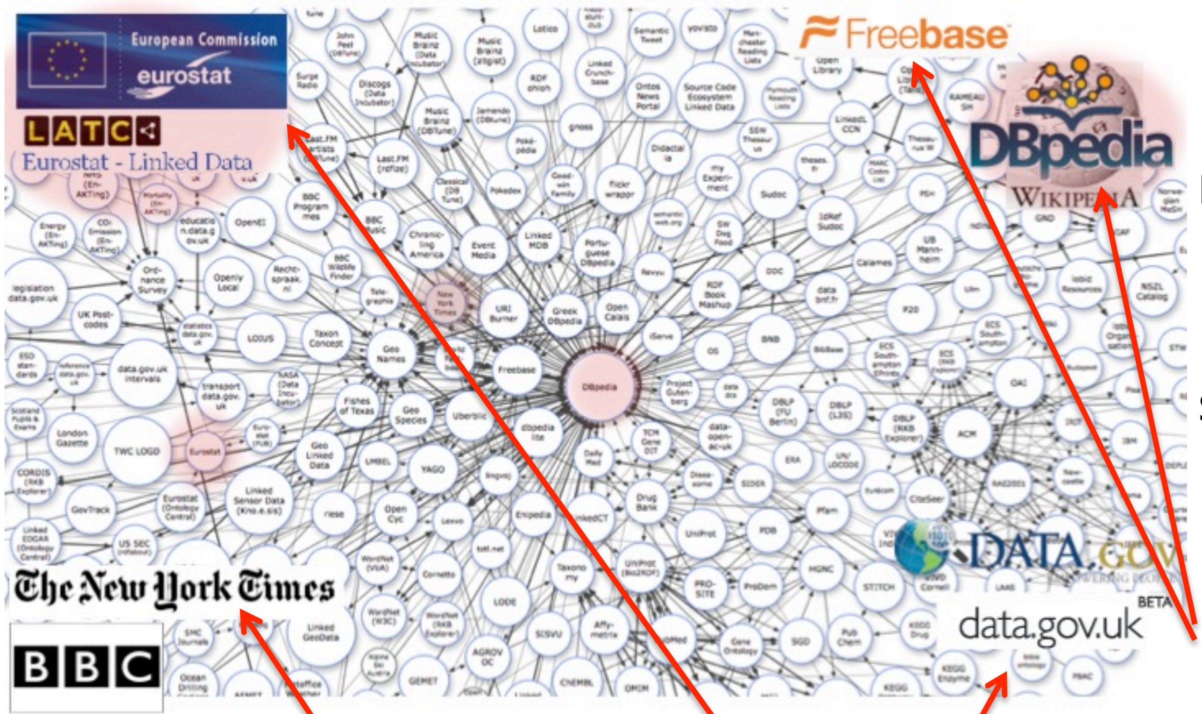
Buzzword Bingo 3/3: Open Data vs. Linked Data

This talk is NOT about DL Reasoning over Linked Data:

Status: OWLED2013 talk …

7

Linked Data on the Web: Adoption



LOD is till growing, but OD is growing faster and challenges aren't necessarily the exactly same…

So. let's focus on Open Data in general…

Alternatives in the meantime: (wikidata...)

LD efforts discontinued?!

LOD in OGD growing, but slowly

Now: Can ontological reasoning help me to integrate Open Data?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

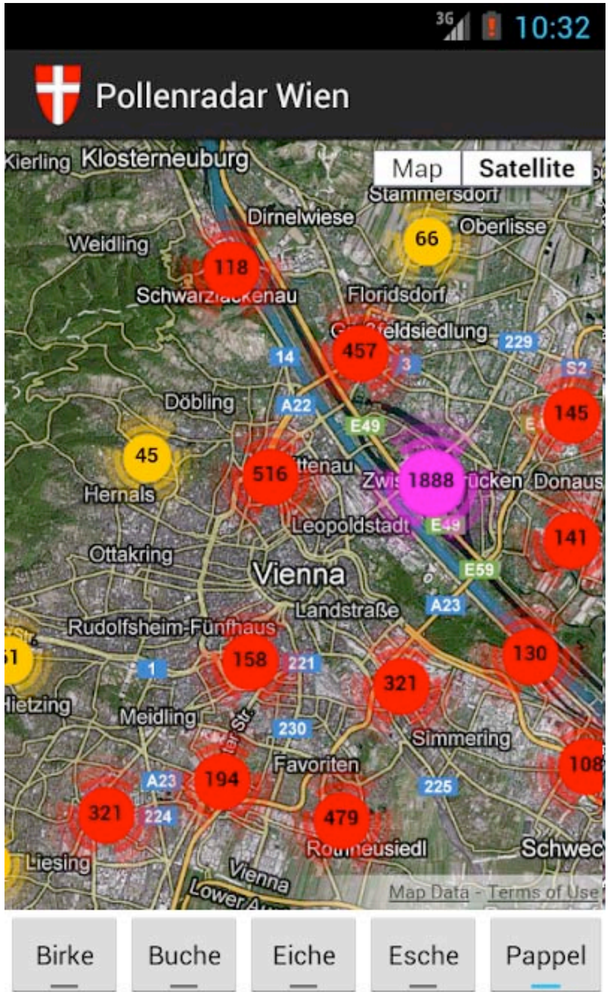
short answer: yes, but ...

long answer: no, but ...

In more detail:

- Is Open Data useful at all?
- Are ontology languages expressive enough?
- Which ontologies could I use?
- Is there enough data at all?
- How to tackle inconsistencies?
- Where to find the right data?

Is Open Data useful at all?
Beyond "single dataset Apps"...



Great stuff, but limited potential...

Is Open Data useful at all?
A concrete use case:

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

The results

The complete results from the index, including the overall result of each city as well as the individual rankings within the eight categories.

Overall

	City	Score
1	Copenhagen	87,31
2	Stockholm	86,65
3	Oslo	83,98
4	Vienna	83,34
5	Amsterdam	83,03
6	Zurich	82,31
7	Helsinki	79,29
8	Berlin	79,01
9	Brussels	78,01
10	Paris	73,21
11	London	71,56
12	Madrid	67,08
13	Vilnius	62,77
14	Rome	62,58
15	Riga	59,57
16	Warsaw	59,04
17	Budapest	57,55
18	Lisbon	57,25
19	Ljubljana	56,39
20	Bratislava	56,09
21	Dublin	53,98
22	Athens	53,09
23	Tallinn	52,98
24	Prague	49,78
25	Istanbul	45,20
26	Zagreb	42,36
27	Belgrade	40,03
28	Bucharest	39,14
29	Sofia	36,85
30	Kiev	32,33

CO_2

	City	Score
1	Oslo	9,58
2	Stockholm	8,99
3	Zurich	8,48
4	Copenhagen	8,35
5	Brussels	8,32
6	Paris	7,81
7	Rome	7,57
8	Vienna	7,53
9	Madrid	7,51
10	London	7,34
11	Helsinki	7,30
12	Amsterdam	7,10
13	Berlin	6,75
14	Ljubljana	6,67
15	Riga	5,55
16	Istanbul	4,86
=17	Athens	4,85
=17	Budapest	4,85
19	Dublin	4,77
20	Warsaw	4,65
21	Bratislava	4,54
22	Lisbon	4,05
23	Vilnius	3,91
24	Bucharest	3,65
25	Prague	3,44
26	Tallinn	3,40
27	Zagreb	3,20
28	Belgrade	3,15
29	Sofia	2,95
30	Kiev	2,49

Energy

	City	Score
1	Oslo	8,71
2	Copenhagen	8,69
3	Vienna	7,76
4	Stockholm	7,61
5	Amsterdam	7,08
6	Zurich	6,92
7	Rome	6,40
8	Brussels	6,19
9	Lisbon	5,77
10	London	5,64
11	Istanbul	5,55
12	Madrid	5,52
13	Berlin	5,48
14	Warsaw	5,29
15	Athens	4,94
16	Paris	4,66
17	Belgrade	4,65
18	Dublin	4,55
19	Helsinki	4,49
20	Zagreb	4,34
21	Bratislava	4,19
22	Riga	3,53
23	Bucharest	3,42
24	Prague	3,26
25	Budapest	2,43
26	Vilnius	2,39
27	Ljubljana	2,23
28	Sofia	2,16
29	Tallinn	1,70
30	Kiev	1,50

Buildings

	City	Score
=1	Berlin	9,44
=1	Stockholm	9,44
3	Oslo	9,22
4	Copenhagen	9,17
5	Helsinki	9,11
6	Amsterdam	9,01
7	Paris	8,96
8	Vienna	8,62
9	Zurich	8,43
10	London	7,96
11	Lisbon	7,34
12	Brussels	7,14
13	Vilnius	6,91
14	Sofia	6,25
15	Rome	6,16
16	Warsaw	5,99
17	Madrid	5,68
18	Riga	5,43
19	Ljubljana	5,20
20	Budapest	5,01
21	Bucharest	4,79
22	Athens	4,36
23	Bratislava	3,54
24	Dublin	3,39
25	Zagreb	3,29
26	Prague	3,14
27	Belgrade	2,89
28	Istanbul	1,51
29	Tallinn	1,06
30	Kiev	0,00

Transport

	City	Score
1	Stockholm	8,81
2	Amsterdam	8,44
3	Copenhagen	8,29
4	Vienna	8,00
5	Oslo	7,92
6	Zurich	7,83
7	Brussels	7,49
8	Bratislava	7,16
9	Helsinki	7,08
=10	Budapest	6,64
=10	Tallinn	6,64
12	Berlin	6,60
13	Ljubljana	6,17
14	Riga	6,16
15	Madrid	6,01
16	London	5,55
17	Athens	5,48
18	Rome	5,31
=19	Kiev	5,29
=19	Paris	5,29
=19	Vilnius	5,29
=19	Zagreb	5,29
23	Istanbul	5,12
24	Warsaw	5,11
25	Lisbon	4,73
26	Prague	4,71
27	Sofia	4,62
28	Bucharest	4,55
29	Belgrade	3,98
30	Dublin	2,89

Water

	City	Score
1	Amsterdam	9,21
2	Vienna	9,13
3	Berlin	9,12
4	Brussels	9,05
=5	Copenhagen	8,88
=5	Zurich	8,88
7	Madrid	8,59
8	London	8,58
9	Paris	8,55
10	Prague	8,39
11	Helsinki	7,92
12	Tallinn	7,90
13	Vilnius	7,71
14	Bratislava	7,65
15	Athens	7,26
=16	Dublin	7,14
=16	Stockholm	7,14
18	Budapest	6,97
19	Rome	6,88
20	Oslo	6,85
21	Riga	6,43
22	Kiev	5,96
23	Istanbul	5,59
24	Lisbon	5,42
25	Warsaw	4,90
26	Zagreb	4,43
27	Ljubljana	4,19
28	Bucharest	4,07
29	Belgrade	3,90
30	Sofia	1,83

Waste and land use

	City	Score
1	Amsterdam	8,98
2	Zurich	8,82
3	Helsinki	8,69
4	Berlin	8,63
5	Vienna	8,60
6	Oslo	8,23
7	Copenhagen	8,05
8	Stockholm	7,99
9	Vilnius	7,31
10	Brussels	7,26
11	London	7,16
12	Paris	6,72
13	Dublin	6,38
14	Prague	6,30
15	Budapest	6,27
16	Tallinn	6,15
17	Rome	5,96
18	Ljubljana	5,95
19	Madrid	5,85
20	Riga	5,72
21	Bratislava	5,60
22	Lisbon	5,34
23	Athens	5,33
24	Warsaw	5,17
25	Istanbul	4,86
26	Belgrade	4,30
27	Zagreb	4,04
28	Bucharest	3,62
29	Sofia	3,32
30	Kiev	1,43

Air quality

	City	Score
1	Vilnius	9,37
2	Stockholm	9,35
3	Helsinki	8,84
4	Dublin	8,62
5	Copenhagen	8,43
6	Tallinn	8,30
7	Riga	8,28
8	Berlin	7,86
9	Zurich	7,70
10	Vienna	7,59
11	Amsterdam	7,48
12	London	7,34
13	Paris	7,14
14	Ljubljana	7,03
15	Oslo	7,00
16	Brussels	6,95
17	Rome	6,56
18	Madrid	6,52
19	Warsaw	6,45
20	Prague	6,37
21	Bratislava	5,96
22	Budapest	5,85
23	Istanbul	5,56
24	Lisbon	4,93
25	Athens	4,82
26	Zagreb	4,74
27	Bucharest	4,54
28	Belgrade	4,48
29	Sofia	4,45
30	Kiev	3,97

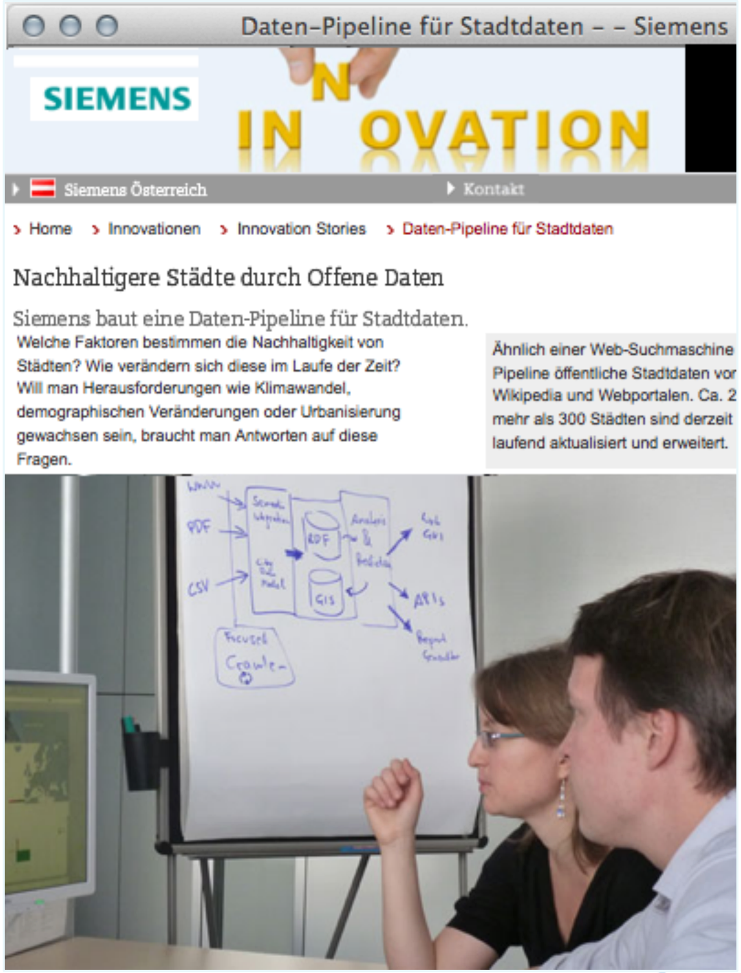
Environmental governance

	City	Score
=1	Brussels	10,00
=1	Copenhagen	10,00
=1	Helsinki	10,00
=1	Stockholm	10,00
=5	Oslo	9,67
=5	Warsaw	9,67
=7	Paris	9,44
=7	Vienna	9,44
9	Berlin	9,33
10	Amsterdam	9,11
11	Zurich	8,78
12	Lisbon	8,22
=13	Budapest	8,00
=13	Madrid	8,00
=15	Ljubljana	7,67
=15	London	7,67
17	Vilnius	7,33
18	Tallinn	7,22
19	Riga	6,56
20	Bratislava	6,22
=21	Athens	5,44
=21	Dublin	5,44
=23	Kiev	5,22
=23	Rome	5,22
25	Belgrade	4,67
26	Zagreb	4,56
27	Prague	4,22
28	Sofia	3,89
29	Istanbul	3,11
30	Bucharest	2,67

A concrete use case: The "City Data Pipeline"

Idea – a "classic" Semantic **Web** use case!

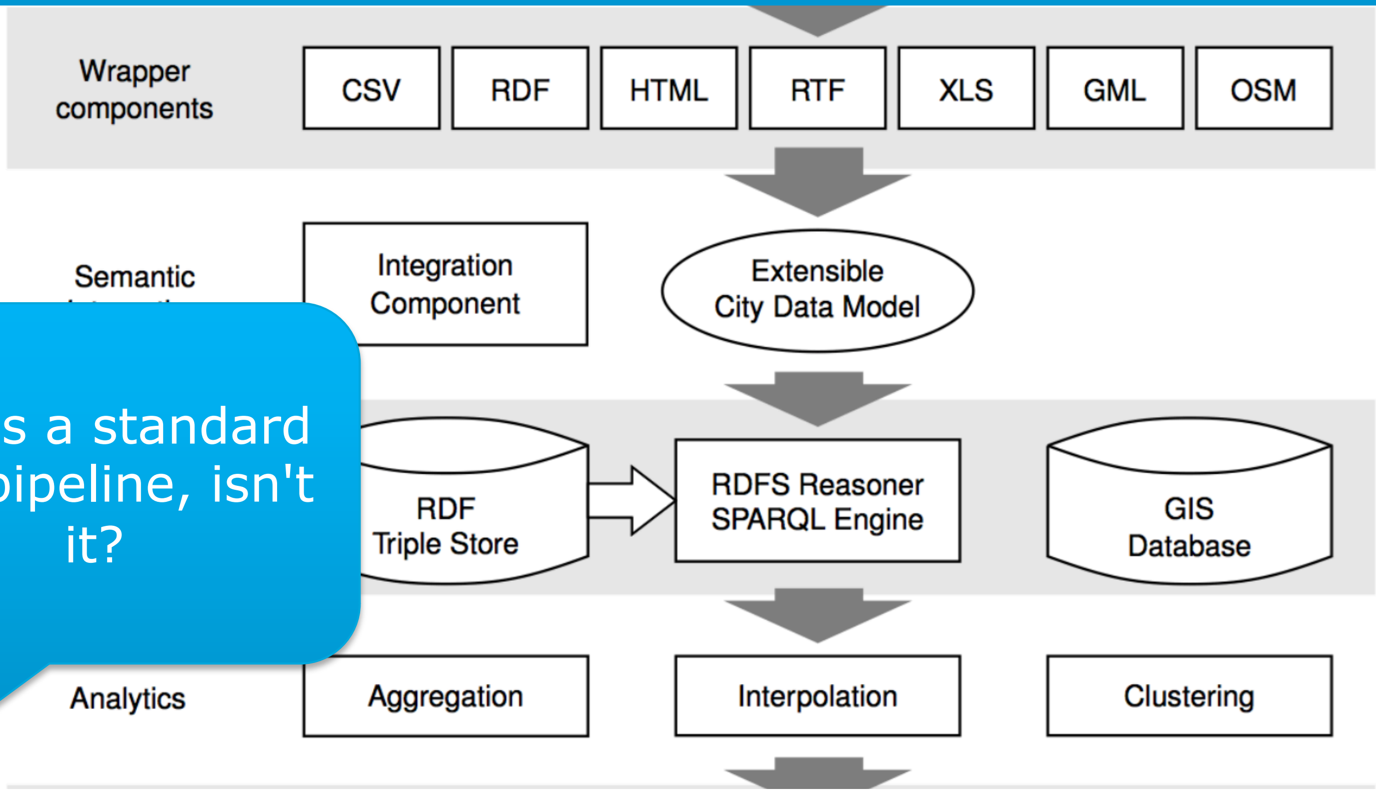
- Regularly integrate various relevant Open Data sources (e.g. eurostat, UNData, ...)
- Make integrated data available for re-use



(How) can ontologies help me?

- Are ontology languages expressive enough?
- Which ontologies could I (re-)use?
- Is there enough data at all?
- Where to find the right data?
- How to tackle inconsistencies?

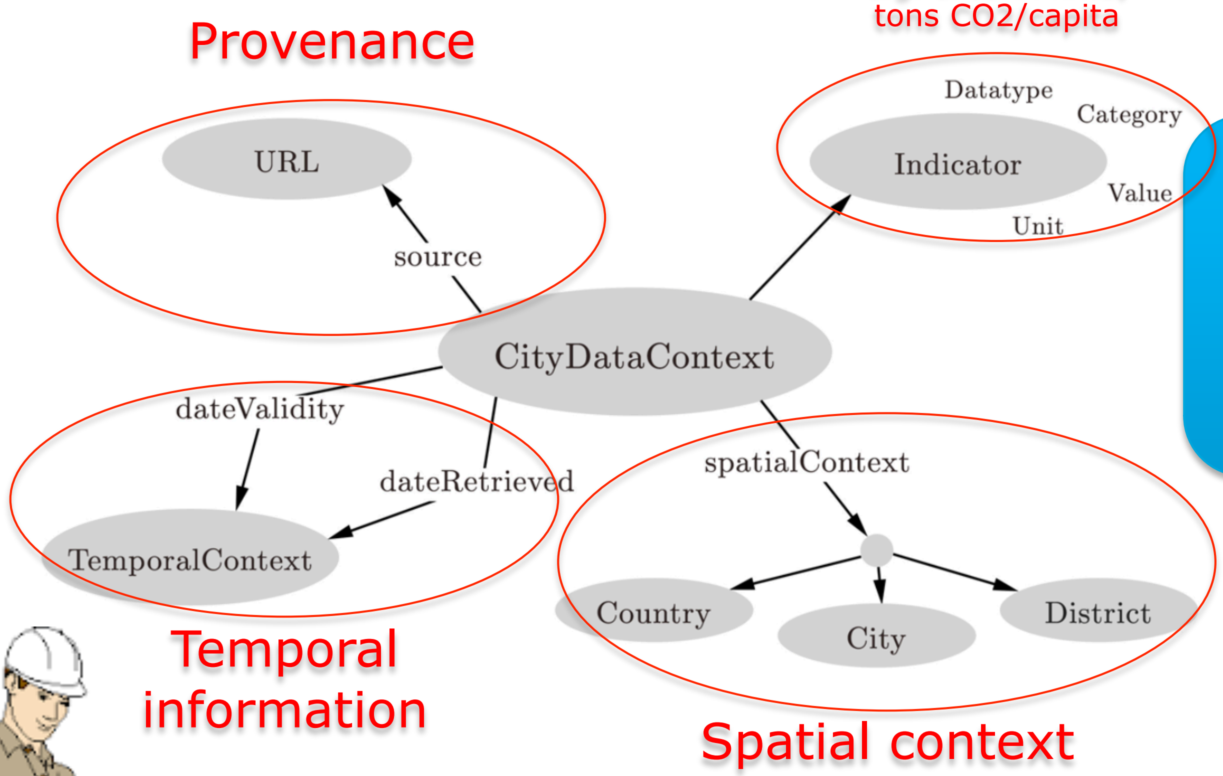
A concrete use case: The "City Data Pipeline"



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

A concrete use case: The "City Data Pipeline"

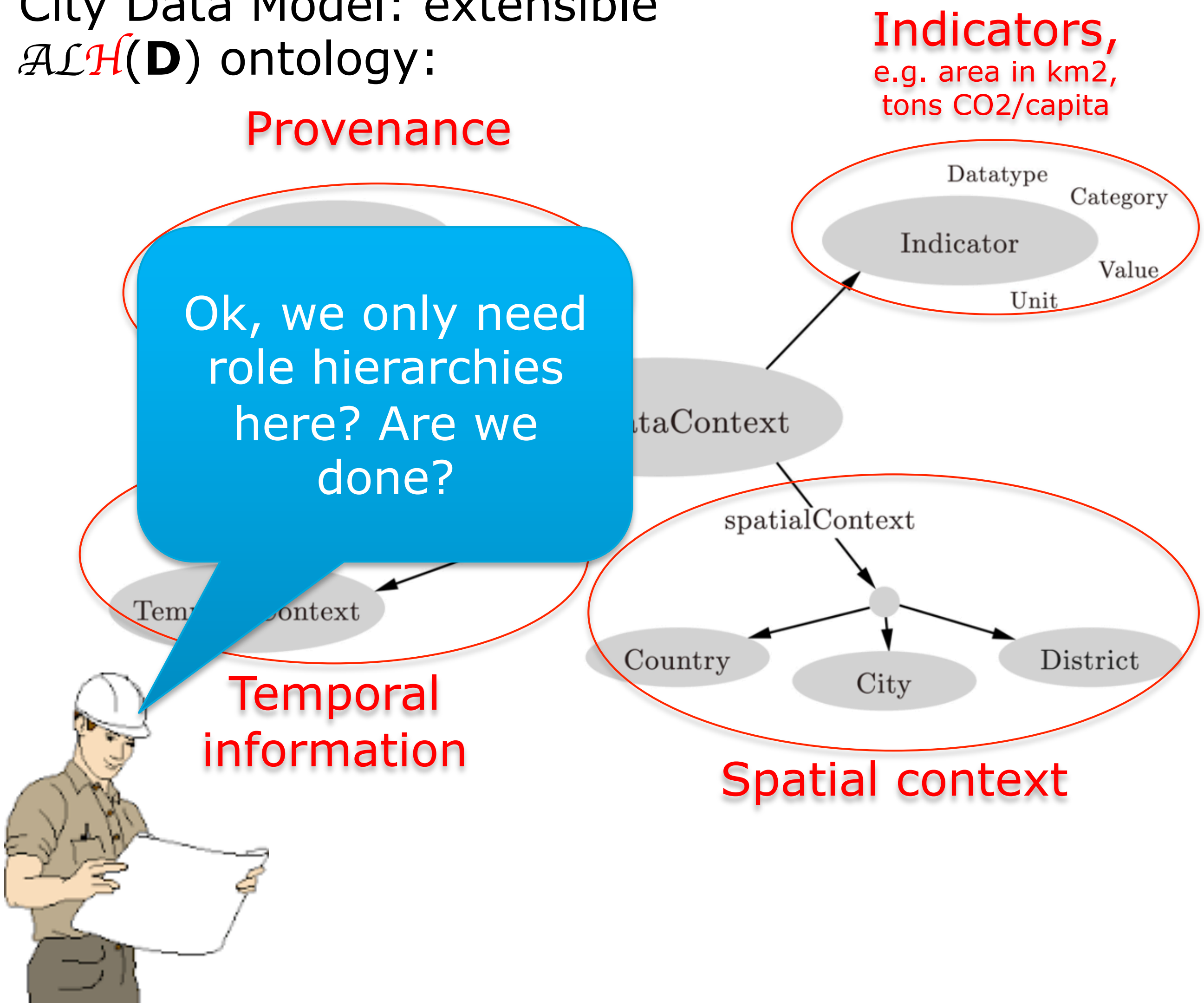
City Data Model: extensible $\mathcal{ALH}(\mathbf{D})$ ontology:



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

A concrete use case: The "City Data Pipeline"

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

Dbpedia:areakm2 $\sqsubseteq$:area

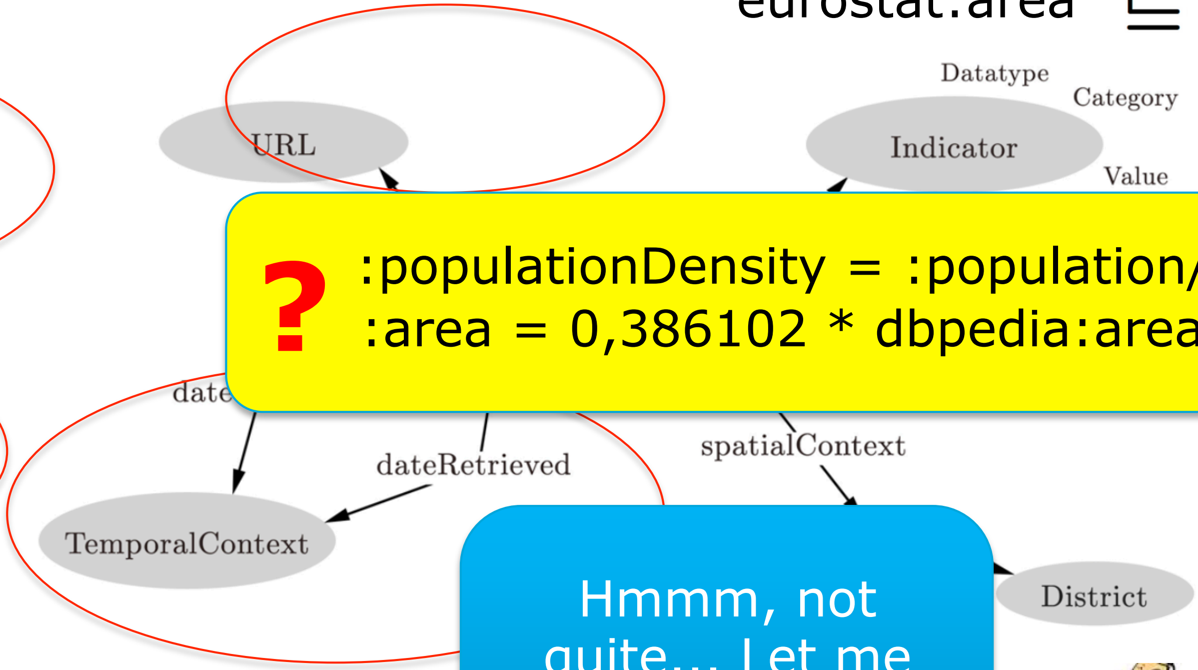
eurostat:area $\sqsubseteq$:area

Indicators, e.g. area in km2, tons CO_2/capita

Provenance

Temporal information

Spatial context



? :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?

RDFS with Attribute Equations via SPARQL Rewriting

Stefan Bischof[1,2] and Axel Polleres[1]

[1] Siemens AG Österreich, Siemensstraße 90, 1210 Vienna, Austria
[2] Vienna University of Technology, Favoritenstraße 9, 1040 Vienna, Austria

Abstract. In addition to taxonomic knowledge about concepts and properties typically expressible in languages such as RDFS and OWL, implicit information in an RDF graph may be likewise determined by arithmetic equations. The main use case here is exploiting knowledge about functional dependencies among numerical attributes expressible by means of such equations. While some of this knowledge can be encoded in rule extensions to ontology languages, we provide an arguably more flexible framework that treats attribute equations as first class citizens in the ontology language. The combination of ontological reasoning and attribute equations is realized by extending query rewriting techniques already successfully applied for ontology languages such as (the DL-Lite-fragment of) RDFS or OWL, respectively. We deploy this technique for rewriting SPARQL queries and discuss the feasibility of alternative implementations, such as rule-based approaches.

1 Introduction

A wide range of literature has discussed completion of data represented in RDF with implicit information through ontologies, mainly through taxonomic reasoning within a

Stefan Bischof, Axel Polleres. ESWC2013

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:

:area = 0,386102 * dbpedia:areaMi2

:areaMi2 = 2,589988 * :area

Can equational knowledge co-exist with OWL?

- An Interpretation $\mathcal{I}$ interprets datatype properties U as binary relations between domain elements and Data-Values *(for simple equations rational numbers are sufficient)*:

  $$U^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \mathbb{Q}$$

  dbo:populationToal rdfs:subPropertyOf :population .

- Interpretations of inclusion axioms are as usual, e.g.
  - A sub-property axiom **sp**

    dbr:Athens dbo:populationTotal 664046.

    U_1 **rdfs:subPropertyOf** U_2 $\quad U_1 \sqsubseteq U_2$

    dbr:Athens :population 664046.

    is satisfied in $\mathcal{I}$ *if* $U_1^{\mathcal{I}} \subseteq U_2^{\mathcal{I}}$

:populationDensity :definedByEquation ":population / :area" .

- **NEW:** A property equation axiom ***e***

  U_0 **:defineByEquation** "$f(U_1,\ldots U_n)$" .

  dbr:Athens :population 664046.
  dbr:Athens :area 38.964 .

  is satisfied in $\mathcal{I}$

  $$\text{if } \forall x, y_1, \ldots, y_n \left(\bigwedge_{i=1}^{n} (x, y_i) \in U_i^{\mathcal{I}}\right) \wedge \text{defined}(f(U_1/y_1, \ldots, U_n/y_n))$$

  dbr:Athens dbo:populationDensity 17042.55 .

  $$\Rightarrow (x, \text{eval}(f(U_1/y_1, \ldots, U_n/y_n)) \in U_0^{\mathcal{I}}$$

- An interpretation $\mathcal{I}$ is a model it satisfies
  - all inclusion axioms
  - ***all variants of*** all equation axioms

Can equational knowledge co-exist with OWL?

- An Interpretation $\mathcal{I}$ interpret datatype properties U as binary relations between domain elements and Data-Values (for our simple equations rational numbers are sufficient): $U^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \mathbb{Q}$
- Interpretations of inclusion axioms are as usual, e.g.
  - A sub-property axiom **sp**

  U_1 **rdfs:subPropertyOf** U_2 $\quad U_1 \sqsubseteq U_2$

  is satisfied in $\mathcal{I}$ *if* $U_1^{\mathcal{I}} \subseteq U_2^{\mathcal{I}}$

:populationDensity :definedByEquation “:population / :area” .

- **NEW:** A property equation axiom ***e***

  U_0 **:defineByEquation** “$f(U1,...U_n)$” .

  is satisfied in $\mathcal{I}$

$$\text{if } \forall x, y_1, \ldots, y_n (\bigwedge_{i=1}^{n} (x, y_i) \in U_i^{\mathcal{I}}) \wedge \text{defined}(f(U_1/y_1, \ldots, U_n/y_n))$$

dbr:Athens :population 664046.
dbr:Athens :area 0 .

:population :definedByEquation “:populationDensity * :area”.
:area:definedByEquation “:population / :populationDensity” .

- An interpretation $\mathcal{I}$ is a model if it satisfies
  - all inclusion axioms
  - ***all variants of*** all equation axioms

Can materialization and/or query rewriting be used?

- Rule-based Materialization:

$(S, \text{popDensity}, PD) \leftarrow (S, \text{population}, P), (S, \text{area}, A), \; PD := P/A, \; A \neq 0.$

$(S, \text{area}, PD) \leftarrow (S, \text{population}, P), (S, \text{popDensity}, PD), \; A := P/PD, PD \neq 0.$

$(S, \text{population}, P) \leftarrow (S, \text{area}, A), (S, \text{popDensity}, PD), \; P := A * PD.$

dbr:Athens dbo:population **2**.
dbr:Athens dbo:area **3**.

dbr:Athens dbo:popDensity 0.66666666.

dbr:Athens dbo:area 3.00000000003.

dbr:Athens dbo:population 1.99999998002.

… potentially infinite values by rounding errors.

Similarly, for ambiguous values (assume 2 population values for Athens)

Can materialization and/or query rewriting be used?

- Rewriting? Again consider clausal form of all variants of equations:

$(S, \text{popDensity}, PD) \leftarrow (S, \text{population}, P), (S, \text{area}, A), \; PD := P/A$

$(S, \text{area}, PD) \leftarrow (S, \text{population}, P), (S, \text{popDensity}, PD), \; A := P/PD$

$(S, \text{population}, P) \leftarrow (S, \text{area}, A), (S, \text{popDensity}, PD), \; P := A * PD$

```
dbr:Athens dbo:Athens 664046.
dbr:Athens dbo:area 38.964 .
```

Finally, the resulting UCQs with assignments can be rewritten back to SPARQL using BIND

```
SELECT ?PD WHERE { :Athens dbo:popDensity ?PD}
```

$q(PD) \leftarrow (S, \text{popDensity}, PD)$

$q(PD) \leftarrow (S, \text{population}, P), (S, \text{area}, A), PD := P/A$

~~$q(PD) \leftarrow (S, \text{popDensity}, PD'), (S, \text{area}, A'), (S, \text{area}, A), PD := P/A, P := PD' * A'$~~

.. infinite expansion even if only 1 equation is considered.

Solution: "blocking" recursive expansion of the same equation for the same value.

```
SELECT ?PD WHERE { {:Athens dbo:popDensity ?PD }
                   UNION
                   { :Athens dbo:population ?P ; dbo:area ?A .
                     BIND (?P/?A AS ?PD )}
                 }
```

Algorithm:

- "Down-stripped" version of PerfectRef [Calvanese, 2007] which handles equations by keeping "adornments" of attributes during rewriting:

Algorithm 1: Rewriting algorithm $\mathsf{PerfectRef}_{\mathsf{E}}$

Input: Conjunctive query q, TBox $\mathcal{T}$
Output: Union (set) of conjunctive queries

$P := \{q\}$
repeat
    $P' := P$
    foreach $q \in P'$ **do**
        foreach *g in q* **do** // expansion
            foreach *inclusion axiom I in $\mathcal{T}$* **do**
                if *I is applicable to g* **then**
                    $P := P \cup \left\{ q[g/\,\mathrm{gr}(g, I)] \right\}$
            foreach *equation axiom E in $\mathcal{T}$* **do**
                if $g = U^{\mathrm{adn}(g)}(x, y)$ *is an (adorned) attribute atom and* $\mathrm{vars}(E) \cap \mathrm{adn}(g) = \emptyset$ **then**
                    $P := P \cup \left\{ q[g/\,\mathrm{expand}(g, E)] \right\}$
until $P' = P$
return P

Can materialization and/or query rewriting be used?

- Rule-based Materialization:

$$(S, \text{popDensity}, PD) \leftarrow (S, \text{population}, P), (S, \text{area}, A),\ PD := P/A,\ A \neq 0.$$
$$(S, \text{area}, PD) \leftarrow (S, \text{population}, P), (S, \text{popDensity}, PD),\ A := P/PD, PD \neq 0.$$
$$(S, \text{population}, P) \leftarrow (S, \text{area}, A), (S, \text{popDensity}, PD),\ P := A * PD.$$

dbr:Athens dbo:population 2.
dbr:Athens dbo:area 3.

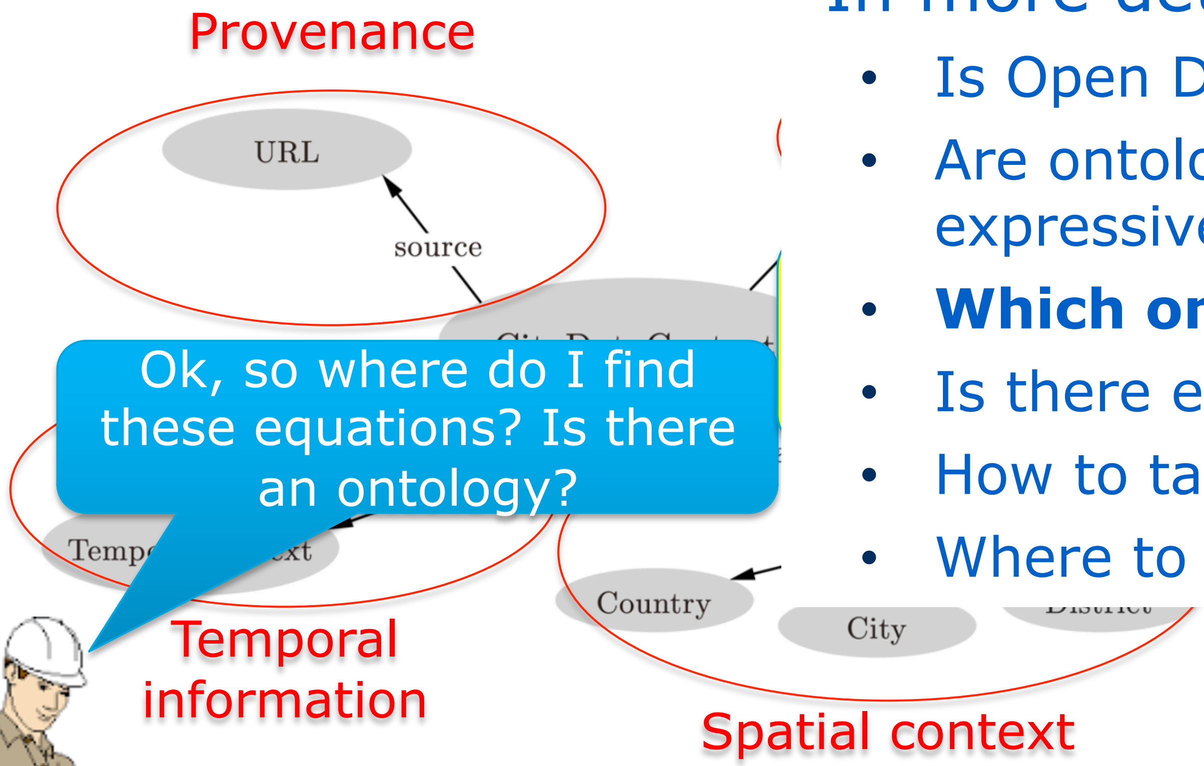
Similar blocking possible in some rule systems, e.g. Jena Rules:

```
[ (?C :area ?A) (?C :population ?P)
  notEqual(?A, 0) quotient(?P, ?A, ?PD)
  noValue(?C, :populationDensity)  -> (?C :populationDensity ?D)]

[ (?C :populationDensity ?PD) (?city :population ?P)
  notEqual(?PD, 0) quotient(?P, ?PD, ?A)
  noValue(?C, :area)  -> (?city :area ?A)]

[ (?C :area ?A) (?C :populationDensity ?P)  product(?A, ?PD, ?P)
  noValue(?city, :population)  -> (?city :population ?P)]
```

Side remark: Experiments in our ESWC2013 paper favor rewriting approach.

A concrete use case: The "City Data Pipeline"

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



In more detail:

- Is Open Data useful at all?
- Are ontology languages expressive enough?
- **Which ontologies could I use?**
- Is there enough data at all?
- How to tackle inconsistencies?
- Where to find the right data?

Equational knowledge:

- Eurostat/Urbanaudit:
  - http://ec.europa.eu/regional_policy/archive/urban2/urban/audit/ftp/vol3.pdf

Domain	N°	Variables	Indicator Name	Presentation of Indicator						Calculations required
				YB Sum	YB CT	ICA				
						City	WTU	SC1	SC2	
Crime	8	Total number of recorded crimes within city (per year)	Total recorded crimes (per 1000 population per year)	X	X	X	X		X	(Total crimes recorded x 1000)/Total resident population

Equational knowledge: Unit conversion

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

http://qudt.org/

http://www.wurvoc.org/vocabularies/om-1.8/

QUDT

QUDT - Quantities, Units, Dimensions and Data Types Ontologies

March 18, 2014

Authors:
Ralph Hodgson, TopQuadrant, Inc.
Paul J. Keller, NASA AMES Research Center
Jack Hodges
Jack Spivak

Overview

The QUDT Ontologies, and derived XML Vocabularies, are being developed by TopQuadrant and NASA. Originally, they were developed for the NASA Exploration Initiatives Ontology Models (NExIOM) project, a Constellation Program initiative at the AMES Research Center (ARC). They now for the basis of the NASA QUDT Handbook to be published by NASA Headquarters.

Ontology of units of Measure (OM)

search concepts in this ontology

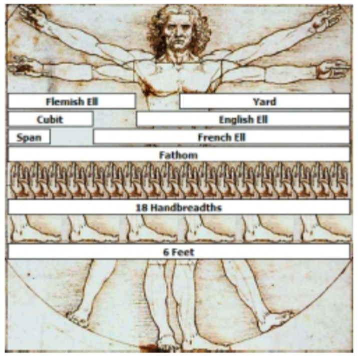
OK

download this ontology

RDF/XML OK

description

The Ontology of units of Measure and related concepts (OM) models concepts and relations important to scientific research. It has a strong focus on units and quantities, measurements, and dimensions.



creator

Hajo Rijgersberg, Mark van Assem, Don Willems, Mari Wigham, Jeen Broekstra, Jan Top

version info

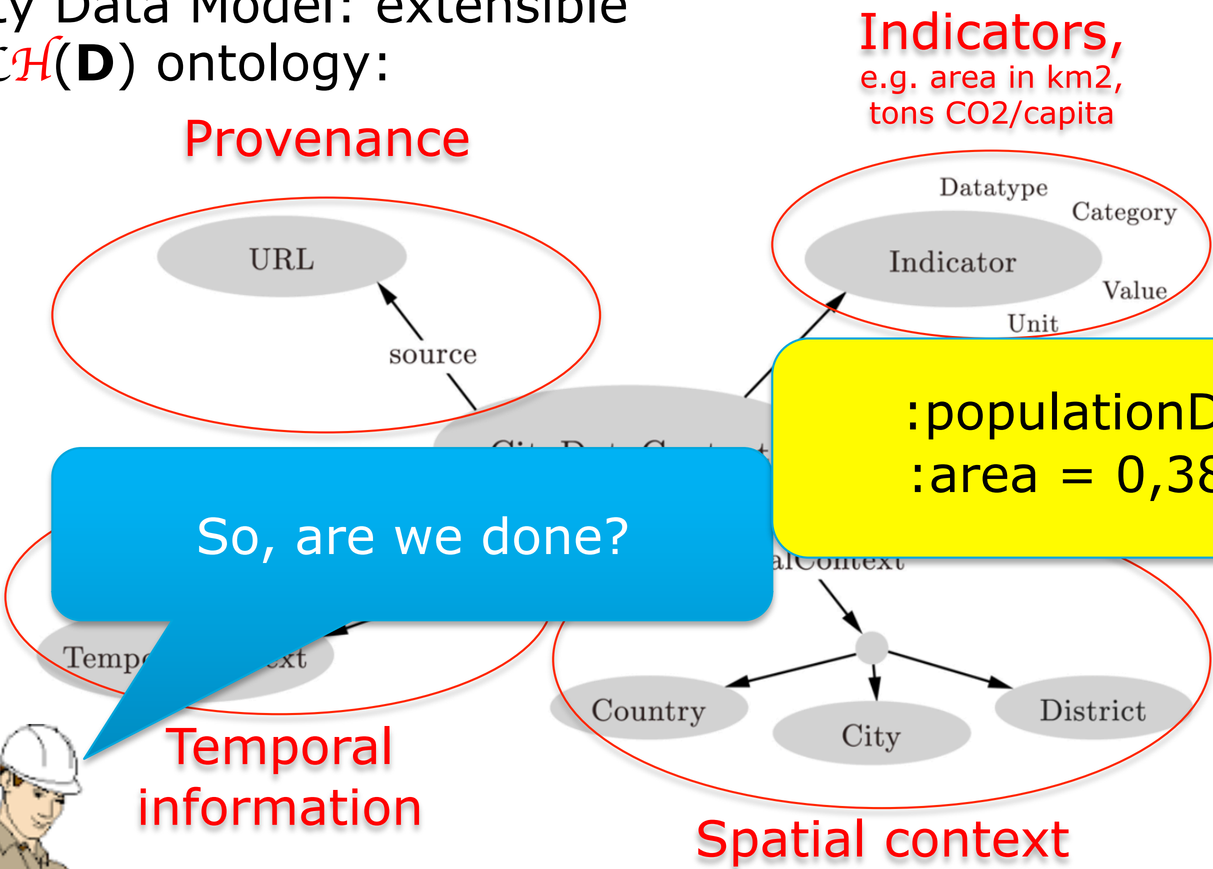
1.8.0

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

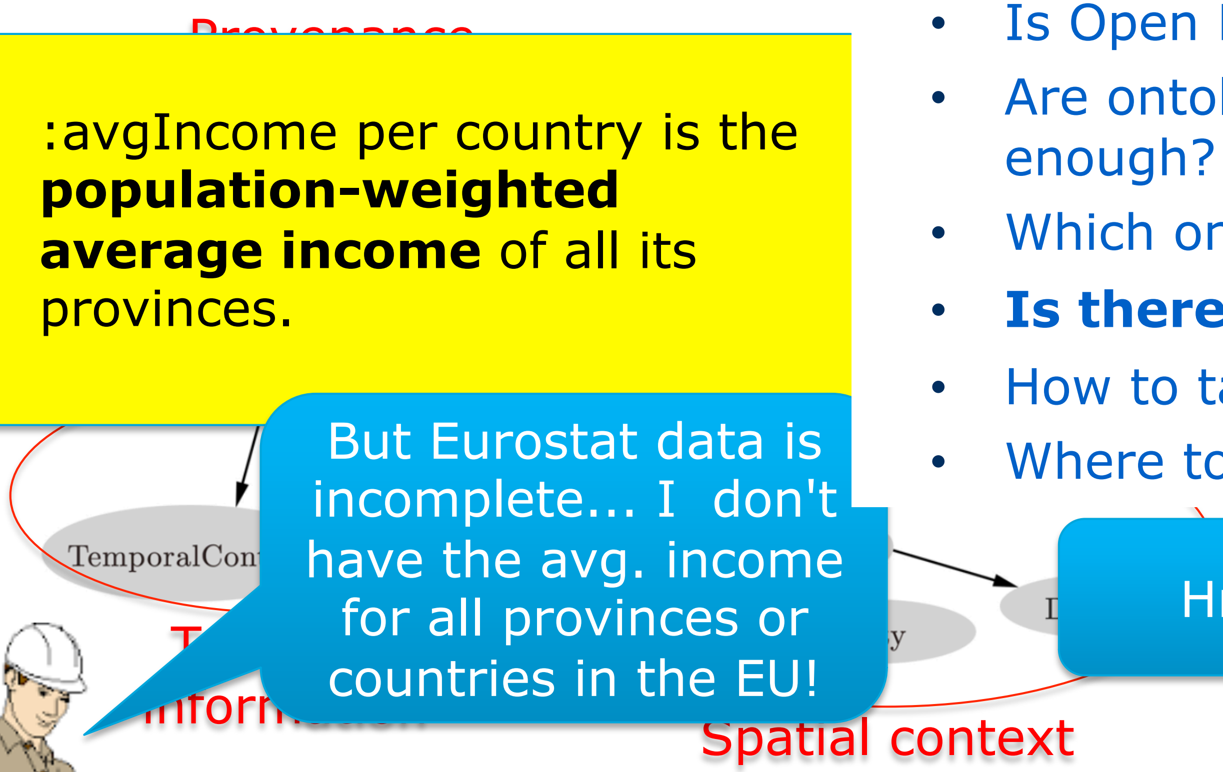


:populationDensity = :population/:area
:area = 0,386102 * dbpedia:areaMi2

So, are we done?

A concrete use case: The "City Data Pipeline"

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



:avgIncome per country is the **population-weighted average income** of all its provinces.

But Eurostat data is incomplete... I don't have the avg. income for all provinces or countries in the EU!

Hmmm...

In more detail:

- Is Open Data useful at all?
- Are ontology languages expressive enough?
- Which ontologies could I use?
- **Is there enough data at all?**
- How to tackle inconsistencies?
- Where to find the right data?

Challenge – Missing values

- Found a huge 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
1990	177	121	2 480	18 937	88.4
2000	477	156	10 347	64 065	85.0
2005	651	167	23 494	85 223	78.4
2010	905	202	90 490	92 320	50.5
2004 - 2012	943	215	531 146	1 293 559	70.9
All (1990 - 2012)	943	215	638 934	4 024 201	86.3

Table 2: United Nations Data Set

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

Challenges – Missing values

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- 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

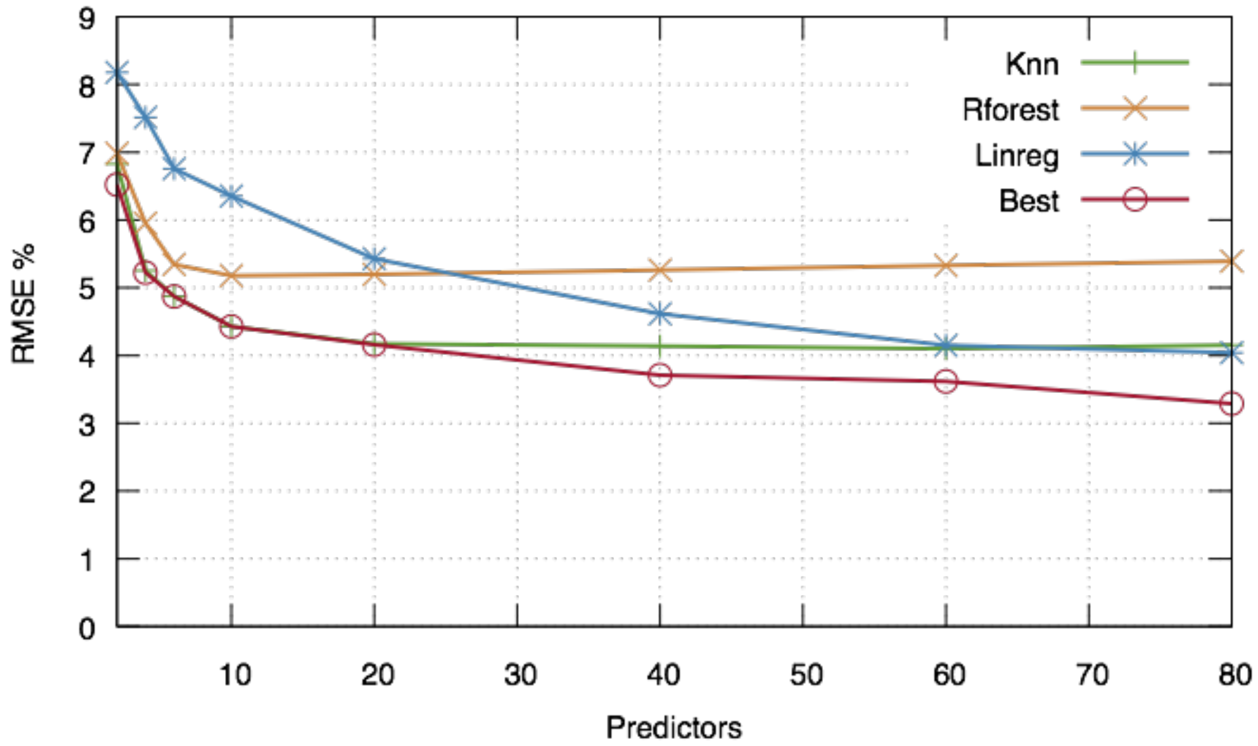
	Vienna	Augsburg	Valletta	Marbella	Stockholm	Funchal
Cars	655806	111561	95858			
Nationals	1342704	216289	203657			
Women per 1000 Men	109.8	108.7	101.9			
Available Beds per 1000				138.3	14969	166.1
Average area of living				36.42	37.24	38.16
Cinema Seats				4691	12751	2676

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:

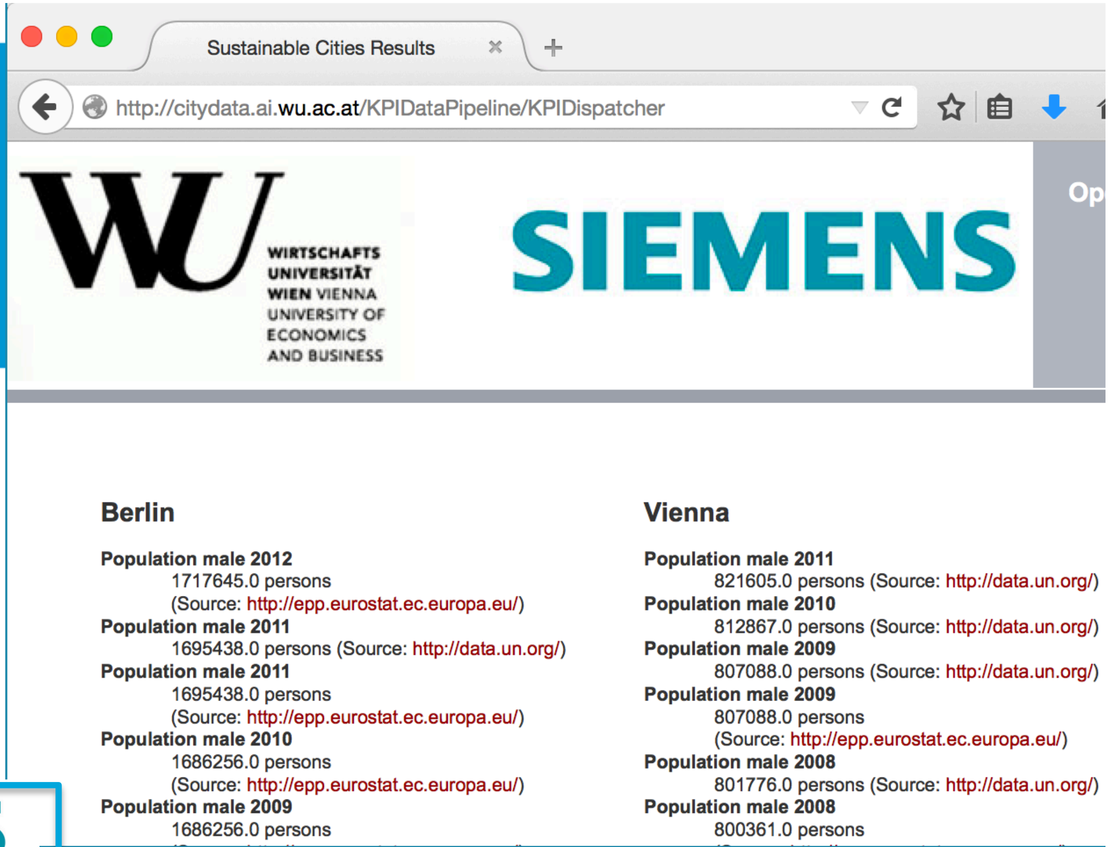
- Instead of using indicators directly we use Principle Components, built from the indicators
- For buidling the PCs, fill in missing data points with neutral values → predict all rows

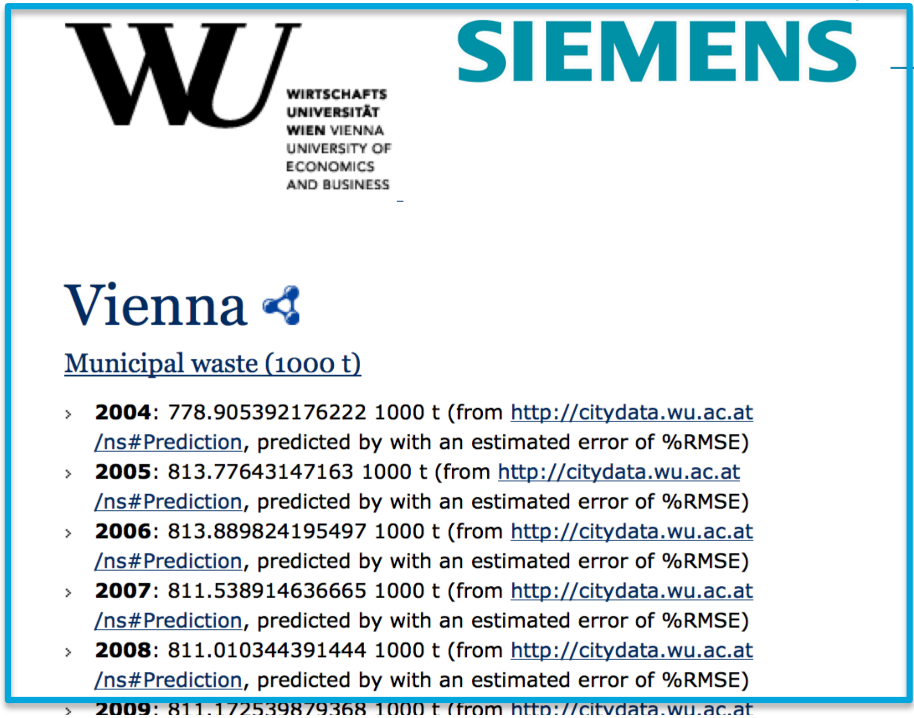


City Data Pipeline

citydata.wu.ac.at

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





...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[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

So, are we done?

...cess to high quality and recent data is crucial both for deci- ... 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.

Lesson(s) learnt?

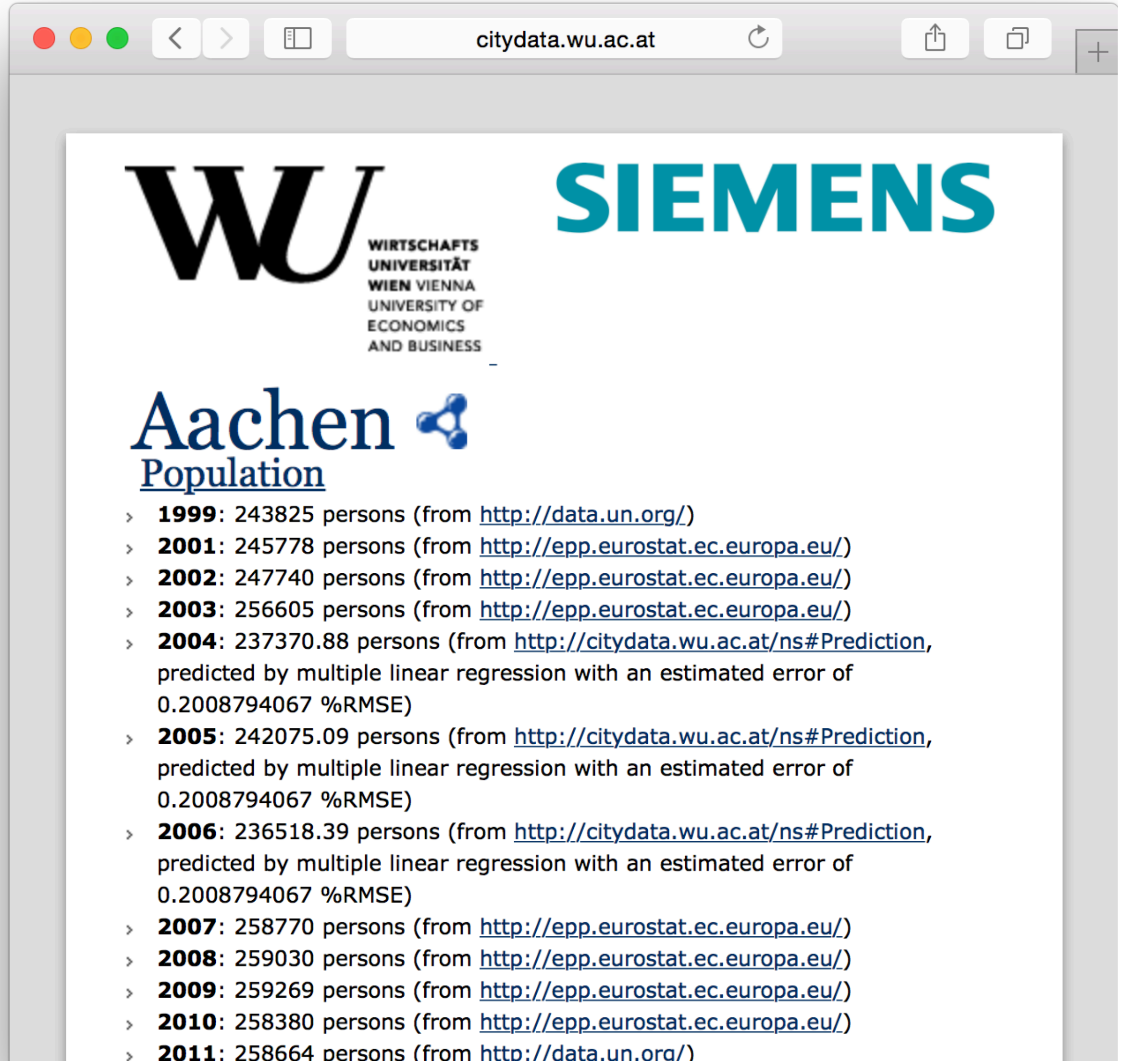
Hmmm, still, lots of open challenges!

- Time series analysis is necessary
- Open Data is incomparable
- Still not great coverage of all available sources
- Open Data **Quality** is an issue
- Still unanswered:
  - Is Open Data useful at all?
  - Are ontology languages expressive enough?
  - Is there enough data at all?
  - Which ontologies could I use?
  - How to tackle inconsistencies?
  - Where to find the right data?

Time series analysis is necessary

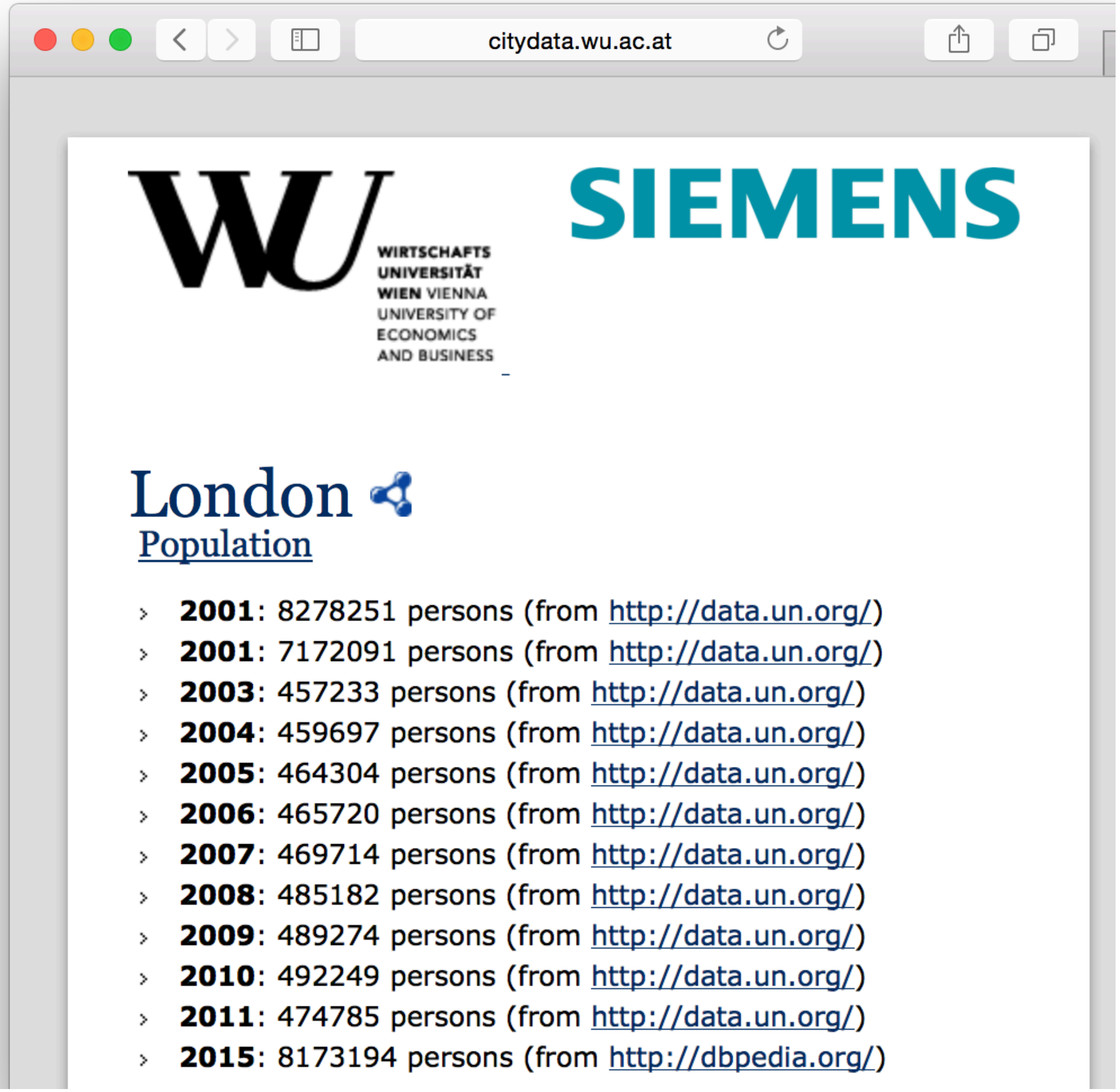
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- Predictions on time series are partially very bad at the moment:
- Most of the data we look at is time series data/data chaning over time.



Open Data is incomparable

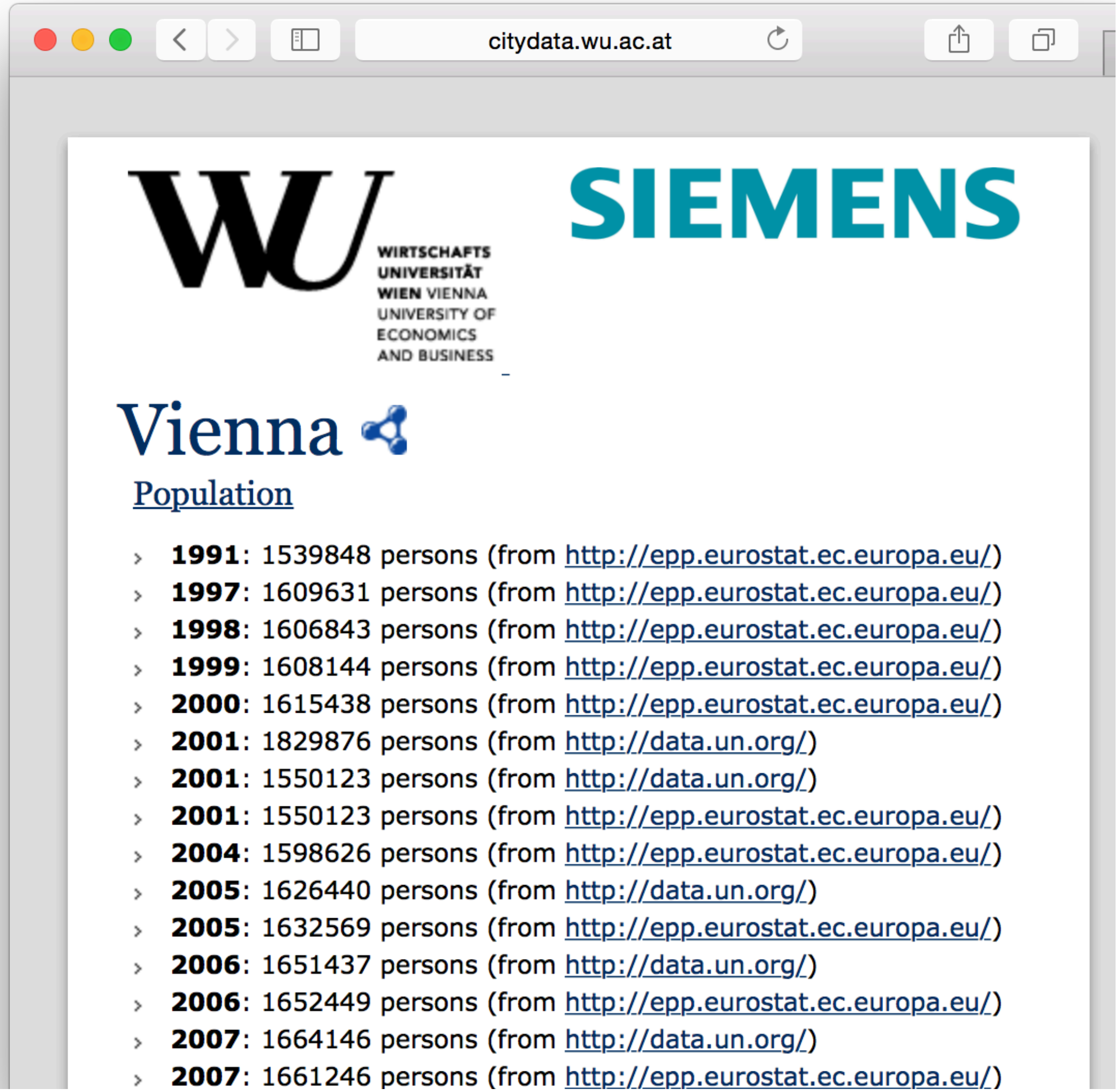
- More surprising maybe, how much obviously weird data you find:
  - Inconsistencies across and within datasets



Open Data is incomparable

- More surprising maybe, how much obviously weird data you find:
  - Inconsistencies across and within datasets
  - Still, some datasets match quite well on certain indicators
  - Open: (How) can we exploit this?

→ *Ontology learning!*



Worthwhile related work to look at… Paulheim, 2012 (ESWC), Nickel et al. 2012 (WWW)

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Generating Possible Interpretations for Statistics from Linked Open Data

Heiko Paulheim

Technische Universität Darmstadt
Knowledge Engineering Group
paulheim@ke.tu-darmstadt.de

Abstract. Statistics are very present in our daily lives. Every day, new statistics are published, showing the perceived quality of living in different cities, the corruption index of different countries, and so on. Interpreting those statistics, on the other hand, is a difficult task. Often, statistics collect only very few attributes, and it is difficult to come up with hypotheses that explain, e.g., *why* the perceived quality of living in one city is higher than in another. In this paper, we introduce *Explain-a-LOD*, an approach which uses data from Linked Open Data for generating hypotheses that explain statistics. We show an implemented prototype and compare different approaches for generating hypotheses by analyzing the perceived quality of those hypotheses in a user study.

Factorizing YAGO

Scalable Machine Learning for Linked Data

Maximilian Nickel
Ludwig-Maximilians University Munich
Oettingenstr. 67
Munich, Germany
nickel@dbs.ifi.lmu.de

Volker Tresp
Siemens AG
Corporate Technology
Otto-Hahn Ring 6
Munich, Germany
volker.tresp@siemens.com

Hans-Peter Kriegel
Ludwig-Maximilians University Munich
Oettingenstr. 67
Munich, Germany
kriegel@dbs.ifi.lmu.de

ABSTRACT

Vast amounts of structured information have been published in the Semantic Web's Linked Open Data (LOD) cloud and their size is still growing rapidly. Yet, access to this information via reasoning and querying is sometimes difficult, due to LOD's size, partial data inconsistencies and inherent noisiness. Machine Learning offers an alternative approach to exploiting LOD's data with the advantages that Machine Learning algorithms are typically robust to both noise and data inconsistencies and are able to efficiently utilize non-deterministic dependencies in the data. From a Machine Learning point of view, LOD is challenging due to its relational nature and its scale. Here, we present an efficient approach to relational learning on LOD data, based on the factorization of a sparse tensor that scales to data consisting of millions of entities, hundreds of relations and billions of known facts. Furthermore, we show how ontological knowledge can be incorporated in the factorization to improve learning results and how computation can be distributed across multiple nodes. We demonstrate that our approach is able to factorize the YAGO 2 core ontology and globally predict statements for this large knowledge base using a single dual-core desktop computer. Furthermore, we show experimentally that our approach achieves good results in several relational learning tasks that are relevant to Linked Data. Once a factorization has been computed, our model is able to predict efficiently, and without any additional training, the likelihood of any of the $4.3 \cdot 10^{14}$ possible triples in the YAGO 2 core ontology.

1. INTRODUCTION

The Semantic Web's Linked Open Data (LOD) [6] cloud is growing rapidly. At the time of this writing, it consists of around 300 interlinked databases, where some of these databases store billions of facts in form of RDF triples.[1] Thus, for the first time, relational data from heterogeneous, interlinked domains is publicly available in large amounts, which provides exciting opportunities for Machine Learning. In particular, much progress has been made in recent years in the subfield of *Relational* Machine Learning to learn efficiently from attribute information and information about the entities' relationships in interlinked domains. Some Relational Machine Learning approaches can exploit contextual information that might be more distant in the relational graph, a capability often referred to as collective learning. State-of-the-art collective learning algorithms can therefore be expected to utilize much of the information and patterns that are present in LOD data. Moreover, the Semantic Web itself can benefit from Machine Learning. Traditional Semantic Web approaches such as formal semantics, reasoning or ontology engineering face serious challenges in processing data in the LOD cloud, due to its size, inherent noisiness and inconsistencies. Consider, for example, that `owl:sameAs` is often misused in the LOD cloud, leading to inconsistencies between different data sources [13]. Further examples include malformed datatype literals, undefined classes and properties, misuses of ontological terms [16] or the modeling of a simple fact such as *Nancy Pelosi voted in favor of the Health Care Bill* using eight RDF triples [15]. Partial inconsistencies in the data or noise such as duplicate enti-

Lesson(s) learnt?

Hmmm, still, lots of open challenges!

- Time Series analysis is necessary
- Open Data is incomparable
- **Open Data Quality is an issue**

- Still unanswered:
  - Is Open Data useful at all?
  - Are ontology languages expressive enough?
  - Is there enough data at all?
  - Which ontologies could I use?
  - **How to tackle inconsistencies?**
  - **Where to find the right data?**

Data Quality issues:

- Missing
- Outdated data
- Wrong data
- Ambiguous Data
- Wrong meta-data
- Data source offline/not reachable

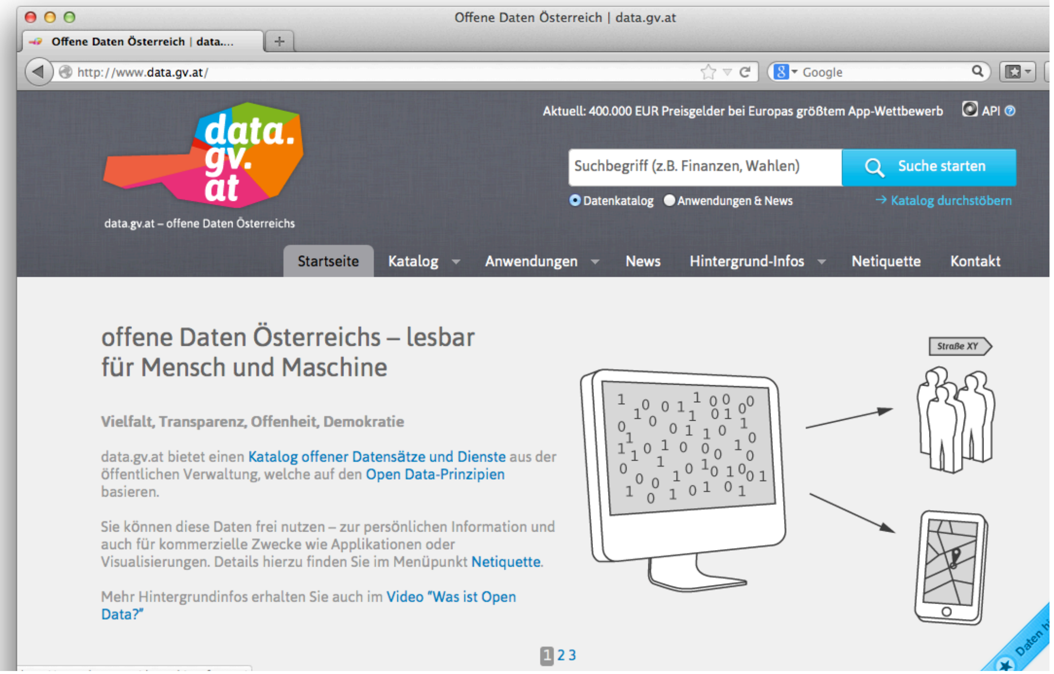
Open Data Portals

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/

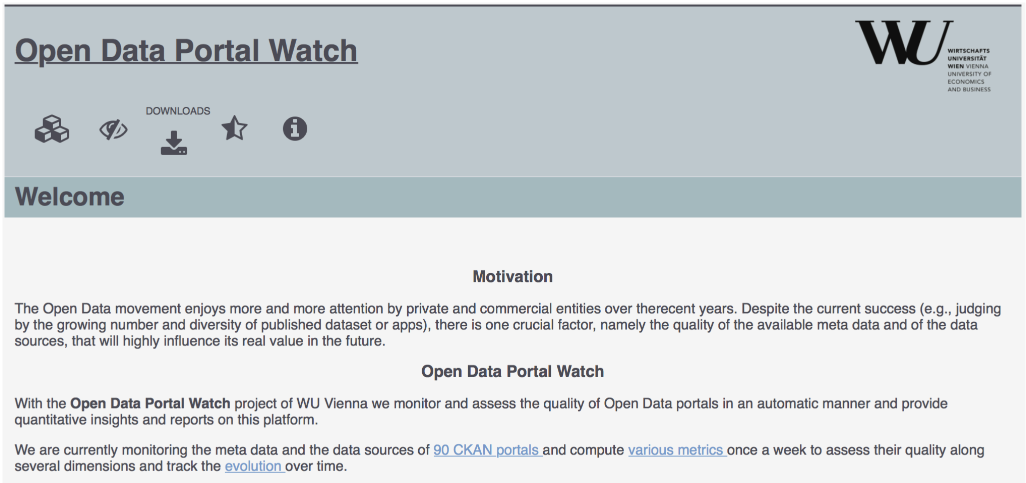
- machine-processable? ...
  ... **partially**



Open Data Portal Watch
... a first step.

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



Open Data Portal list

Open Data Portal Watch

Brief overview of 89 Open Data CKAN portals

Sort by: Domain | Country | Datasets | Resources — Filter: Filter by portal or country — Tile view | Table View

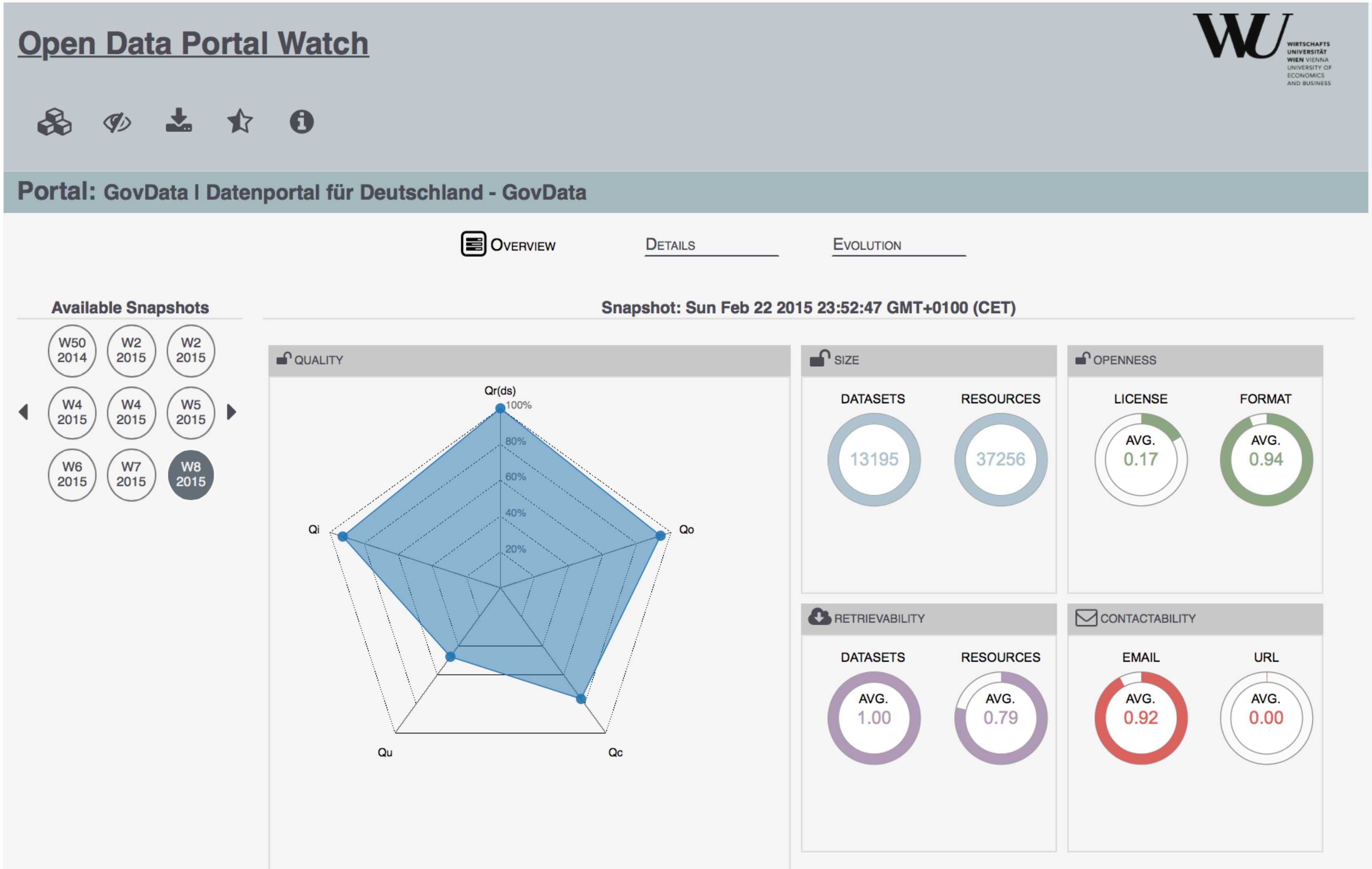
Portal	Country	Datasets	Resources
annuario.comune.fi.it	Italy	358 DATASETS	1363 RESOURCES
catalogue.datalocale.fr	France	303 DATASETS	751 RESOURCES
dados.gov.br	Brazil	501 DATASETS	4344 RESOURCES
data.buenosaires.gob.ar	Argentina	123 DATASETS	626 RESOURCES
data.edostate.gov.ng	Nigeria	164 DATASETS	207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

QUALITY DIMENSIONS

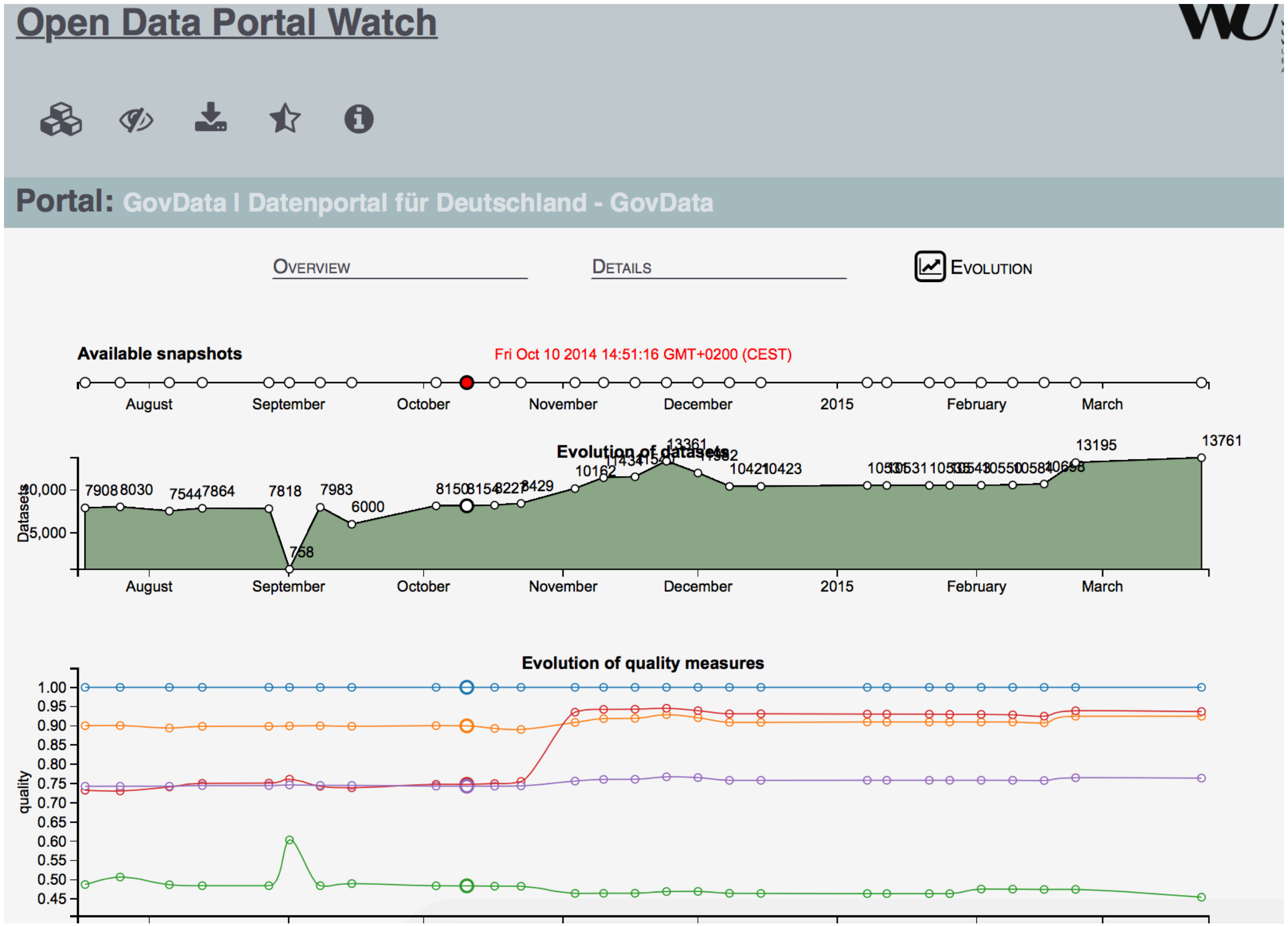
DIMENSION	DESCRIPTION
Retrievability	The extent to which meta data and resources can be retrieved.
Usage	The extent to which available meta data keys are used to describe a dataset.
Completeness	The extent to which the used meta data keys are non empty.
Accuracy	The extent to which certain meta data 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



ODP Evolution



ODP CHANGES

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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 10 entries

Search:

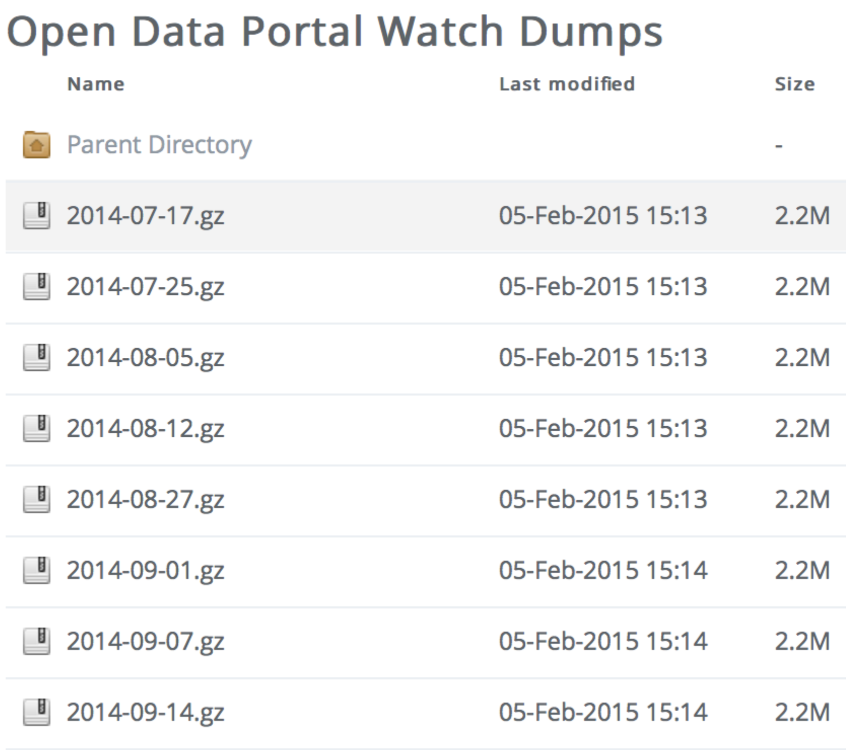
PORTAL	FROM	TO	CHANGE	↓CHANGE PERCENTAGE
data.sa.gov.au *(2014-07-17)→ (2015-03-15)*	484	5721	5237	1082.02%
datos.codeandomexico.org *(2014-07-17)→ (2015-03-15)*	94	715	621	660.64%
data.opendataportal.at *(2014-07-17)→ (2015-03-16)*	46	323	277	602.17%
annuario.comune.fi.it *(2014-08-07)→ (2015-03-15)*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

Open Data Portal Watch

Towards assessing the quality evolution of Open Data portals

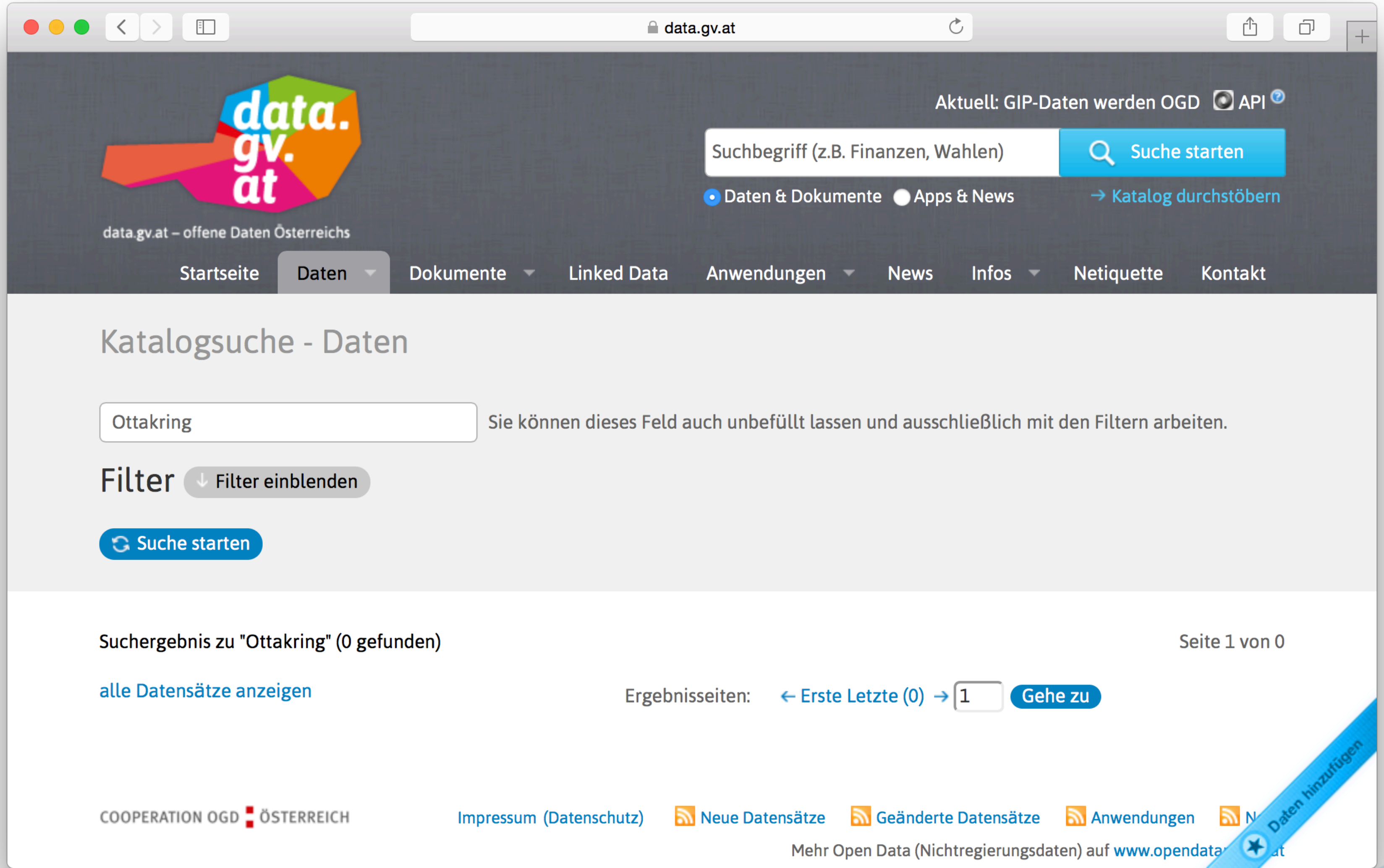
Jürgen Umbrich, Sebastian Neumaier, Axel Polleres
Vienna University of Economics and Business, Vienna, Austria

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:
  - Significantly varying quality acrosss portals
  - Rapid growth for some portals
  - Huge variety and range of datasets
  - Open Data Portal **search** is a big problem.

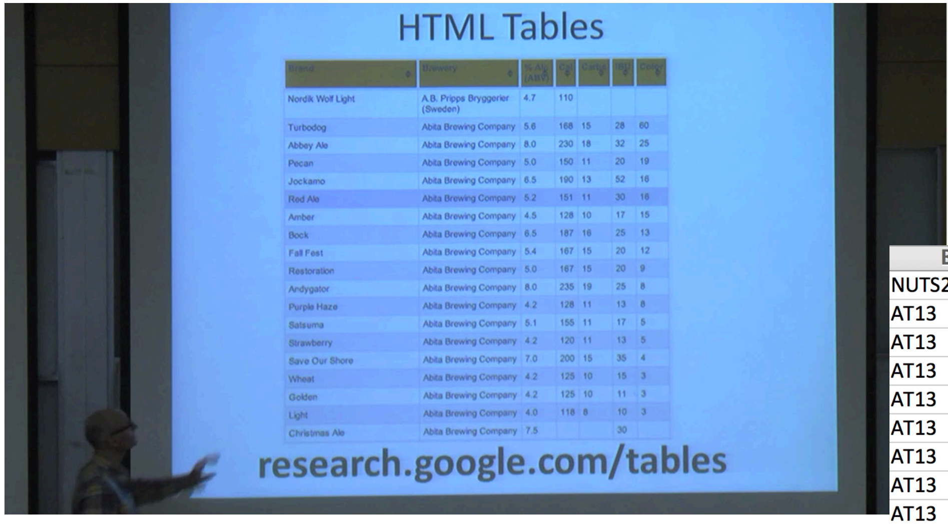
Open Data Portal search is a big problem… Why?

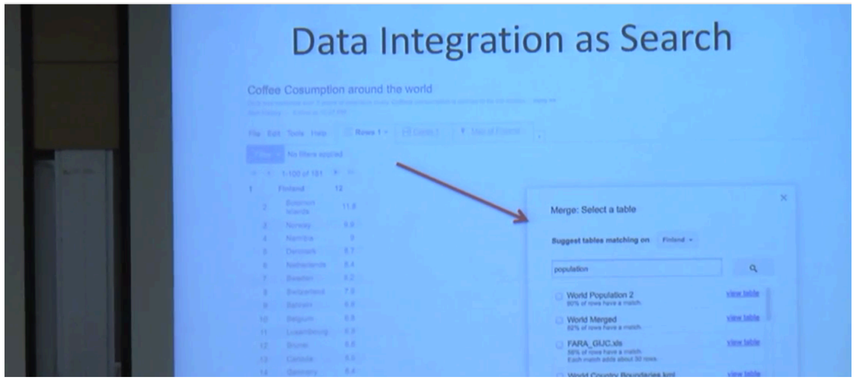


Open Data integration as Search?

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
A							014
A							014
A							014
A							014
A							014
A							014
A							014
AT13	AT130	92001	0	84305	41200	43105	01.01.2014
AT13	AT130	92101	0	148947	71633	77314	01.01.2014

Disclaimer: Won't attempt to compete, *but …*

a) This looks like a slightly different problem…

b) Can linking to "Open" knowledge graphs help? (wikidata, dbpedia?) … Probably.

What's next? Research roadmap to make Open Data usage more effective:

- Improving Open Data Quality, make OD better searchable...
- https://www.data.gv.at/wp-content/uploads/2012/03/Mission-Statement-AG-Qualitaetssicherung-OpenData-Portale.pdf

COOPERATION OGD ÖSTERREICH

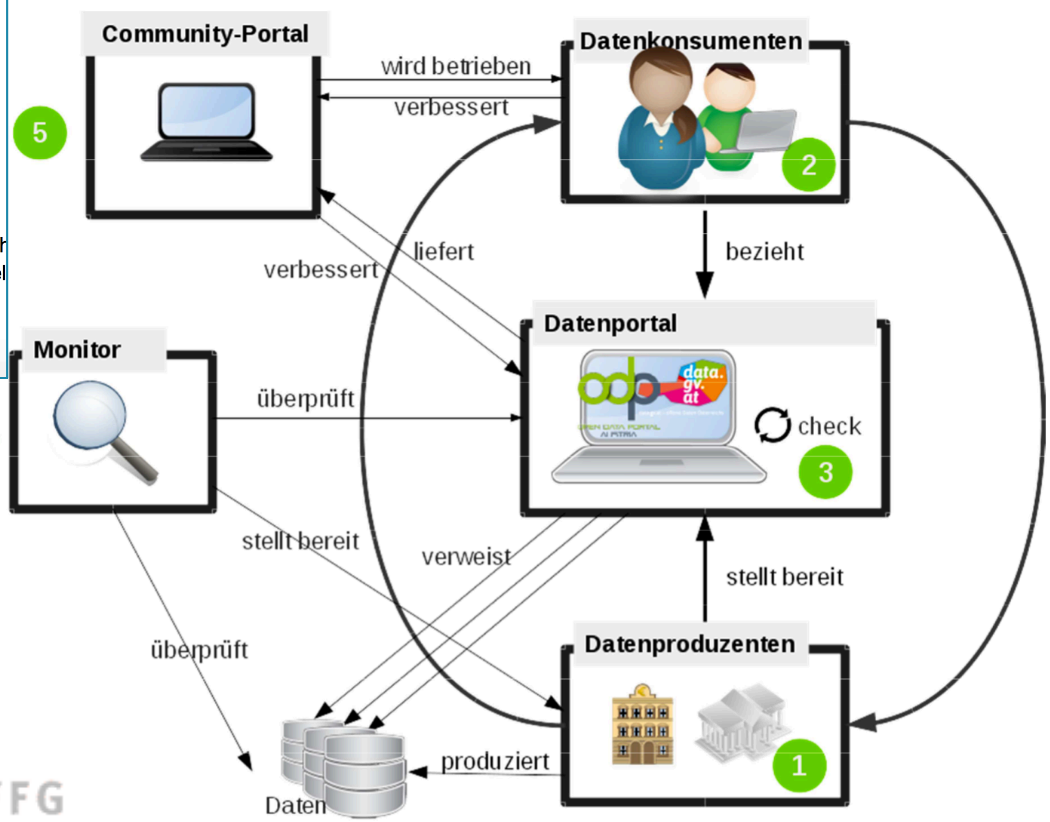
Datenqualität und Veröffentlichungsprozesse

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 technisch Maßnahmen und die Erstellung von Leitfäden zur empfohlenen Praxis die Datenqualität aktuel 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:

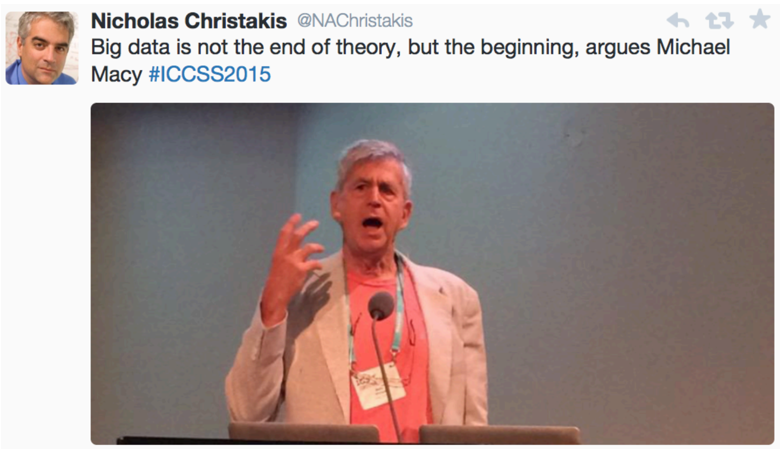
ADEQUATe: Analytics & Data Enrichment to improve the QUAliTy of Open Data

Project Start: Fall 2015

FFG

Integrating Open Data: (How) Can Description Logics Help me?

... even the computational social scientists don't buy that:



- Expressive ontology languages (plus e.g. equational knowledge) needed
- combination of reasoning about formal background knowledge & statistical methods needed
- **temporal aspects** need to be taken into account, but also **provenance**
- soundness/completeness (KRR) vs. coverage/accuracy (ML)

- "NoLD"... not only Linked Data

Integrating Open Data: (How) Can Description Logics Help me?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Temporal aspects:

- On Implementing Temporal Query Answering in DL-Lite (extended abstract) Veronika Thost, Jan Holste, Özgür Özcep (DL2015)
- The Complexity of Temporal Description Logics with Rigid Roles and Restricted TBoxes: In Quest of Saving a Troublesome Marriage Víctor Gutiérrez Basulto, Jean Christoph Jung, Thomas Schneider (DL2015)
- Temporal Query Answering in EL. Stefan Borgwardt, Veronika Thost (DL2015)
- Interval Temporal Description Logics. Alessandro Artale, Roman Kontchakov, Vladislav Ryzhikov, Michael Zakharyaschev (DL2015)
- Temporal OBDA with LTL and DL-Lite. Alessandro Artale, Roman Kontchakov, Alisa Kovtunova, Vladislav Ryzhikov, Frank Wolter, Michael Zakharyaschev (DL2014)
- Comp… 233-2…
- Tempo… Jared …
- Tempo… Kontch… (DL20…
- …

Inconsistency handling/ paraconsistent reasoning:

- Reasoning Efficiently with Ontologies and Rules in the Presence of Inconsistencies (Extended Abstract) Tobias Kaminski, Matthias Knorr, Joao Leite (DL2015)
- Explaining Query Answers under Inconsistency-Tolerant Semantics over Description Logic Knowledge Bases (Extended Abstract) Meghyn Bienvenu, Camille Bourgaux, François Goasdoué (DL2015)
- OBDA Using RL Reasoners and Repairing 729-733 Giorgos Stoilos (DL2014)
- Querying Inconsistent Description Logic Knowledge Bases under Preferred Repair Semantics 96-99 Camille Bourgaux, Meghyn Bienvenu, François Goasdoué (DL2014)
- …eylan,
- …adir
- …ng in …ina

Numerical Reasoning? Equations?

Closest related work on **DLs with concrete domains…**

- Snorocket 2.0: Concrete Domains and Concurrent Classification 32-38. Alejandro Metke-Jimenez, Michael Lawley (ORE2013)

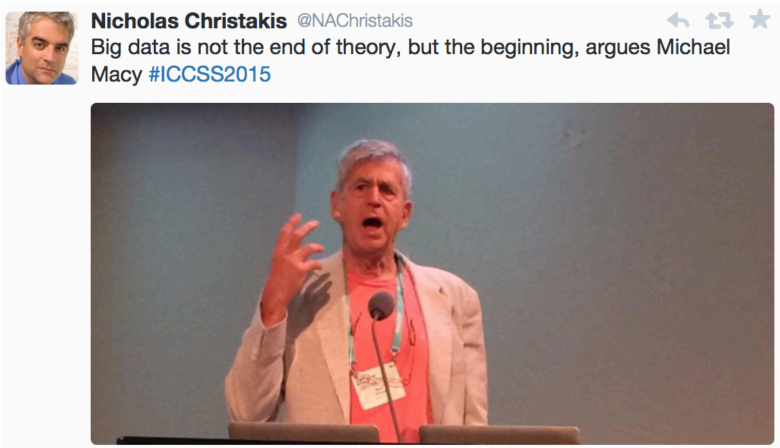
Concrete domains also supported in HERMIT, Fact++.

- Most foundational works 2005 and before…? E.g.: Tableau Algorithm for DLs with Concrete Domains and GCIs – DL2005 Carsten Lutz, Maja Milicic:

Integrating Open Data:
(How) Can Description Logics Help me?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

... even the computational social scientists don't buy that:



- Expressive ontology languages (plus e.g. equational knowledge) needed
- combination of reasoning about formal background knowledge & statistical methods needed
- **temporal aspects** need to be taken into account, but also **provenance**
- soundness/completeness (KRR) vs. coverage/accuracy (ML)
- "NoLD"... not only Linked Data

- Maybe you find our datasets useful:
- data.wu.ac.at/portalwatch
- citydata.wu.ac.at

EFMD EQUIS ACCREDITED