WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

EFMD
EQUIS
ACCREDITED

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

Axel Polleres, Sebastian Neumaier 13th Reasoning Web Summer School, 2017

# Target audience…

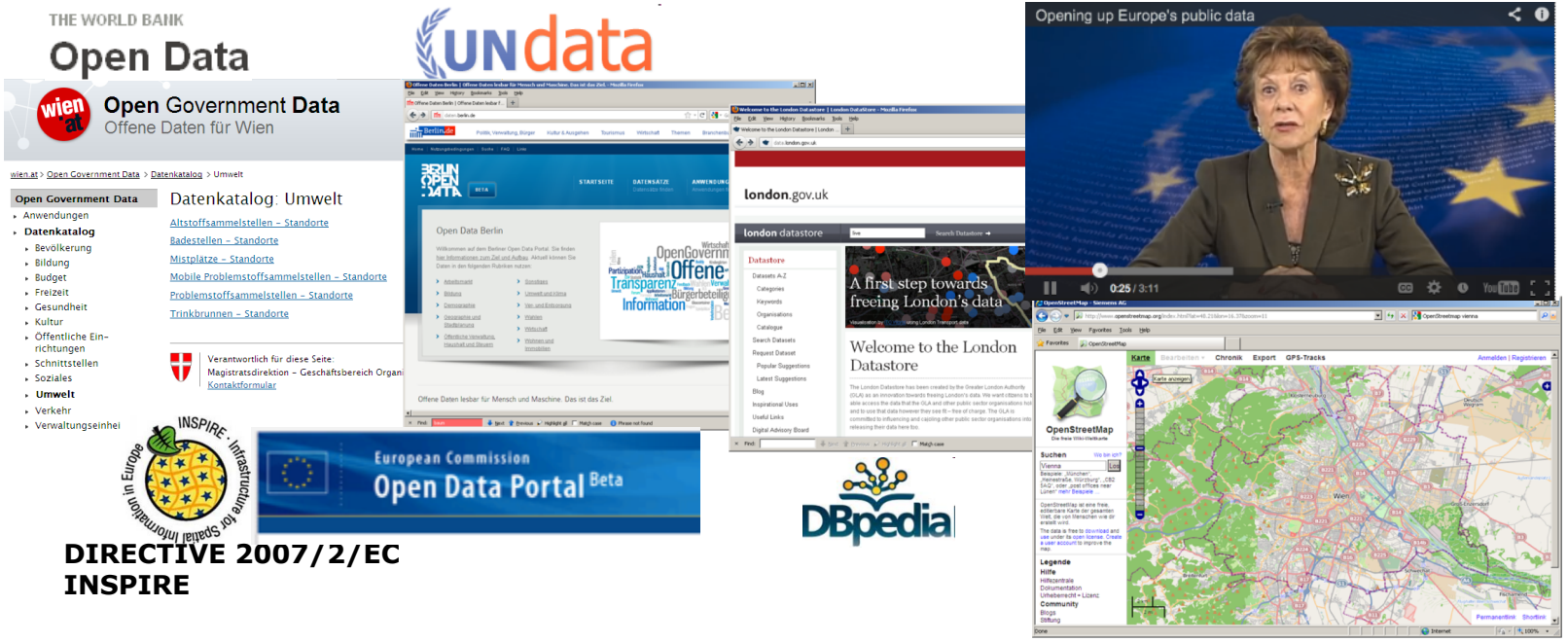
- **My vague assumptions:**
  - You all know about Logics, you know about Semantics and Reasoning…
  - … probably also about this thing called the “Semantic Web”
  - … maybe even about Linked Data?
  - But what if you actually start integrating (**Open**) data from the **Web**?
    … *Problems start at least one level below already*

# Goal of today's lecture: Talk about things which we have NOT yet solved!

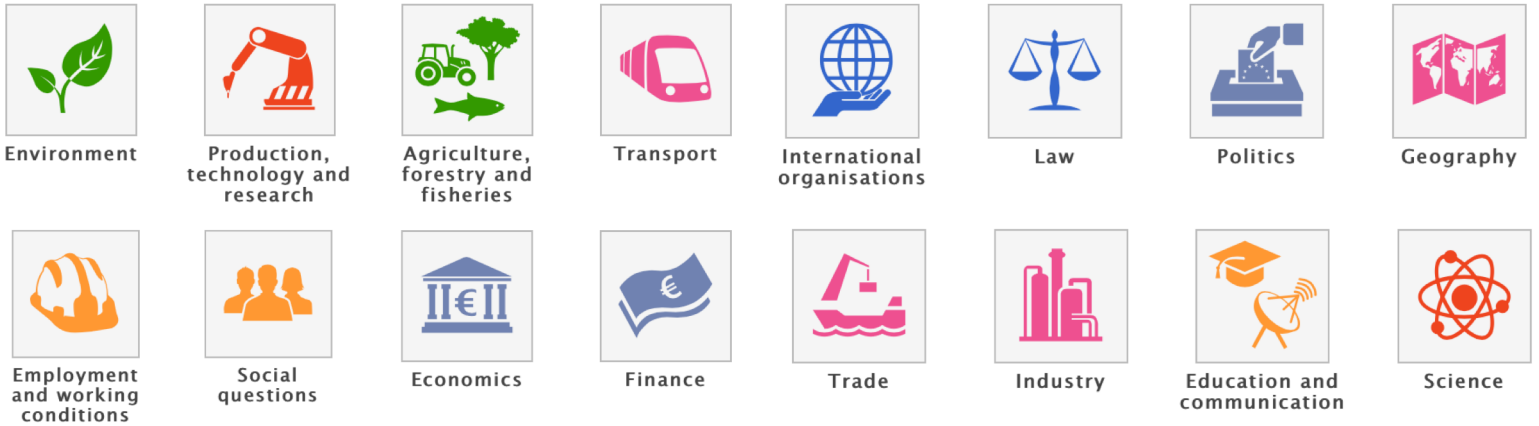
- **Part I:**
  - Where to find Open Data?
  - Dealing with "Low-level" data heterogeneity – Which formats are there on the Web?
  - Licenses and Provenance – Which data is actually "open"?
  - Quality Issues in Open Data
  - How to find Open data: Search over Open Data
- Part II:
  - How does reasoning help? A motivating Use Case.
  - Let's discuss how Rules & Reasoning can help? Group work!

# Open Data is a Global Trend!

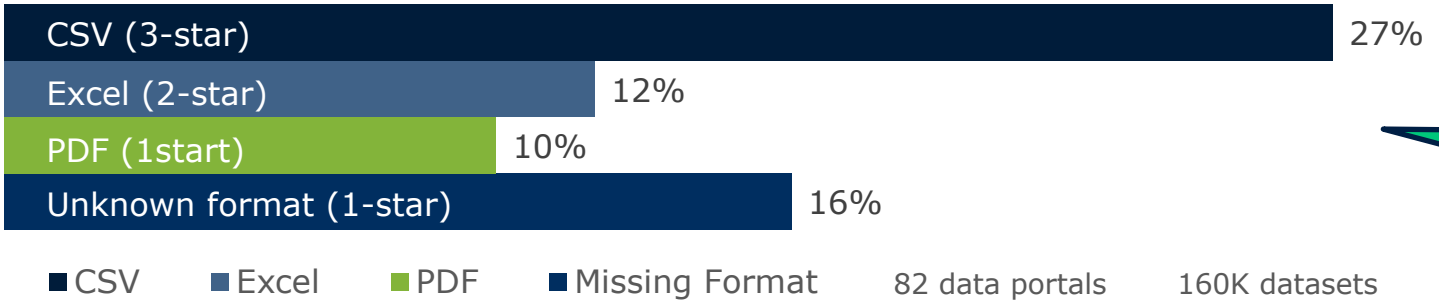
- EU & Austria are pushing Open Data!



# Open Data – Linked Data?




- Available data is only partially structured and not linked [1]:

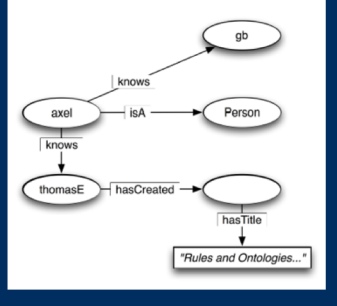


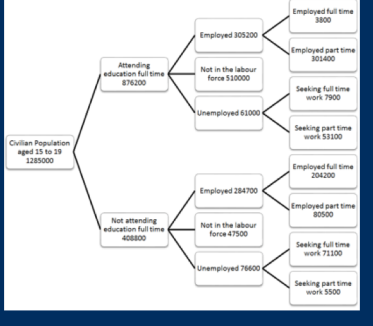

RDF? Not signicfiant

[1] Umbrich, J., Neumaier, S., Polleres, A.: Quality assessment & evolution of open data portals. International Conference on Open and Big Data (2015)

# Data Formats on the Web


| | A | B | C |
|---|---|---|---|
| 1 | roomcode | category_en | capacity |
| 2 | LC.0.000 | Event space | 1500 |
| 3 | TC.0.10 | Event space | 650 |
| 4 | LC.0.110 | Event space | 400 |
| 5 | SC.2.733 | Event space | 200 |
| 6 | LC.0.132 | Event space | 200 |
| 7 | TC.0.02 | Auditorium | 180 |
| 8 | TC.0.01 | Auditorium | 180 |
| 9 | TC.0.03 | Auditorium | 180 |
| 10 | TC.0.04 | Auditorium | 180 |
| 11 | TC.1.02 | Auditorium | 120 |
| 12 | TC.1.01 | Auditorium | 120 |
| 13 | | | |



# Different Formats

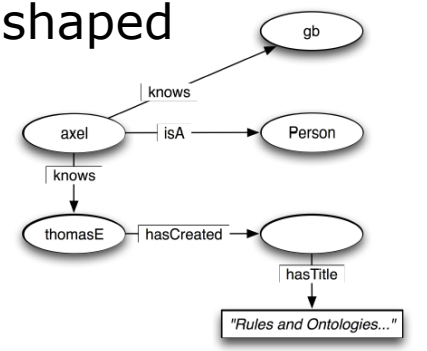
```
A N W E S E N D E .

VORSITZENDE:    Bürgermeister Mag. Siegfried NAGL
                      Gemeinderätin Gerda GESEK

Weiters 48 Mitglieder des Gemeinderates, und zwar:

BERGMANN Ingeborg
BRAUNERSREUTHER Christine
DREISIEBNER Karl
EBER Manfred
FABISCH Andreas Mag.
FRÖLICH Klaus Mag.
```

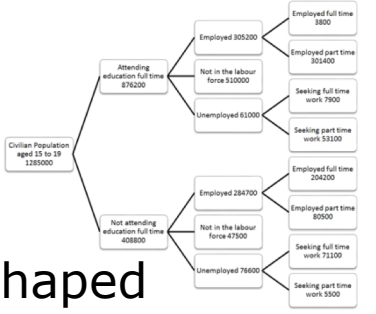
plain text

tabular

| | A | B | C |
|---|---|---|---|
| 1 | **roomcode** | **category_en** | **capacity** |
| 2 | LC.0.000 | Event space | 1500 |
| 3 | TC.0.10 | Event space | 650 |
| 4 | LC.0.110 | Event space | 400 |
| 5 | SC.2.733 | Event space | 200 |
| 6 | LC.0.132 | Event space | 200 |
| 7 | TC.0.02 | Auditorium | 180 |
| 8 | TC.0.01 | Auditorium | 180 |
| 9 | TC.0.03 | Auditorium | 180 |
| 10 | TC.0.04 | Auditorium | 180 |
| 11 | TC.1.02 | Auditorium | 120 |
| 12 | TC.1.01 | Auditorium | 120 |
| 13 | | | |

graph-shaped



tree-shaped

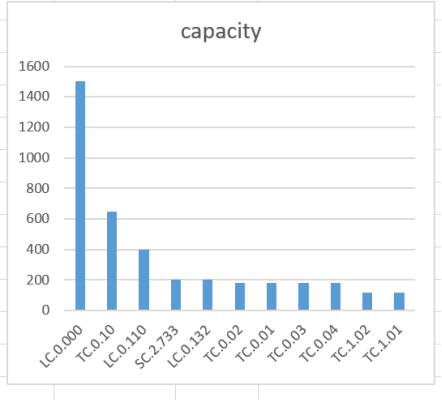
# Tabular

- CSV
  - Comma-separated values
  - Plain text
  - Variations on the format

```
course_id,semester,name
4000,15S,Marketing
4001,15S,Marketing
4002,15S,"Personal, Führung, Organisation"
4004,15S,Grundlagen der Volkswirtschaftslehre
4006,15S,Mathematik
4007,15S,Statistik
```

- Excel
  - .xls, .xlsx, .xlsm, … (binary)
  - Widely used but **proprietary**
  - Built-in visualisation

| | A | B | C |
|---|---|---|---|
| 1 | **roomcode** | **category_en** | **capacity** |
| 2 | LC.0.000 | Event space | 1500 |
| 3 | TC.0.10 | Event space | 650 |
| 4 | LC.0.110 | Event space | 400 |
| 5 | SC.2.733 | Event space | 200 |
| 6 | LC.0.132 | Event space | 200 |
| 7 | TC.0.02 | Auditorium | 180 |
| 8 | TC.0.01 | Auditorium | 180 |
| 9 | TC.0.03 | Auditorium | 180 |
| 10 | TC.0.04 | Auditorium | 180 |
| 11 | TC.1.02 | Auditorium | 120 |
| 12 | TC.1.01 | Auditorium | 120 |
| 13 | | | |



# Tree-based: Hierachies

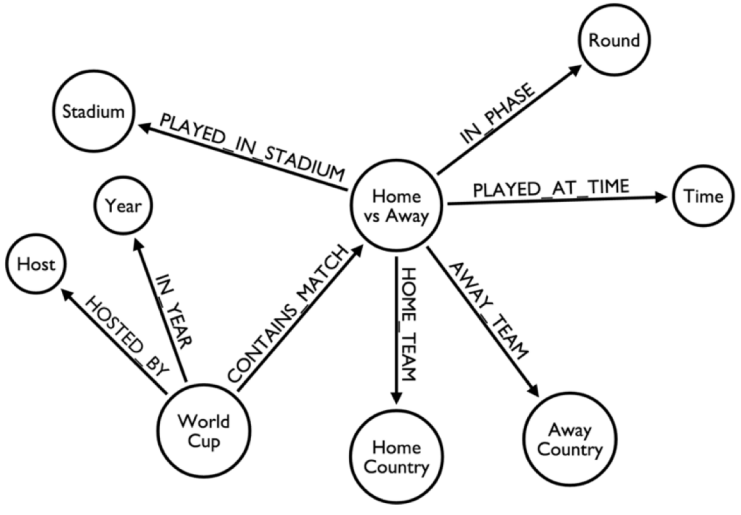
- XML
  - Markup language
  - Tag based (not predefined like HTML)
  - Human readable

- JSON
  - **J**ava**S**cript **O**bject **N**otation
  - Identical to JavaScript objects
  - Name/value pairs
  - Easy-to-use alternative to XML

```
<Gemeindedaten>
    <Gemeinde>
        <bgmname>Mag. Siegfried NAGL</bgmname>
        <email>stadtverwaltung@graz.at</email>
        <gemfax>+43 (316) 872-2369</gemfax>
        <gemname>Graz</gemname>
        <gemtel>+43 (316) 872-0</gemtel>
        <ort>Graz</ort>
        <plz>8010</plz>
        <strasse>Hauptplatz 1</strasse>
        <vbgmname>Mag.a Dr.in Martina SCHRÖCK</vbgmname>
        <webadr>http://www.graz.at</webadr>
    </Gemeinde>
</Gemeindedaten>
```

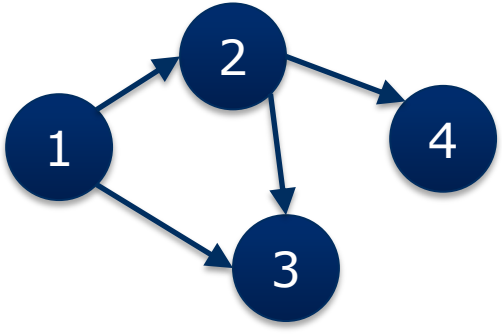
```
{
    id: "UNIVERSITAETOGD.762955",
    geometry: {
        type: "Point",
        coordinates: [
            16.408860446502068,
            48.21384341856702
        ]
    },
    properties: {
        NAME: "Wirtschaftsuniversität Wien",
        BEZEICHNUNG: "LC - Hauptgebäude"
    }
}
```

# Graph-shaped data formats





# Graph representation




- Edge-list
  1 2
  1 3
  2 3
  2 4

- Adjacency-list
  1: 2,3
  2: 3,4

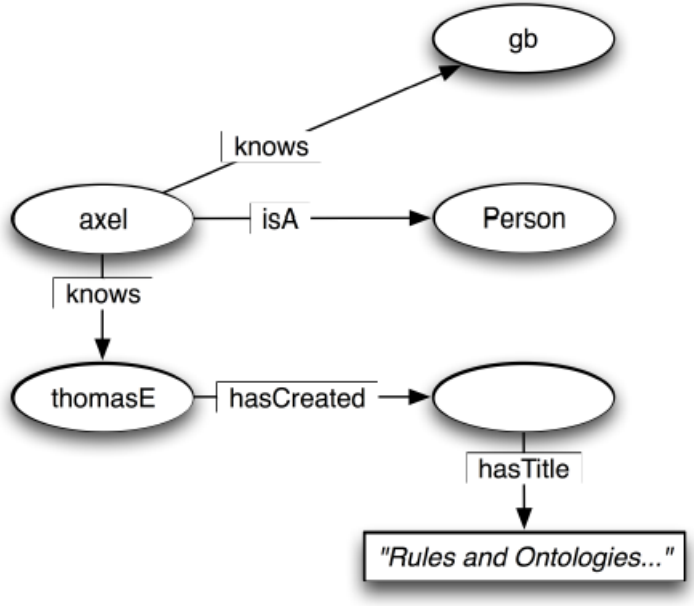
- Matrix

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 2 | 1 | 0 | 1 | 1 |
| 3 | 1 | 1 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 |

# Graph

- RDF (**R**esource **D**escription **F**ramework)
  - Describing resources per triples/statements
  - **S**ubject **P**redicate **O**bject

axel isA Person .
axel hasName “Axel Polleres”.
axel knows gb .
axel knows thomas.
thomas hasCreated an Article
titled “Rules and Ontologies …”.



# RDF

- Different syntaxes for RDF:
  - **RDF/XML**
  - Turtle
  - RDFa
  - JSON-LD
  - HDT (compressed)
- Accessing RDF:
  - Query language: SPARQL
  - Parser

```
<?xml version="1.0" encoding="UTF-8"?>

<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/" >
  <foaf:Person rdf:ID="me">
    <foaf:name>Axel Polleres</foaf:name>
    <foaf:title>Dr</foaf:title>
    <foaf:firstName>Axel</foaf:firstName>
    <foaf:lastName>Polleres</foaf:lastName>

    <foaf:knows>
      <foaf:Person>
        <foaf:name>Thomas Eiter</foaf:name>
      </foaf:Person>
    </foaf:knows>

  </foaf:Person>

</rdf:RDF>
```

RDF/XML

# RDF

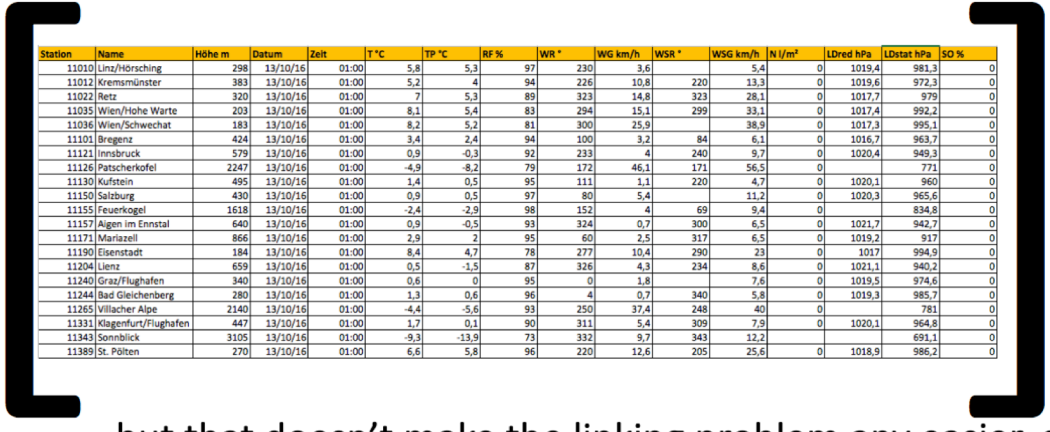
- Different syntaxes for RDF:
  - RDF/XML
  - **Turtle**
  - RDFa
  - JSON-LD
  - HDT (compressed)
- Accessing RDF:
  - Query language: SPARQL
  - Parser

```
@prefix : <http://polleres.net/foaf.df> .
@prefix rdfs: <http://…> .
@prefix foaf: <http://…> .
@prefix dc: <http://…>

:me a Person .
:me foaf:name “Axel Polleres”.
:me foaf:knows gb .
:me dc:creator [ rdfs:label “Rules and
Ontologies …” ] .
```

RDF/XML

# All formats are intertranslateable, but...

- Does this make sense to translate everything to RDF triples? e.g., I can trivially turn tabular data into RDF/Turtle



| Station | Name | Höhe m | Datum | Zeit | T °C | TP °C | RF % | WR ° | WG km/h | WSR ° | WSG km/h | N l/m² | LDred hPa | LDstat hPa | SO % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11010 | Linz/Hörsching | 298 | 13/10/16 | 01:00 | 5,8 | 5,3 | 97 | 230 | 3,6 | | 5,4 | 0 | 1019,4 | 981,3 | 0 |
| 11012 | Kremsmünster | 383 | 13/10/16 | 01:00 | 5,2 | 4 | 94 | 226 | 10,8 | 220 | 13,3 | 0 | 1019,6 | 972,3 | 0 |
| 11022 | Retz | 320 | 13/10/16 | 01:00 | 7 | 5,3 | 89 | 323 | 14,8 | 323 | 28,1 | 0 | 1017,7 | 979 | 0 |
| 11035 | Wien/Hohe Warte | 203 | 13/10/16 | 01:00 | 8,1 | 5,4 | 83 | 294 | 15,1 | 299 | 33,1 | 0 | 1017,4 | 992,2 | 0 |
| 11036 | Wien/Schwechat | 183 | 13/10/16 | 01:00 | 8,2 | 5,2 | 81 | 300 | 25,9 | | 38,9 | 0 | 1017,3 | 995,1 | 0 |
| 11101 | Bregenz | 424 | 13/10/16 | 01:00 | 3,4 | 2,4 | 94 | 100 | 3,2 | 84 | 6,1 | 0 | 1016,7 | 963,7 | 0 |
| 11121 | Innsbruck | 579 | 13/10/16 | 01:00 | 0,9 | -0,3 | 92 | 233 | 4 | 240 | 9,7 | 0 | 1020,4 | 949,3 | 0 |
| 11126 | Patscherkofel | 2247 | 13/10/16 | 01:00 | -4,9 | -8,2 | 79 | 172 | 46,1 | 171 | 56,5 | 0 | | 771 | 0 |
| 11130 | Kufstein | 495 | 13/10/16 | 01:00 | 1,4 | 0,5 | 95 | 111 | 1,1 | 220 | 4,7 | 0 | 1020,1 | 960 | 0 |
| 11150 | Salzburg | 430 | 13/10/16 | 01:00 | 0,9 | 0,5 | 97 | 80 | 5,4 | | 11,2 | 0 | 1020,3 | 965,6 | 0 |
| 11155 | Feuerkogel | 1618 | 13/10/16 | 01:00 | -2,4 | -2,9 | 98 | 152 | 4 | 69 | 9,4 | 0 | | 834,8 | 0 |
| 11157 | Aigen im Ennstal | 640 | 13/10/16 | 01:00 | 0,9 | -0,5 | 93 | 324 | 0,7 | 300 | 6,5 | 0 | 1021,7 | 942,7 | 0 |
| 11171 | Mariazell | 866 | 13/10/16 | 01:00 | 2,9 | 2 | 95 | 60 | 2,5 | 317 | 6,5 | 0 | 1019,2 | 917 | 0 |
| 11190 | Eisenstadt | 184 | 13/10/16 | 01:00 | 8,4 | 4,7 | 78 | 277 | 10,4 | 290 | 23 | 0 | 1017 | 994,9 | 0 |
| 11204 | Lienz | 659 | 13/10/16 | 01:00 | 0,5 | -1,5 | 87 | 326 | 4,3 | 234 | 8,6 | 0 | 1021,1 | 940,2 | 0 |
| 11240 | Graz/Flughafen | 340 | 13/10/16 | 01:00 | 0,6 | 0 | 95 | 0 | 1,8 | | 7,6 | 0 | 1019,5 | 974,6 | 0 |
| 11244 | Bad Gleichenberg | 280 | 13/10/16 | 01:00 | 1,3 | 0,6 | 96 | 4 | 0,7 | 340 | 5,8 | 0 | 1019,3 | 985,7 | 0 |
| 11265 | Villacher Alpe | 2140 | 13/10/16 | 01:00 | -4,4 | -5,6 | 93 | 250 | 37,4 | 248 | 40 | 0 | | 781 | 0 |
| 11331 | Klagenfurt/Flughafen | 447 | 13/10/16 | 01:00 | 1,7 | 0,1 | 90 | 311 | 5,4 | 309 | 7,9 | 0 | 1020,1 | 964,8 | 0 |
| 11343 | Sonnblick | 3105 | 13/10/16 | 01:00 | -9,3 | -13,9 | 73 | 332 | 9,7 | 343 | 12,2 | | | 691,1 | 0 |
| 11389 | St. Pölten | 270 | 13/10/16 | 01:00 | 6,6 | 5,8 | 96 | 220 | 12,6 | 205 | 25,6 | 0 | 1018,9 | 986,2 | 0 |

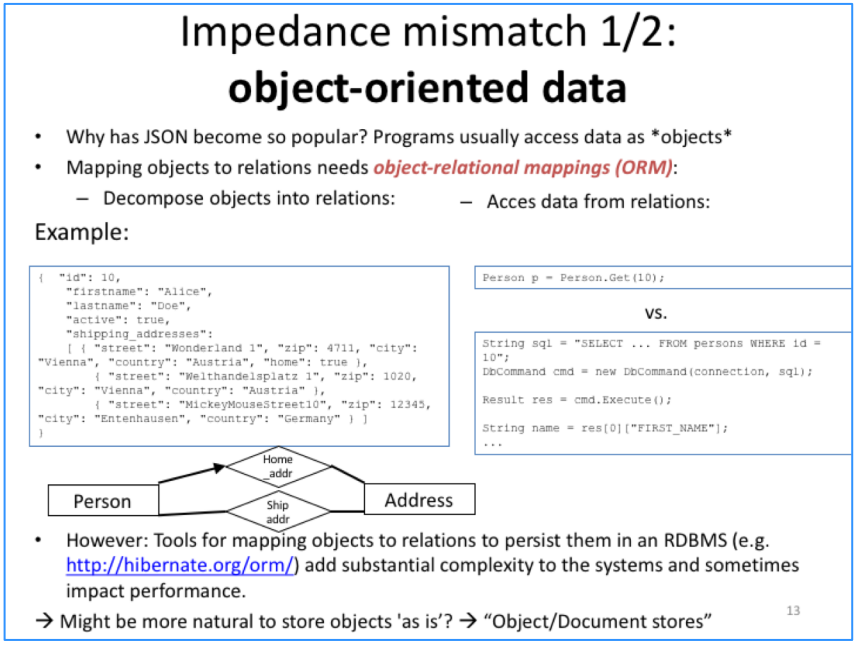
```
@prefix : <http://www.example.org/mynamespace/> .
:col1 rdfs:label "Station".
:col2 rdfs:label "Name".
[ :col1 "10010" ; :col2 "Linz/Hörsching"; … ].
```

… but that doesn't make the linking problem any easier, does it?

https://polleres.net/presentations/20161018COL2016D_Panel.pdf

# All formats are intertranslateable, but...

- Does it make sense to translate everything to tables?
- Does it make sense to translate everything to trees?
  - Probably neither… (recall your Databases 101: ***Impedance mismatch***)



Impedance mismatch 1/2:

**object-oriented data**

- Why has JSON become so popular? Programs usually access data as *objects*
- Mapping objects to relations needs *object-relational mappings (ORM)*:
  - Decompose objects into relations:
  - Acces data from relations:

Example:

```
{  "id": 10,
    "firstname": "Alice",
    "lastname": "Doe",
    "active": true,
    "shipping_addresses":
    [ { "street": "Wonderland 1", "zip": 4711, "city":
"Vienna", "country": "Austria", "home": true },
      { "street": "Welthandelsplatz 1", "zip": 1020,
"city": "Vienna", "country": "Austria" },
      { "street": "MickeyMouseStreet10", "zip": 12345,
"city": "Entenhausen", "country": "Germany" } ]
}
```

```
Person p = Person.Get(10);
```

vs.

```
String sql = "SELECT ... FROM persons WHERE id =
10";
DbCommand cmd = new DbCommand(connection, sql);

Result res = cmd.Execute();

String name = res[0]["FIRST_NAME"];
...
```



- However: Tools for mapping objects to relations to persist them in an RDBMS (e.g. http://hibernate.org/orm/) add substantial complexity to the systems and sometimes impact performance.

→ Might be more natural to store objects 'as is'? → "Object/Document stores"

13

*Bottomline: it makes sense to use and query data "as is" in many cases!*

# Linked Data…?

https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/

| | |
|---|---|
| ★ | Available on the web (whatever format) *but with an open licence, to be Open Data* |
| ★★ | Available as machine-readable structured data (e.g. excel instead of image scan of a table) |
| ★★★ | as (2) plus non-proprietary format (e.g. CSV instead of excel) |
| ★★★★ | All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff |
| ★★★★★ | All the above, plus: Link your data to other people's data to provide context |

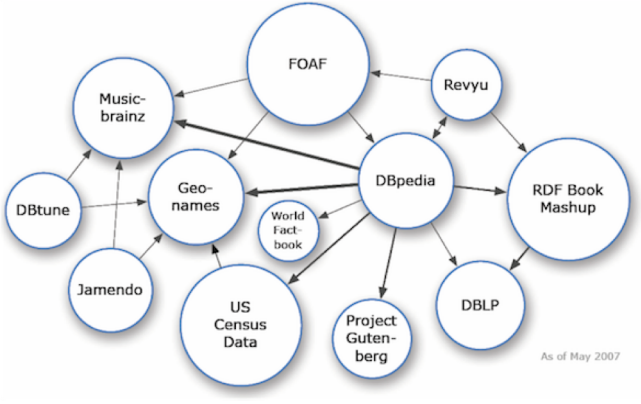
+

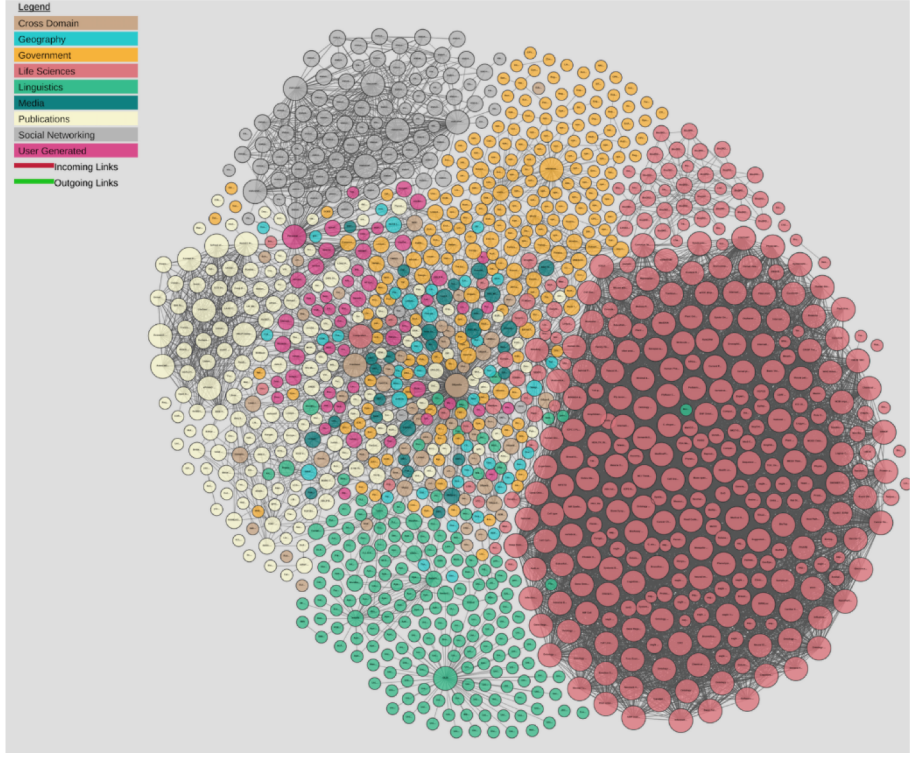
**Linked Data Principles**

- **LDP1:** use URIs as names for things
- **LDP2:** use HTTP URIs so those names can be dereferenced
- **LDP3:** return useful – RDF? – information upon dereferencing those URIs
- **LDP4:** include links using externally dereferenceable URIs.

https://www.w3.org/DesignIssues/LinkedData.html

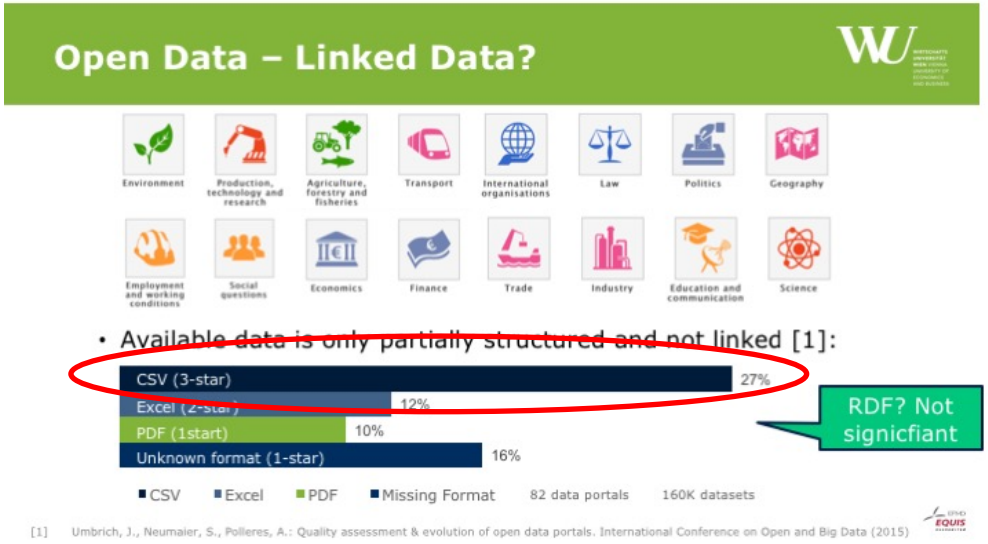
# Linked Data... growth since ~10 years



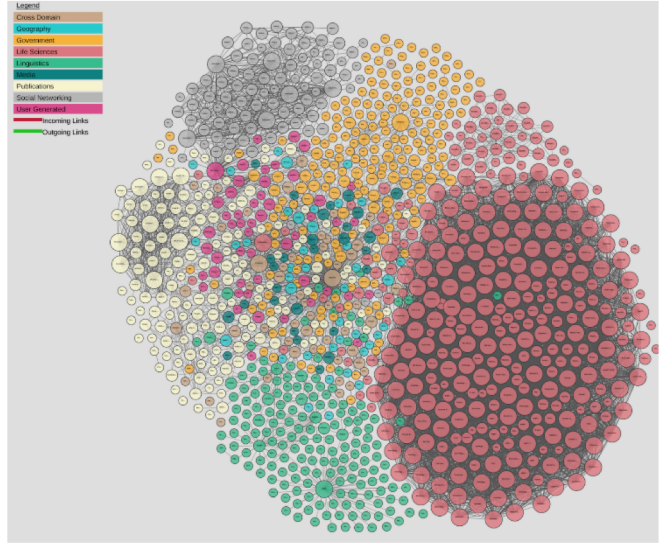




Linking Open Data cloud diagram 2007-2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

# Open Data... growing faster!




**413GB** ... Tabular **CSV data only**, on Open Data Portals, cf. http://data.wu.ac.at/portalwatch/



149,423,660,620 triples according to LODSTATS (**325GB** in compressed HDT format, cf. https://datahub.io/dataset/lod-a-lot)

Linking Open Data cloud diagram 2007-2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

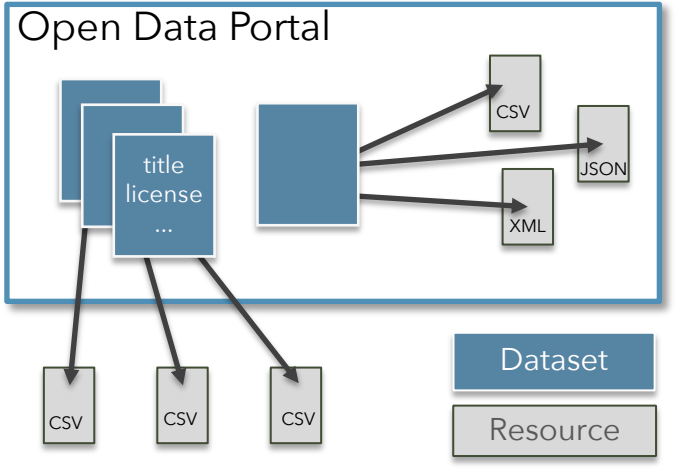
# Where to find Web Data?

# Open Data Portals

- Catalogues for datasets
  - "rich" meta data about the content of the data
- Offer search & filter functionality
  - Search **about** datasets (tags, license, title)
  - No search **in** datasets (e.g., datasets containing the word "Bloomsbury")
- E.g. >140 CKAN instances around the world, in total we index E.g. over 260 active Open Data catalogues in our portalwatch project: http://data.wu.ac.at/portalwatch/

# Open Data Portals

- Single point of access
  - APIs
- Meta data
  - Licenses, Provenance, Formats, …
- Typical software

# E.g.: data.gv.at (CKAN)

Open Data Portal by the Austrian Government

| Titel | Veröffentlichende Stelle / Datenverantwortliche Stelle | Veröffentlicht auf data.gv.at am | Letzte Änderung | Format |
|---|---|---|---|---|
| Bevölkerung nach Alter und Geschlecht<br>Bevölkerung nach Alter, Geschlecht und Wohngemeinde zum 1.1. des jeweiligen Jahres | Land Niederösterreich / Abteilung Raumordnung und Regionalpolitik-Statistik | 11.04.2013 | 10.05.2015 | CSV |
| Bevölkerung nach Gemeinden - Volkszählungen<br>Bevölkerung nach Wohngemeinden zum Volkszählungsstichtag des jeweiligen Jahres | Land Niederösterreich / Abteilung Raumordnung und Regionalpolitik-Statistik | 11.04.2013 | 10.05.2015 | CSV |
| Landtagswahl 2013<br>Wahlergebnisse der Landtagswahl 2013 \\ der Hinweis in der Attributbeschreibung ! N a c h f o l g e ... | Land Niederösterreich / Abteilung Staatsbürgerschaft und Wahlen | 28.10.2014 | 10.05.2015 | CSV |
| Wanderungen nach Gemeinden<br>Zuzüge, Wegzüge und Gemeindebinnenwanderungen im jeweiligen Jahr | Land Niederösterreich / Abteilung Raumordnung und Regionalpolitik-Statistik | 11.04.2013 | 10.05.2015 | CSV |

https://www.data.gv.at/katalog/api/rest/dataset


```
[
    "abaenderungsantrag-nr",
    "abfallentsorgung-containerstandpl-tze-engerwitzdorf",
    "abfallentsorgung-engerwitzdorf",
    "abfluss-und-seepegel-land-salzburg",
    "abwasser-behandlungsanlagen-land-salzburg",
    "abwasserwirtschaft-engerwitzdorf",
    ressen-tirol",
    ressinformationen-tirol",
    ressliste-gemeinnutziger-bauvereinigungen",
    rzte-engerwitzdorf",
    rarstrukturerhebung-2010-uberblick",
    tuelleozondatenoesterreich",
    tuelle-verkehrsbehinderungen",
    koholberatungsstellen-in-oberosterreich",
    lg-bildende-pflichtschulen-in-oo",
    lgemein-bildende-h-here-schulen",
    lgemein-bildende-pflichtschulen-in-tirol",
    ten-und-pflegeheime-in-oberosterreich",
    ternative-energiegewinnungsanlagen-in-oberosterreich-bewil
    phibienwanderstrecken-an-no-strassen",
    sarbeitsmarktdaten",
    sausundweiterbildung",
```

# CKAN Metadata (JSON)




```
{
 "license_title": "Creative Commons Namensnennung",
 "maintainer": "Stadtvermessung Graz",
 "author": "",
 "author_email": "stadtvermessung@stadt.graz.at",
 "resources": [
   {
   "size": "6698",
       "format": "CSV",
       "mimetype": "",
       "url": "http://data.graz.gv.at/.../Bibliothek.csv"
   }
 ],
 "tags": [
     "bibliothek", "geodaten", "graz", "kultur", "poi"
 ],
 "license_id": "CC-BY-3.0",
 "organization": null,
 "name": "bibliotheken",
 "notes": "Standorte der städtischen Bibliotheken...",
 "extras": {
     "Sprache des Metadatensatzes": "ger/deu Deutsch"
 },
   "license_url": "http://creativecommons.org/.../by/3.0/at/",
}
```


core keys

resource keys

extra keys

Possible challenges?

# CKAN Metadata mapped to RDF (DCAT)

http://data.wu.ac.at/portalwatch/api/v1/memento/data_gv_at/4904b8e4-a002-4bcd-bda3-d496eab8fda4/dcat

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix vcard: <http://www.w3.org/2006/vcard/ns#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://www.data.gv.at/katalog/dataset/4904b8e4-a002-4bcd-bda3-d496eab8fda4> a dcat:Dataset ;
    dcterms:description """Standorte der städtischen Bibliotheken.  \r
(Informationen zum [Koordinatensystem] (http://data.graz.gv.at/koordinaten))."""";
    dcterms:identifier "4904b8e4-a002-4bcd-bda3-d496eab8fda4" ;
    dcterms:issued "2014-09-07T23:02:41.374532"^^xsd:dateTime ;
    dcterms:modified "2017-02-20T00:04:23.775498"^^xsd:dateTime ;
    dcterms:publisher <http://www.data.gv.at/katalog/organization/fdedbd18-1dab-46be-b356-71278176c27c> ;
    dcterms:title "Bibliotheken" ;
    dcat:contactPoint [ a vcard:Organization ;
            vcard:fn "http://www.graz.at/cms/beitrag/10020470/311392/" ;
            vcard:hasEmail "stadtvermessung@stadt.graz.at" ] ;
    dcat:distribution <http://www.data.gv.at/katalog/dataset/4904b8e4-a002-4bcd-bda3-d496eab8fda4/resource/0cbd59c1-6c8d-4ea
        <http://www.data.gv.at/katalog/dataset/4904b8e4-a002-4bcd-bda3-d496eab8fda4/resource/1621a9c1-15f3-40cc-913c-2ad649a
    dcat:keyword "bibliothek",
        "geodaten",
        "graz",
        "kultur",
        "poi" .

<http://www.data.gv.at/katalog/dataset/4904b8e4-a002-4bcd-bda3-d496eab8fda4/resource/1621a9c1-15f3-40cc-913c-2ad649a37e6e> a
    dcterms:description "Standorte der städtischen Bibliotheken" ;
    dcterms:format [ a dcterms:MediaTypeOrExtent ;
            rdfs:label "CSV" ] ;
    dcterms:license <https://creativecommons.org/licenses/by/3.0/at/deed.de> ;
    dcterms:title "Standorte der städtischen Bibliotheken" ;
    dcat:accessURL "http://data.graz.gv.at/katalog/bildung und forschung/Bibliothek.csv" ;
    dcat:byteSize 6698.0 .
```

# Selected CKAN Examples

- Datahub.io
  - Free, powerful data management platform
  - Many public datasets
- European Data Portal
  - Single point of access to data from several national CKAN data portals
- OpenData@WU
  - First open data portal of an Austrian University

# Many other interesting Open Data sources, examples:

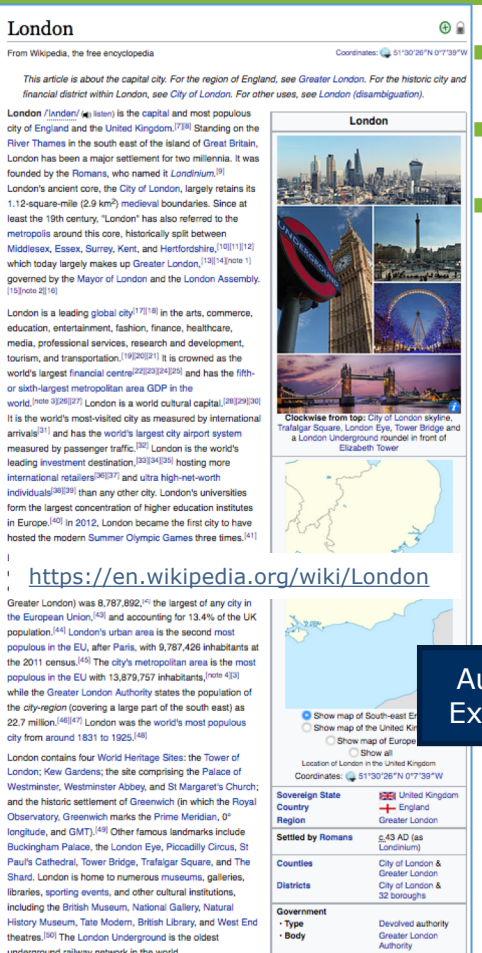

- DBpedia: Structured data from Wikipedia


- WIKIDATA: Open, shared database of the world's knowledge


- GEONAMES: geographical database covers all countries and contains over eight million placenames

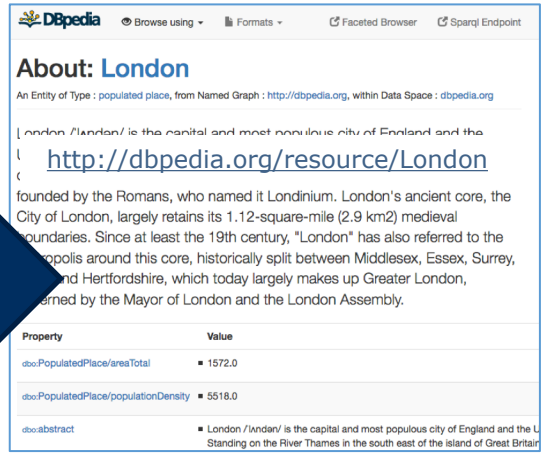

- OPENSTREETMAP: Free, editable map of the world

# DBpedia




- One of the central datasets of the LOD-Cloud
- RDF extracted from Wikipedia-Infoboxes
- SPARQL endpoint, e.g.:
  - „Cities in the UK with more than 1M population“:

Automatic Exctractors



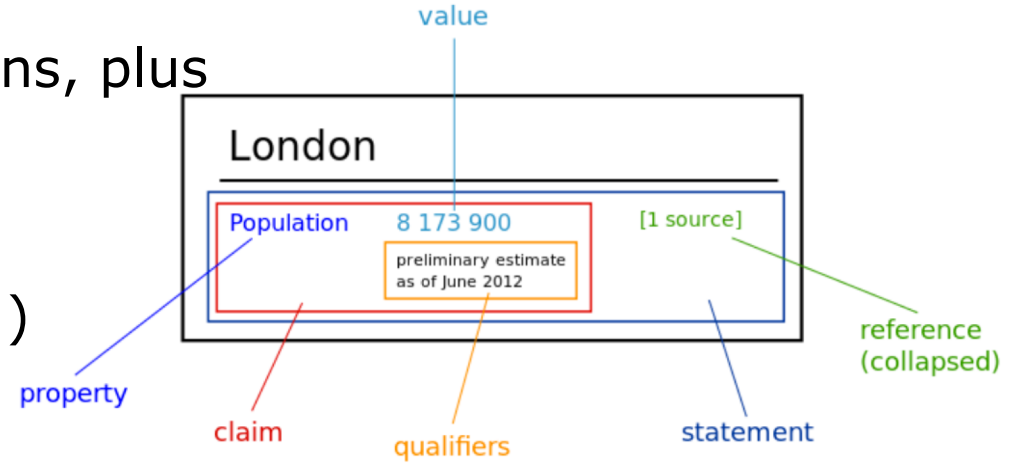

## Structured queries (SPARQL):

http://dbpedia.org/sparql

```
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX yago: <http://dbpedia.org/class/yago/>

SELECT DISTINCT ?city ?pop WHERE {
   ?city a yago:City108524735 .
   ?city dbo:country :United_Kingdom.
   ?city dbo:populationTotal ?pop

   FILTER ( ?pop > 1000000 )
}
```

# Wikidata

- Different concept:
  - “data **for** infoboxes", instead of "data **from** infoboxes"
  - Crowd-curated database of observations
- Statememnts
  - Entity-attribute-value observations, plus
  - versioning, ie. timestamps
  - provenance, ie. Who? Source?
  - Export to RDF (needs reification!)




https://de.slideshare.net/_Emw/an-ambitious-wikidata-tutorial/

# Wikidata as RDF … can be queried by SPARQL

- "Simple" surface query:

```
SELECT DISTINCT ?city WHERE {
    {
        ?city wdt:P31/wdt:P279* wd:Q515 .
        ?city wdt:P1082 ?population .
        ?city wdt:P17 wd:Q145 .
        FILTER (?population > 1000000)
    }
}
```

- What's this?

**instance of (P31)**
that class of which this subject is a particular example and member. (Subject typically an individual member with Proper Name label.) Different from P279 (subclass of).

**subclass of (P279)**
all instances of these items are instances of those items; this item is a class (subset) of that item. Not to be confused with Property:P31 (instance of).

**city (Q515)**
large and permanent human settlement

**population (P1082)**
number of people inhabiting the place; number of people of subject

**country (P17)**
sovereign state of this item

**United Kingdom (Q145)**
country in Europe

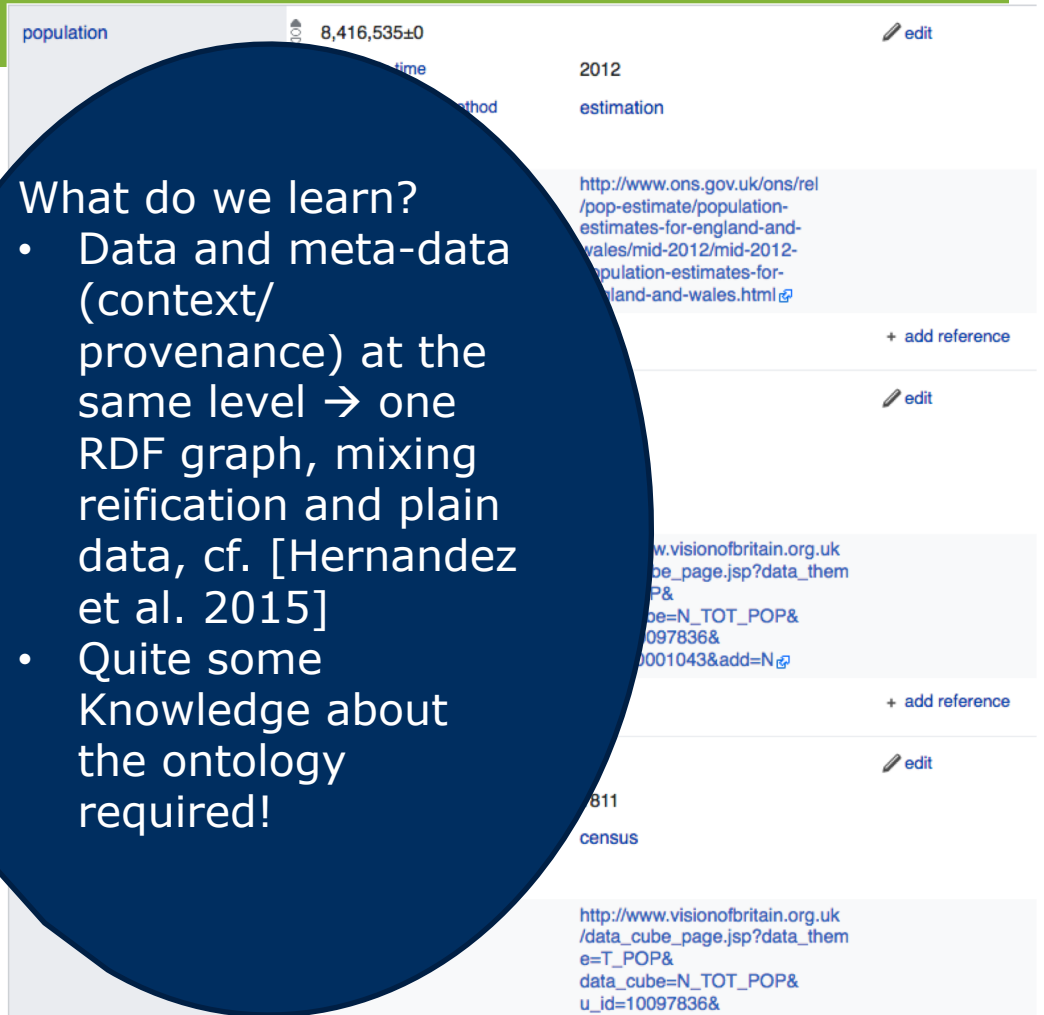
# Wikidata as RDF … can be queried by SPARQL

- However, Wikidata has more complex info: (**temporal** context, **provenance**,…
  - London: https://www.wikidata.org/wiki/Q84

… Can I query that with SPARQL? Yes!

Wikidata Query

```
SELECT ?city (min(?time) as ?year) WHERE {
    ?city wdt:P31/wdt:P279* wd:Q515.
    ?city wdt:P17 wd:Q145 .
    ?city p:P1082 ?statement .
    ?statement <http://www.wikidata.org/prop/statement/value/P1082> ?value;
       <http://www.wikidata.org/prop/qualifier/P585> ?time .
     ?value <http://wikiba.se/ontology#quantityAmount> ?population .
     FILTER (?population > 1000000)
}
GROUP BY ?city
```

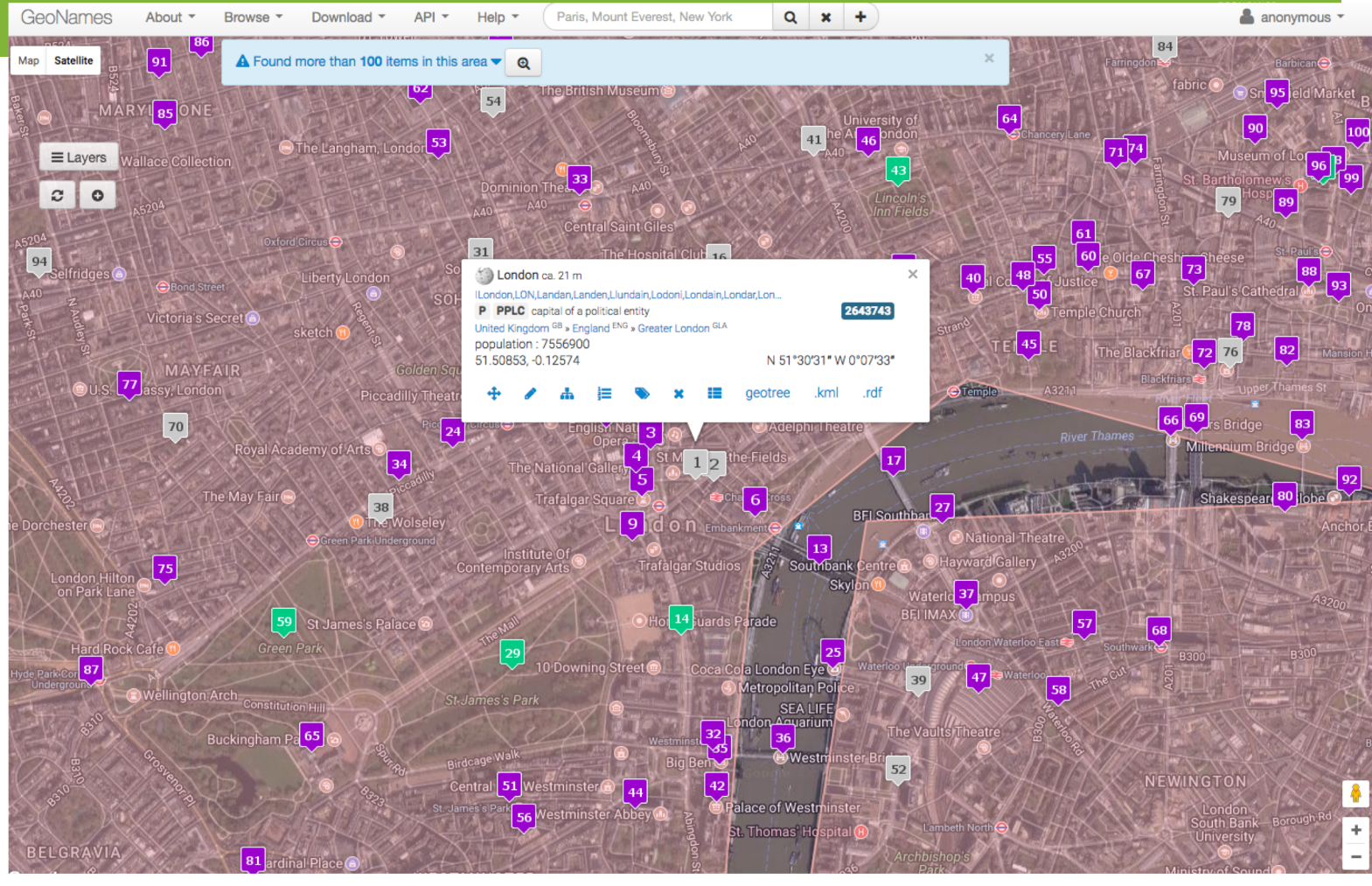



What do we learn?

- Data and meta-data (context/ provenance) at the same level → one RDF graph, mixing reification and plain data, cf. [Hernandez et al. 2015]
- Quite some Knowledge about the ontology required!

# Geonames

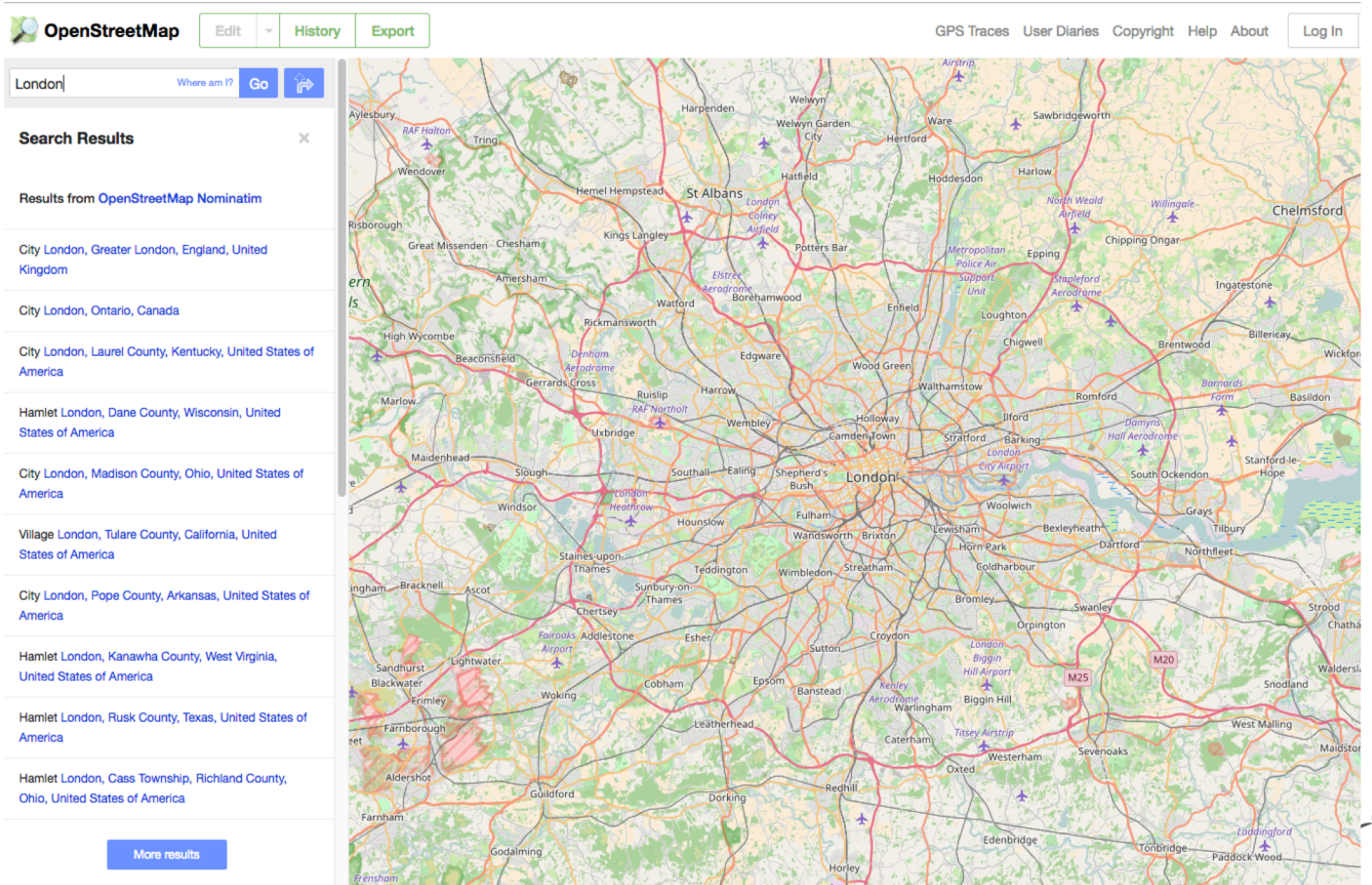
- Also might help for various use cases, API for geo-search:
  - Coordinates, shapes, also population data
  - Plus, it contains an implicit taxonomy of spatial entities (regions, countries, etc.)
  - Accessible via an API (XML, JSON, partially also RDF)



# Open Streetmap

Richer Sourcem, many more features than geonames, available via an XML API

- Land use,
- Road network,
- etc.



WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

# Licensing and Provenance of Data

EFMD EQUIS ACCREDITED

# Motivation: Who does the data on Wikipedia belong to? What license is it?

https://en.wikipedia.org/wiki/Wikipedia:Copyrights

## Wikipedia:Copyrights

From Wikipedia, the free encyclopedia

✓ **This page documents a Wikipedia policy with legal considerations.**

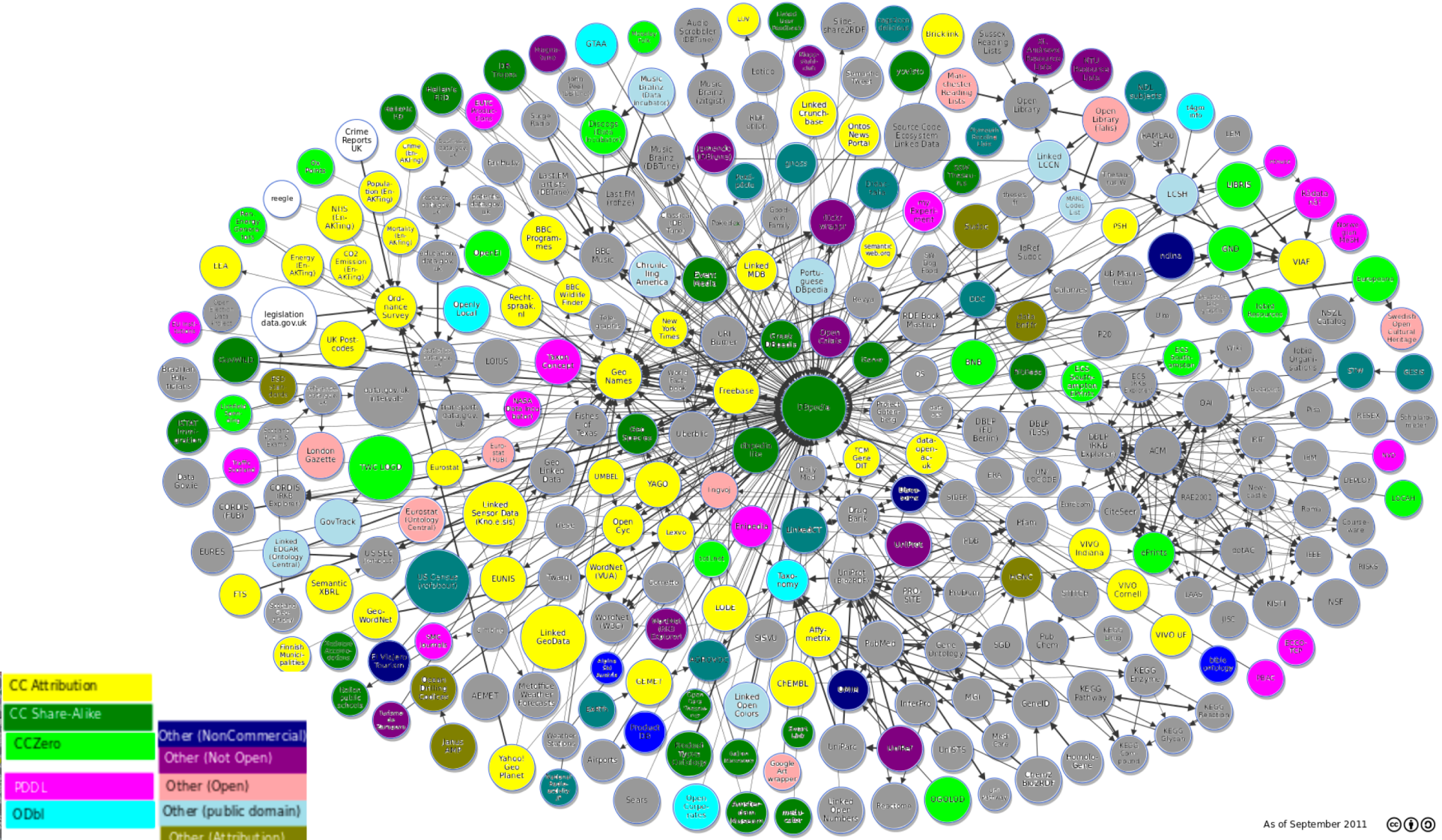
**Shortcuts:**
**WP:C**
**WP:COPY**
**WP:COPYRIGHT**

*"WP:COPY" redirects here. You may be looking for Wikipedia:Copyright Problems (shortcut:WP:CP), Wikipedia:How to copy-edit (shortcut: WP:COPYEDIT) or Wikipedia:Copying within Wikipedia (shortcut: WP:COPYWITHIN).*
*"WP:C" redirects here. You may be looking for Wikipedia:Consensus (shortcut: WP:CON), Wikipedia:Civility (shortcut: WP:CIV), Wikipedia:Categorization (shortcut: WP:CAT), Wikipedia:WikiProject Countries (shortcut: WP:COUNTRIES) or Wikipedia:WikiProject Council (shortcut: WP:COUNCIL).*

**Important note:** The Wikimedia Foundation does not own copyright on Wikipedia article texts or illustrations. **It is therefore pointless to email**

**Wikipedia copyright**

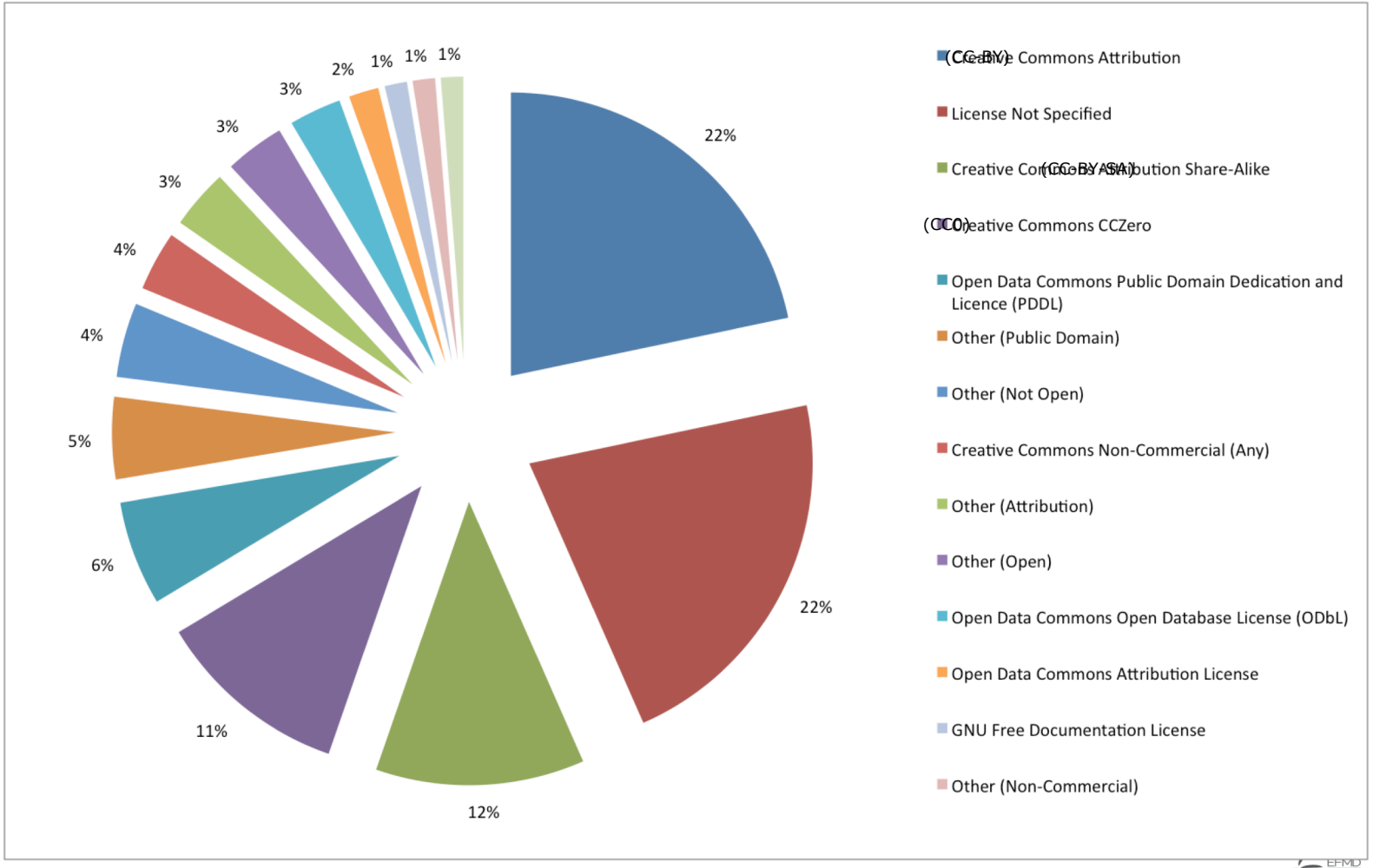
**Policy**




http://ns.inria.fr/l4lod/v2/l4lod_v2.html

# Different Licences in Open Data: *Open* != *Open - LOD*

Most widely used OD-licences in LOD:
https://opendatacommons.org/licenses/pddl/
https://creativecommons.org/




http://ns.inria.fr/l4lod/v2/l4lod_v2.html

# Different Licences in Open Data– *OD portals*

Note: many country-specific versions of https://creativecommons.org/, e.g. CC-BY-AT, OGL-UK, ca-ogl-lgo

(Austrian version of CC-BY, UK Open Government Data License, etc)

Table 4: Top-10 licenses.

| license_id | \|datasets\| | % | \|portals\| |
|---|---|---|---|
| ca-ogl-lgo | 239662 | 32.3 | 1 |
| notspecified | 193043 | 26 | 71 |
| dl-de-by-2.0 | 55117 | 7.4 | 7 |
| **CC-BY-4.0** | 49198 | 6.6 | 84 |
| **us-pd** | 35288 | 4.8 | 1 |
| **OGL-UK-3.0** | 33164 | 4.5 | 18 |
| other-nc | 27705 | 3.7 | 21 |
| **CC0-1.0** | 9931 | 1.3 | 36 |
| dl-de-by-1.0 | 9608 | 1.3 | 6 |
| Europ.Comm.[27] | 8604 | 1.2 | 2 |
| others | 80164 | 10.8 | |

Table 5: Open Definition conformant data licenses [40]

| License |
|---|
| Creative Commons Zero (CC0) |
| Creative Commons Attribution 4.0 (CC-BY-4.0) |
| Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0) |
| Open Data Commons Attribution License (ODC-BY) |
| Open Data Commons Public Domain Dedication and Licence (ODC-PDDL) |
| Open Data Commons Open Database License (ODC-ODbL) |

# Problem: How to License Derivativ/combined Works?


License Spectrum

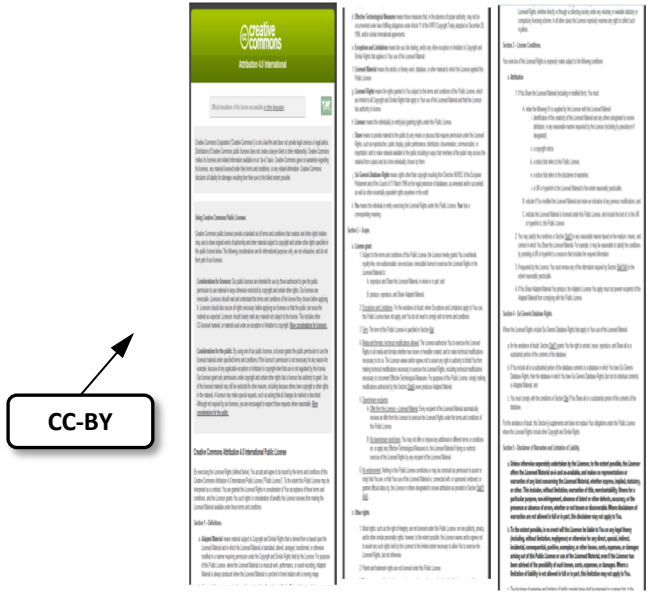
ODbL
CC-BY
CC-BY-SA
IMDB License
TomTom License

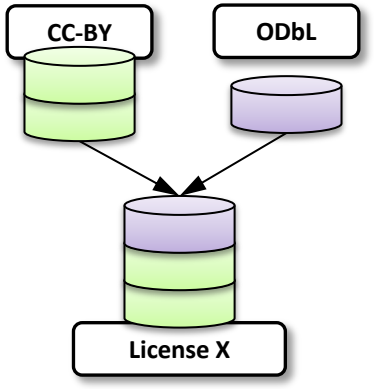
Public Domain | Strongly Protective | Weakly Protective | Permissive | Proprietary

No rights reserved ---------- Some rights reserved ---------- All rights reserved

# Problem: How to License Derivativ/combined Works?

- The reuse of data, software or content is often accompanied with legal uncertainty with respect to intellectual property rights and privacy issues.

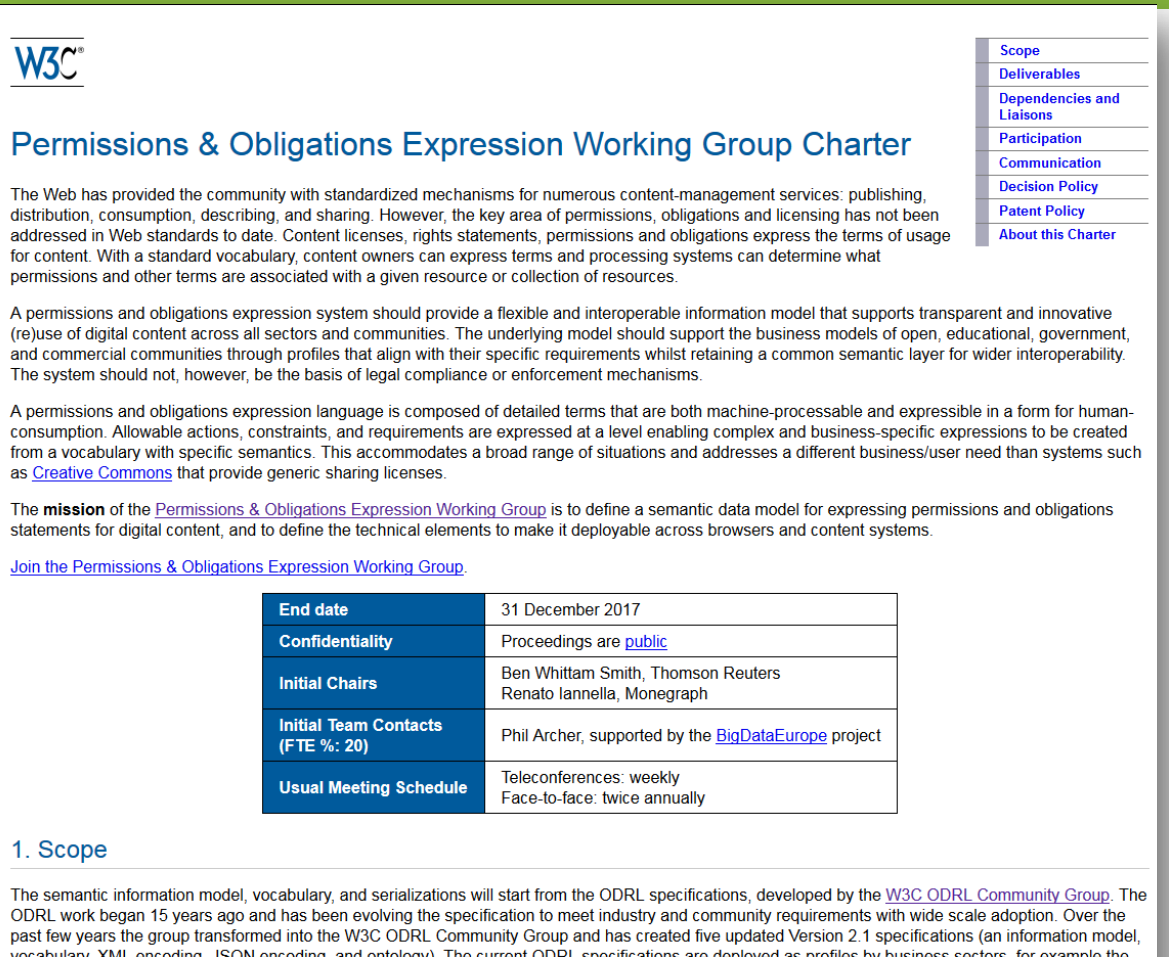



1. Are **CC-BY** and **ODbL** compatible?
2. What is the **semantics** of **License X** = **CC-BY** ∪ **ODBL**

# Data Licence Representation

# Open Digital Rights Language (ODRL)



W3C

## Permissions & Obligations Expression Working Group Charter



The Web has provided the community with standardized mechanisms for numerous content-management services: publishing, distribution, consumption, describing, and sharing. However, the key area of permissions, obligations and licensing has not been addressed in Web standards to date. Content licenses, rights statements, permissions and obligations express the terms of usage for content. With a standard vocabulary, content owners can express terms and processing systems can determine what permissions and other terms are associated with a given resource or collection of resources.

A permissions and obligations expression system should provide a flexible and interoperable information model that supports transparent and innovative (re)use of digital content across all sectors and communities. The underlying model should support the business models of open, educational, government, and commercial communities through profiles that align with their specific requirements whilst retaining a common semantic layer for wider interoperability. The system should not, however, be the basis of legal compliance or enforcement mechanisms.

A permissions and obligations expression language is composed of detailed terms that are both machine-processable and expressible in a form for human-consumption. Allowable actions, constraints, and requirements are expressed at a level enabling complex and business-specific expressions to be created from a vocabulary with specific semantics. This accommodates a broad range of situations and addresses a different business/user need than systems such as Creative Commons that provide generic sharing licenses.

The **mission** of the Permissions & Obligations Expression Working Group is to define a semantic data model for expressing permissions and obligations statements for digital content, and to define the technical elements to make it deployable across browsers and content systems.

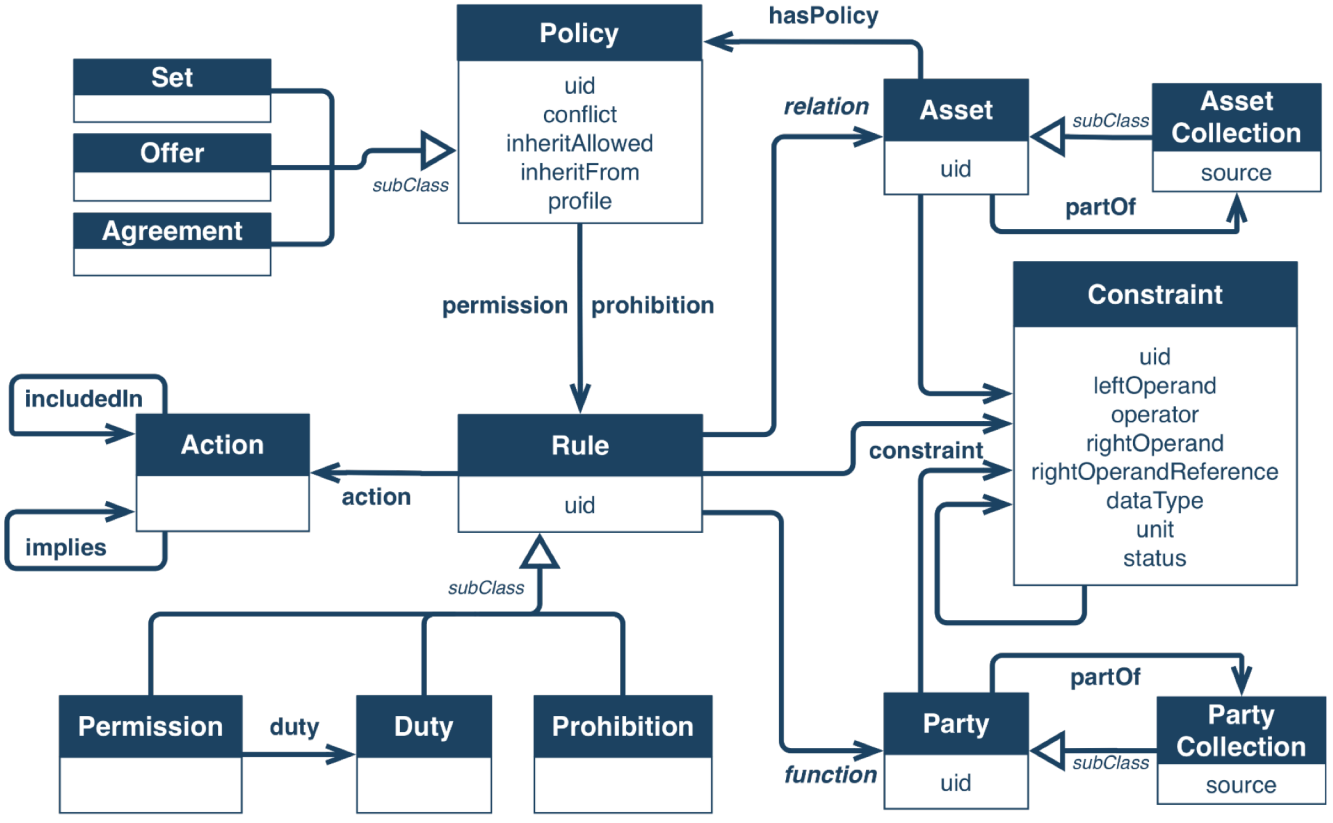
Join the Permissions & Obligations Expression Working Group.

| | |
|---|---|
| **End date** | 31 December 2017 |
| **Confidentiality** | Proceedings are public |
| **Initial Chairs** | Ben Whittam Smith, Thomson Reuters<br>Renato Iannella, Monegraph |
| **Initial Team Contacts (FTE %: 20)** | Phil Archer, supported by the BigDataEurope project |
| **Usual Meeting Schedule** | Teleconferences: weekly<br>Face-to-face: twice annually |

### 1. Scope

The semantic information model, vocabulary, and serializations will start from the ODRL specifications, developed by the W3C ODRL Community Group. The ODRL work began 15 years ago and has been evolving the specification to meet industry and community requirements with wide scale adoption. Over the past few years the group transformed into the W3C ODRL Community Group and has created five updated Version 2.1 specifications (an information model, vocabulary, XML encoding, JSON encoding, and ontology). The current ODRL specifications are deployed as profiles by business sectors, for example the

# Open Digital Rights Language (ODRL)




http://w3c.github.io/poe/model/

# ODRL Concepts 1/5
## Policy

## Policy Types

A **Policy** is the central entity that forms ODRL policy expressions. It can refer to **Permissions** and **Prohibitions** which hold for that **Policy** and be further distinguished into several subtypes such as:

### Offer

A **Policy** which proposes terms of usage from an **Assigner** to possible **Assignees**.

*Alice allows anyone who pays her 20€ to read her dataset.*

### Agreement

A **Policy** which contains all terms of usage between both an **Assigner** and an **Assignee** about an **Asset.**

*Alice allows Bob to read her dataset if he pays her 20€.*

### Request

A **Policy** which proposes terms of usage from an **Assignee** to an **Assigner** (owner of the **Asset).**

*Bob wants to pay Alice 20€ if she allows him to read her dataset.*

### Set

A **Policy** which defines **Prohibitions**, **Permissions** and/or **Duties** for a certain **Asset**.

*Dataset XY is licensed under CC-BY.*
*(i.e. duty to attribute asset owner)*

# ODRL Concepts 2/5
## Permission, Prohibition, and Duty

### Permission
A **Permission** specifies **Actions** which are allowed to be executed on a certain **Asset**.

*Alice allows Bob to read her dataset.*

### Prohibition
**Prohibitions** are used to forbid specific **Actions** on an **Asset** and cannot refer to **Duties**.

*Alice prohibits Bob to distribute her dataset.*

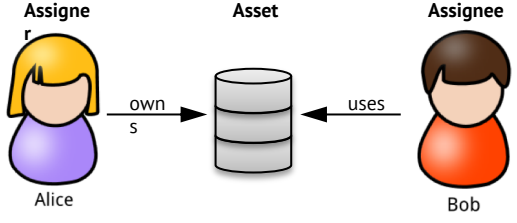
### Duty
**Duties** define **Actions** that have to be perfomed so that sorrounding **Permissions** or **Policies** become valid.

*Alice allows Bob to read her dataset if he pays her 20€.*

# ODRL Concepts 3/5
## Asset and Party

### Asset
An **Asset** is the entity whose terms of usage are restricted by its surrounding policy expression. In the domain of *Linked Data* an **Asset** usually represents a dataset or parts of a dataset (triples) .

### Party
A **Party** can be distinguished into **Assigners** (the parties who propose the policy statements) and **Assignees** (the ones who receive the policy statements).



## Action

**Actions** are operations on **Assets** that a potential **Assignee** is allowed (if related to a **Permission**), is prohibited (if related to a **Prohibition**) or has (if related to a **Duty**) to perform. In a Linked Data scenario actions could be, e.g.:

**use** – generic action for use of an assset

**transfer** – generic action of transferring the ownership of the asset (e.g. sell).

**Others more specific actions (non-ODRL core, ODRL „common vocabulary“ profile):**

**aggregate** – can be used to express the action of querying different datasets and aggregate the retrieved data.

**read** – the act of obtaining data from the asset (e.g. via SPARQL query).

**copy** - the action of copying data is fundamental when defining copyright restrictions and necessary for certain license definitions.

**write** - can be used to represent SPARQL INSERT queries, since it specifies the action of writing to an asset.

**...**

# ODRL Concepts 5/5
## Constraint

### Constraint

**Constraints** offer the possibility to restrict and limit the scope of **Permissions**, **Prohibitions** and **Duties**, using a simple mathematical structure with two operands and one operator:

*The number of query requests must be less or equal than 100 per day.*

operand — operator — operand

Additionally a specific **Purpose** of the constrained **Action** can be defined:

*Reading from a dataset using ASK queries shall be restricted to 100 executions.*

action — purpose — constraint

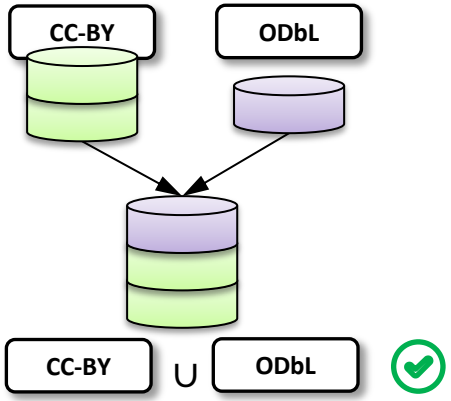
# Machine-readable License - Example:
## CC-BY-SA 3.0

```
<http://purl.org/NET/rdflicense/cc-by-sa3.0>
   a odrl:Policy ;
   rdfs:label "Creative Commons CC-BY-SA" ;
   dct:source <http://creativecommons.org/licenses/by-sa/3.0/> ;
   dct:hasVersion "3.0" ;
   dct:language <http://www.lexvo.org/page/iso639-3/eng> ;
   dct:publisher "Creative Commons" ;

   odrl:permission [
      odrl:action cc:Distribution, cc:Reproduction, cc:DerivativeWorks;
      odrl:duty [
         odrl:action cc:Notice, cc:ShareAlike, cc:Attribution
      ]
   ] .
```

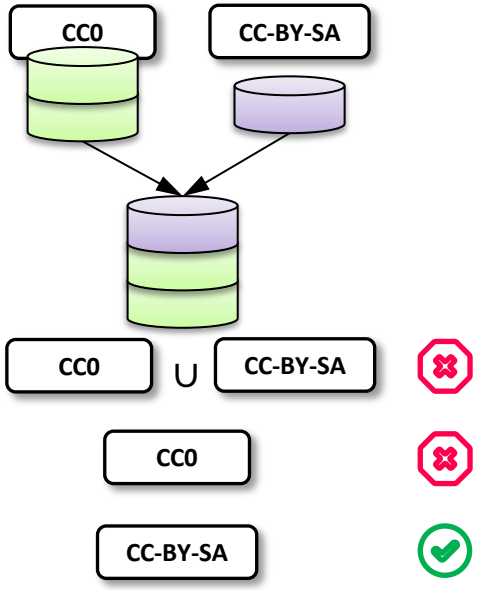
# Reasoning about Licences

# Example Reasoning Task: Policy compatibility

## Conflict Detection

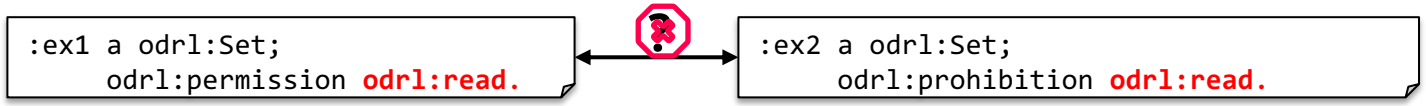



## Conflict Resolution

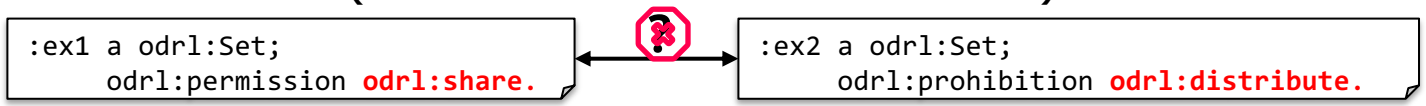


# Dependencies between ODRL Actions

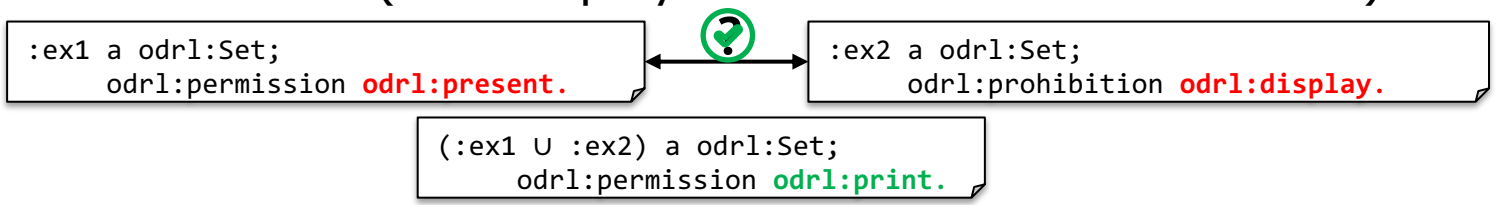
- Policies govern execution of actions over assets.
- Does the **permission** of one action conflict with the **prohibition** of another action?
- **Direct Conflict:**




- **Implicit Conflict** (odrl:share → odrl:distribute)




- **Partial Conflict** (odrl:display odrl:narrowerThan odrl:use)



# Conflict:

```
:ex1 a odrl:Set;
     odrl:permission odrl:read.
```

```
:ex2 a odrl:Set;
     odrl:prohibition odrl:read.
```

- How to deal with conflicting evaluation results?

```
@prefix odrl: <http://w3.org/ns/odrl/2/> .
@prefix : <http://www.example.com/> .

:policy1 a odrl:Agreement ;
              odrl:permission [
                  a odrl:Permission;
                  odrl:assigner :owner;
                  odrl:assignee :alice;
                  odrl:action odrl:read;
                  odrl:target :dataset1;
```

```
@prefix odrl: <http://w3.org/ns/odrl/2/> .
@prefix : <http://www.example.com/> .

:policy2 a odrl:Agreement ;
              odrl:prohibition [
                  a odrl:Prohibition;
                  odrl:assigner :owner;
                  odrl:assignee :alice;
                  odrl:action odrl:read;
                  odrl:target :dataset1;
```

- **ODRL defines three different conflict resolution strategies**
  - perm … Permmission overrides,
  - Prohibit … Prohibition overrides
  - Invalid … inconsistent

# Implicit Dependencies between Actions

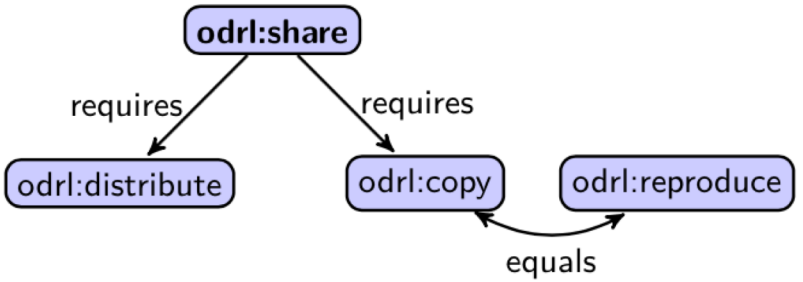
```
:ex1 a odrl:Set;
     odrl:permission odrl:share.
```

```
:ex2 a odrl:Set;
     odrl:prohibition odrl:distribute.
```

- Other dependencies are only implicitly expressed as part of the natural language description of ODRL actions.

- e.g. **odrl:share**
  - Prohibition of either **odrl:reproduce/odrl:copy** or **odrl:distribute** would cause a conflict, if **odrl:share** would be pe

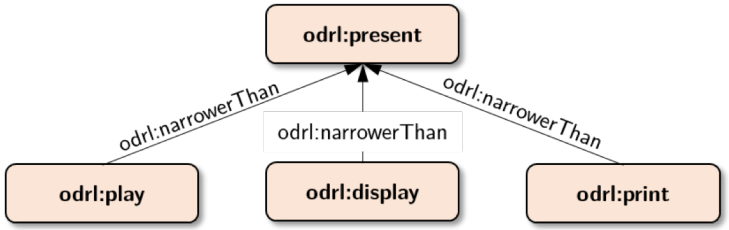
**odrl:share**: *The act of the non-commercial **reproduction** and **distribution** of the asset to third-parties.*



# Vertical Hierarchy of Actions

```
:ex1 a odrl:Set;
     odrl:permission odrl:present.
```

```
:ex2 a odrl:Set;
     odrl:prohibition odrl:display.
```




$$\frac{\rho \text{ odrl:action } \beta \qquad \alpha \text{ odrl:narrowerThan } \beta}{\rho \text{ odrl:action } \alpha}$$

**Raw**

```
<http://example.com/policy:01a>
    a odrl:Policy;
    odrl:permission [
        a odrl:Permission ;
        odrl:target ex:PartA ;
        odrl:action odrl:present ;
        odrl:assignee ex:Bob
    ] ;
    odrl:prohibition [
        a odrl:Prohibition ;
        odrl:target ex:PartB ;
        odrl:action odrl:print ;
        odrl:assignee ex:Bob
    ] .
```

**Inferred**

```
<http://example.com/policy:01b>
    a odrl:Policy;
    odrl:permission [
        a odrl:Permission ;
        odrl:target ex:PartA ;
        odrl:action odrl:present ;
        odrl:action odrl:play ;
        odrl:action odrl:display;
        odrl:action odrl:print ;
        odrl:assignee ex:Bob
    ] ;
    odrl:prohibition [
        a odrl:Prohibition ;
        odrl:target ex:PartB ;
        odrl:action odrl:print ;
        odrl:assignee ex:Bob
    ] .
```

# Details:

- For a formalization of conflict resolution strategies and reasoning:

**Towards Formal Semantics for ODRL Policies***

Simon Steyskal[1,2] and Axel Polleres[1]

[1] Vienna University of Economics and Business, Austria
[firstname.lastname]@wu.ac.at
[2] Siemens AG, Vienna, Austria
[firstname.lastname]@siemens.com

**Abstract.** Most policy-based access control frameworks explicitly model whether execution of certain actions (read, write, etc.) on certain assets should be permitted or denied and usually assume that such actions are disjoint from each other, i.e. there does not exist any explicit or implicit dependency between actions of the domain. This in turn means, that conflicts among rules or policies can only occur if those contradictory rules or policies constrain the same action. In the present paper - motivated by the example of ODRL 2.1 as policy expression language - we follow a different approach and shed light on possible dependencies among actions of access control policies. We propose an interpretation of the formal semantics of general ODRL policy expressions and motivate rule-based reasoning over such policy expressions taking both explicit and implicit dependencies among actions into account. Our main contributions are (i) an exploration of different kinds of ambiguities that might emerge based on explicit or implicit dependencies among actions, and (ii) a formal interpretation of the semantics of general ODRL policies based on a defined abstract syntax for ODRL which shall eventually enable to perform rule-based reasoning over a set of such policies

Simon Steyskal and Axel Polleres. Towards formal semantics for ODRL policies. In *9th International Web Rule Symposium (RuleML2015)*, August 2015. Springer.

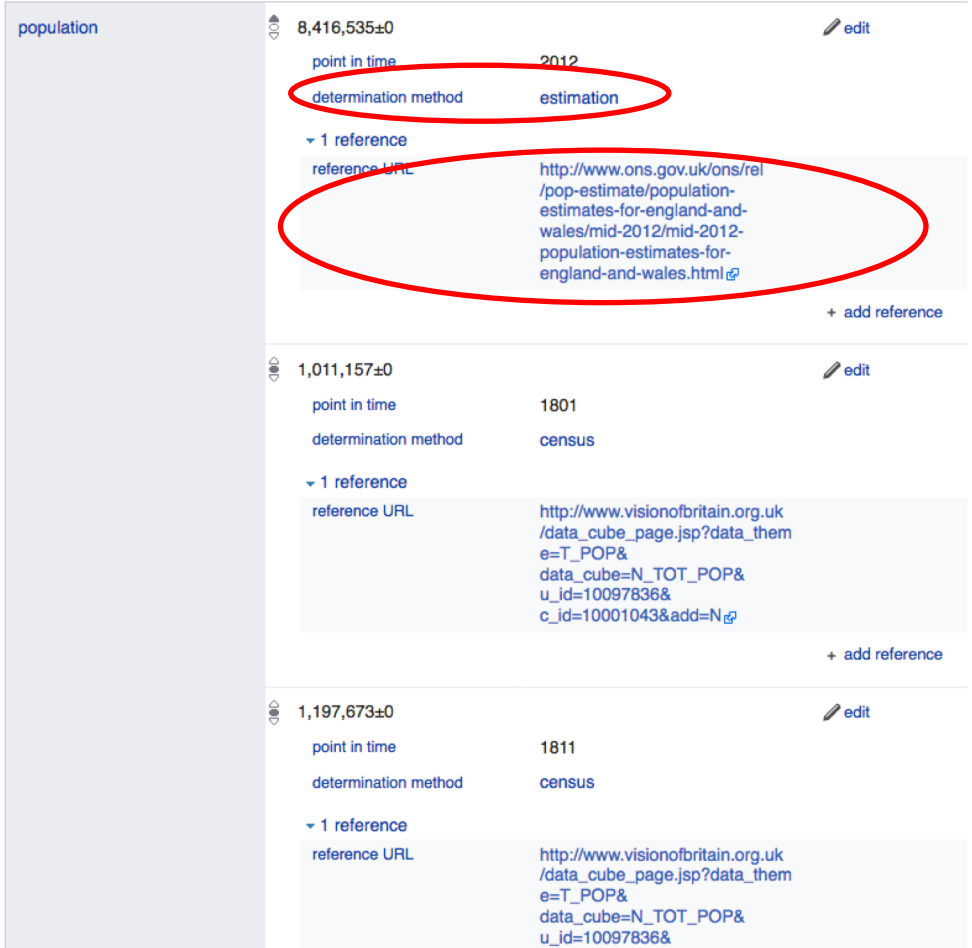
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

# Provenance

*www.provbook.org/tutorial/provenanceweek2014/*

EFMD EQUIS ACCREDITED

# Motivation: Where does the data on wikidata come from? How was it generated?



# Provenance Definition

- Provenance is a record that describes the people, institutions, entities, and activities, involved in producing, influencing, or delivering a piece of data or a thing in the world

- Provenance is crucial in deciding:
  - where information is coming from
  - whether information is to be trusted,
  - how it should be integrated with other sources, and
  - how to give credit to its originators when reusing it.

- Provenance can help users to make trust judgments.

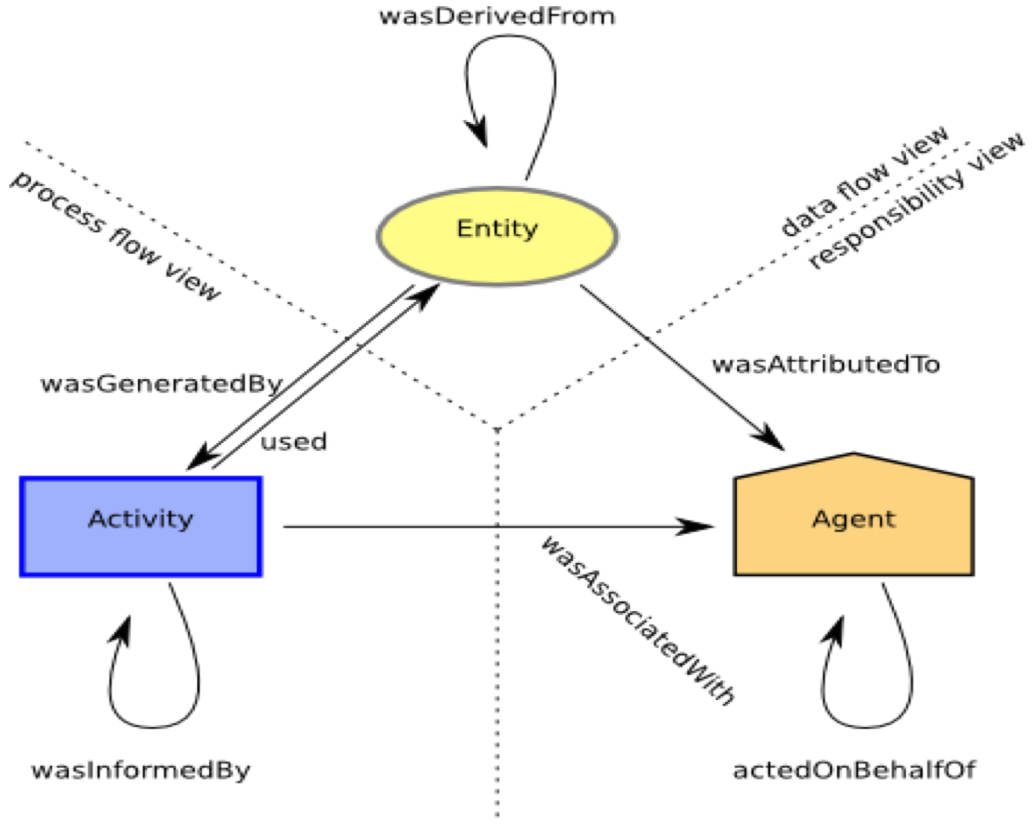
# “Provenance” is not a new subject

- There has been lot of work around provenance in:
  - workflow systems
  - Databases (“lineage”)
  - knowledge representation
  - information retrieval
- There are communities and vocabularies out there
  - Open Provenance Model (OPM)
  - Dublin Core
  - Provenir ontology
  - Provenance vocabulary
  - SWAN provenance ontology
  - etc.

# The fundamental notions of PROV

- Entity
  - the “things” whose provenance we want to describe
- Activity
  - describes how entities are created, changed.
- Agent
  - are responsible for the activities.
- Usage, generation, derivation, attribution,..
  - connections describing how entities, agents, and activities interact

# Three Views of Provenance



# PROV example:

```
ex:observation123 a prov:Entity ;
        prov:generatedAtTime "2017-01-01T01:01:01"^^xsd:dateTime;
        prov:wasGeneratedBy ex:activity456 ;
        prov:wasDerivedFrom ex:obs789 .
ex:activity456 a prov:Activity;
        prov:qualifiedAssociation [ a Association ;
        prov:wasAssociatedWith ex:fred ;
        prov:hadPlan ex:rule397 . ] .
```

- Goal: standardized vocabulary to track how data was created, add references, for which purpose, etc.
- Not yet widely used e.g. on OD portals. ☹
- Neither completely implemented in Wikidata (own vocabulary, only partially uses PROV…)

# PROV Formal Semantics?

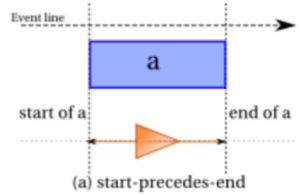
http://www.w3.org/TR/prov-constraints/

## Semantics of the PROV Data Model

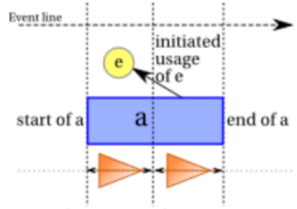
http://www.w3.org/TR/prov-constraints/

## Constraints of the PROV Data Model

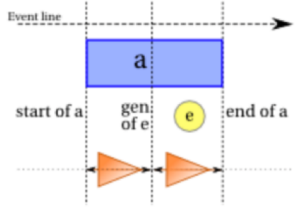
- More or less only "conformance/validity" of PROV descriptions…
- i.e. practical reasoning tasks over PROV descriptions is – AFAIK – a relatively open space!



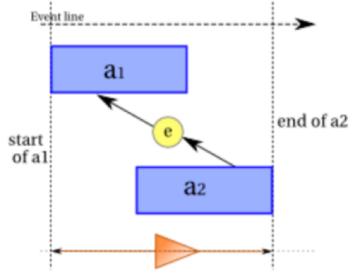
(a) start-precedes-end



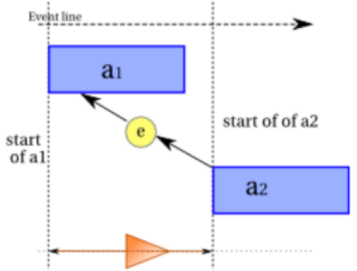
(b) usage-within-activity



(c) generation-within-activity



(d) wasInformedBy-ordering



(e) wasStartedBy-ordering

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

# Metadata Quality Issues and Vocabularies in Open Data Portals

EFMD EQUIS ACCREDITED

# Open Data Portal Watch http://data.wu.ac.at/portalwatch

Quality Assessment and Monitoring of 259 Open Data Portals





## 259 Portals

Filter ... (by URI, Software, ISO)

### data.gov.hk

API Homepage CKAN

53 Snapshots Last Jun 19 - Jun 25, 2017

447 DATASETS 7245 RESOURCES

Dashboard

### data.gov.sk

API Homepage CKAN

52 Snapshots Last Jun 19 - Jun 25, 2017

1136 DATASETS 5256 RESOURCES

Dashboard

# Open Data Portal Watch Project

***Monitoring and QA over evolving data portals (weekly snapshots)***

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

**3/2015 [1]:**
- 90 portals
- Only **CKAN**

**8/2015 [2]:**
- 6 **quality metrics**
- QA

**6/2016 [3]:**
- 260 portals
- **Socrata**, **OpenDataSoft**
- 18 metrics

| | total | CKAN | Socrata | ODSoft | DCAT |
|---|---|---|---|---|---|
| **portals** | 261 | 149 | 99 | 11 | 2 |
| **datasets** | 854,013 | 767,364 | 81,268 | 3,340 | 2,041 |
| **URLs** | 2,057,924 | 1,964,971 | 104,298 | 12,398 | 6,092 |


[1] Towards assessing the quality evolution of open data portals. In ODQ2015: Open Data Quality Workshop, Munich, Germany

[2] Quality assessment & evolution of open data portals. In: International Conference on Open and Big Data, Rome, Italy (2015)

[3] Automated quality assessment of metadata across open data portals. ACM Journal of Data and Information Quality (2016)


EFMD EQUIS ACCREDITED

# Quality Assessment

*5 Quality Dimensions*:

- Existence
- Conformance
- Retrievability
- Accuracy
- Open Data

# Open Data & Conformance dimension

| | | | |
|---|---|---|---|
| **OPEN DATA** | | | |
| *Is the specified format and license information suitable to classify a dataset as open?* | | | |
| OpenFormat | Is the file format based on an open standard? | | `dct:format`<br>`dcat:mediaType` |
| MachineRead | Can the file format be considered as machine readable? | | `dct:format` |
| OpenLicense | Is the used license conform to the open definition? | | `dct:license` |
| **CONFORMANCE** | | | |
| *Does information adhere to a certain format if it exist?* | | | |
| AccessURL* | Are the values of access properties valid HTTP URLs? | | `dcat:accessURL`<br>`dcat:downloadURL` |
| ContactEmail* | Are the values of contact properties valid emails? | `dcat:contactPoint`<br>`dct:publisher` | |
| ContactURL* | Are the values of contact properties valid HTTP URLs? | `dcat:contactPoint`<br>`dct:publisher` | |
| DateFormat | Is date information specified in a valid date format? | `dct:issued`<br>`dcat:modified` | `dct:issued`<br>`dcat:modified` |
| License | Can the license be mapped to the list of licenses reviewed by opendefinition.org? | | `dct:license` |
| FileFormat | Is the specified file format or media type registered by IANA? | | `dct:format`<br>`dcat:mediaType` |

# Existence & Accuracy dimension

| | | | |
|---|---|---|---|
| **EXISTENCE** | | | |
| *Existence of important information (i.e. exist certain metadata keys)* | | | |
| Access* | Is there access information for resources provided? | | `dcat:accessURL`<br>`dcat:downloadURL` |
| Discovery | Is information available that can help to discover/search datasets? | `dct:title`<br>`dct:description`<br>`dcat:keyword` | |
| Contact* | Existence of information that would allow to contact the dataset provider. | `dcat:contactPoint`<br>`dct:publisher` | |
| Rights | Existence of information about the license of the dataset or resource. | | `dct:license` |
| **ACCURACY** | | | |
| *Does information accurately describe the underlying resources?* | | | |
| FormatAccr | Is the specified file format accurate? | | `dct:format`<br>`dcat:mediaType` |
| SizeAccr | Is the specified file size accurate? | | `dcat:byteSize` |

# Data Quality, example data.gov.uk:

http://data.wu.ac.at/portalwatch/portal/data_gov_uk/1725/quality

**Quality Assessment over the DCAT representation**

**Existence Metrics**

| DATE | RIGHTS | PRESERVATION |
|---|---|---|
| 0.72 | 0.76 | 0.26 |
| 43165 DATASETS ANALYSED | 43165 DATASETS ANALYSED | 43165 DATASETS ANALYSED |

| ACCESS | DISCOVERY | TEMPORAL |
|---|---|---|
| 0.89 | 0.76 | 0.0 |
| 43165 DATASETS ANALYSED | 43165 DATASETS ANALYSED | 43165 DATASETS ANALYSED |

| SPATIAL | CONTACT |
|---|---|
| 0.57 | 1.0 |
| 43165 DATASETS ANALYSED | 43165 DATASETS ANALYSED |

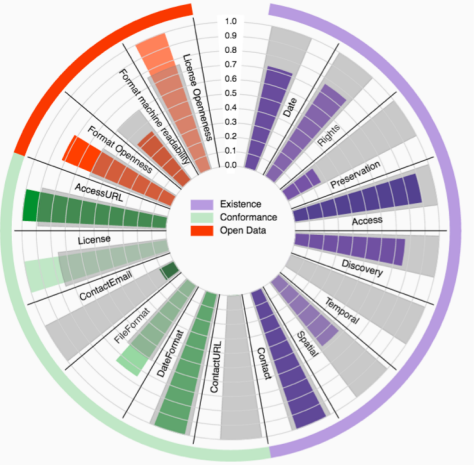
**Conformance Metrics**

| CONTACTURL | DATEFORMAT | FILEFORMAT |
|---|---|---|
| 0.0 | 1.0 | 0.75 |
| 43165 DATASETS ANALYSED | 43165 DATASETS ANALYSED | 26103 DATASETS ANALYSED |

| CONTACTEMAIL | LICENSE | ACCESSURL |
|---|---|---|
| 0.11 | 1.0 | 1.0 |
| 43165 DATASETS ANALYSED | 32750 DATASETS ANALYSED | 38475 DATASETS ANALYSED |

**Open Data Metrics**

| FORMAT OPENNESS | FORMAT MACHINE READABILITY | LICENSE OPENNENESS |
|---|---|---|
| 0.82 | 0.43 | 1.0 |
| 26080 DATASETS ANALYSED | 26080 DATASETS ANALYSED | 32750 DATASETS ANALYSED |



# CKAN (mostly) exports DCAT

- CKAN provides **extension** to export DCAT
  - Mapping of *datasets* and *resources* to `dcat:Dataset` and `dcat:Distribution`
  - Recent version of extension supports DCAT-AP

- 93 of 133 CKAN portals provide DCAT export

# CKAN provides “extra” keys

- CKAN can include additional metadata keys
  - Added by portal provider, or other CKAN extension

ADDITIONAL INFORMATION

| | |
|---|---|
| **Added to data.gov.uk** | 04/04/2011 |
| **Theme** | Environment |
| **Temporal coverage** | 1990 - 2006 |
| **Geographic coverage** | England, Wales |
| **Precision** | nearest per cent |
| **Update frequency** | annual |
| **Temporal granularity** | year |

Example from: https://data.gov.uk/dataset/river-water-quality-regions

# What “extra” keys are available?

- 3607 different extra keys in 514k datasets
- Extra keys in multiple portals:

| Portals | 1 | 2 | 3 – 9 | 10 – 19 | ≥ 20 |
|---|---|---|---|---|---|
| Extra keys | 2269 | 1131 | 172 | 30 | 5 |

- Most frequent extra keys:

| Extra key | Datasets | Portals | Origin | DCAT key |
|---|---|---|---|---|
| harvest_object_id | 268241 | 30 | Harvesting extension | — |
| spatial | 245211 | 33 | Spatial extension | dct:spatial |
| harvest_source_id | 243136 | 29 | Harvesting extension | — |
| harvest_source_title | 243043 | 29 | Harvesting extension | — |
| guid | 150811 | 20 | Spatial extension | — |
| resource-type | 148843 | 15 | Spatial extension | — |
| contact-email | 148671 | 17 | Spatial extension | dcat:contactPoint |
| metadata-date | 141758 | 15 | Spatial extension | dct:issued/modified |

# Current mapping of “extra” keys

3 different cases, depending on version and configuration of CKAN-to-DCAT extension:

- ***Portal-specific mapping***:
  Portal defines mapping for metadata key to property, e.g.:

  ```
  "temporal_coverage" → dct:temporal
  ```

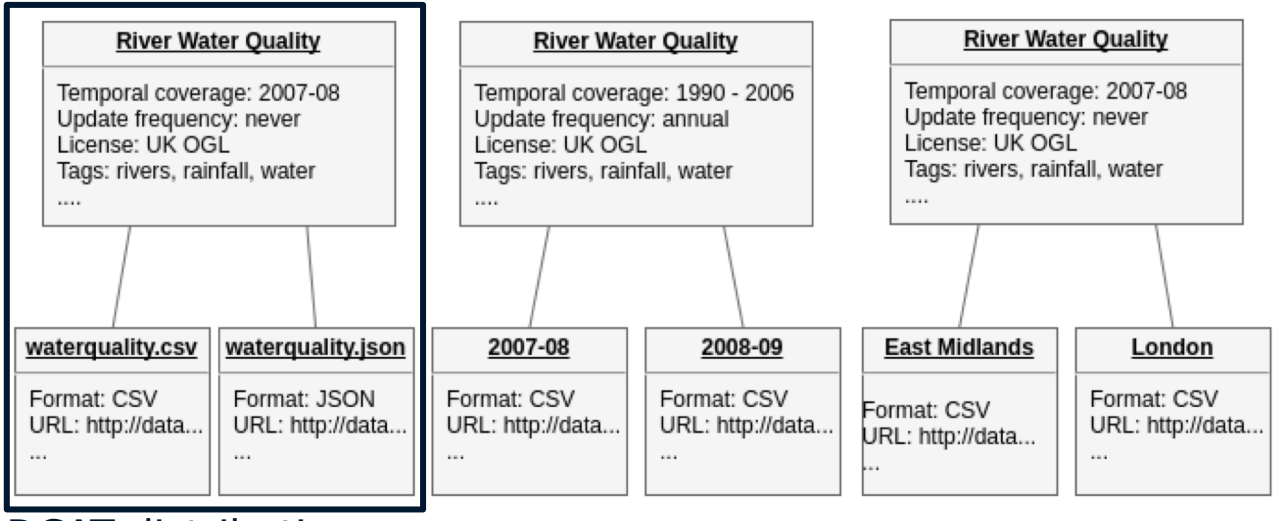
- ***Generic mapping by extension***:
  Pattern for exporting all available extra keys, e.g.:

  ```
  dc:relation [
      rdfs:label "geographic_coverage" ;
      rdf:value "101000: England, Wales"
    ]
  ```

- ***No mapping***:
  Retrieved DCAT description returns no mapping for extra keys

# How to model CKAN resources?

- DCAT distribution:
  - *dataset might be available in different forms, these forms might represent different formats or endpoints*
- Use of resources in CKAN:



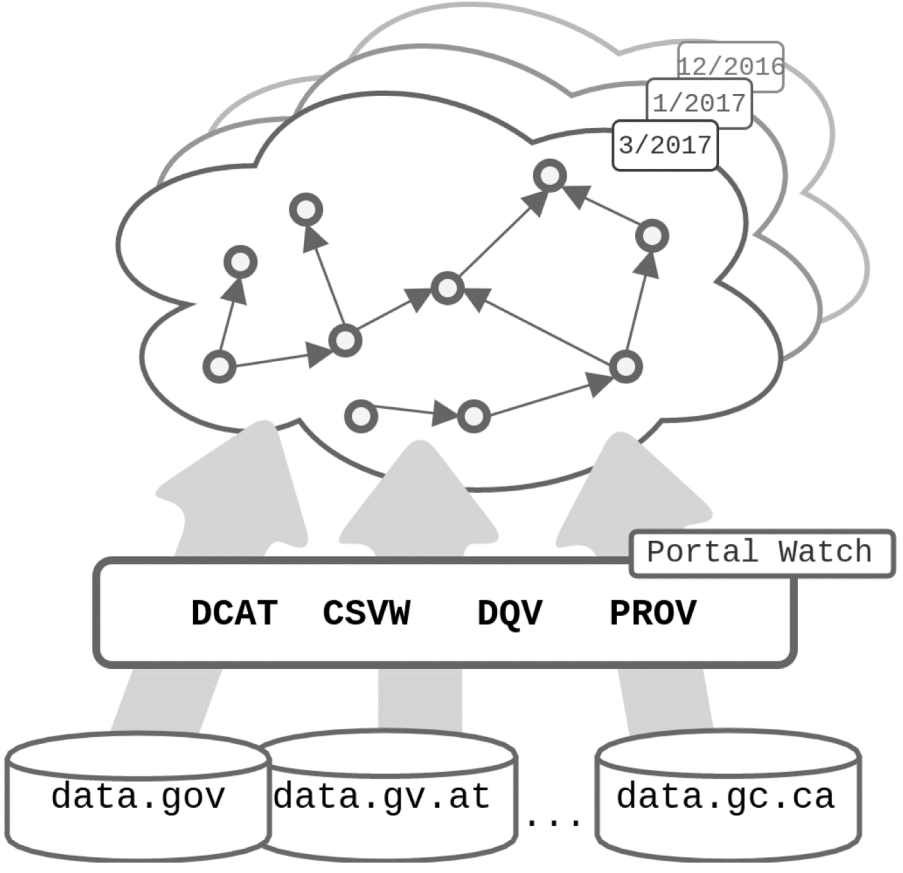

DCAT distribution

Example from: https://data.gov.uk/dataset/river-water-quality-regions

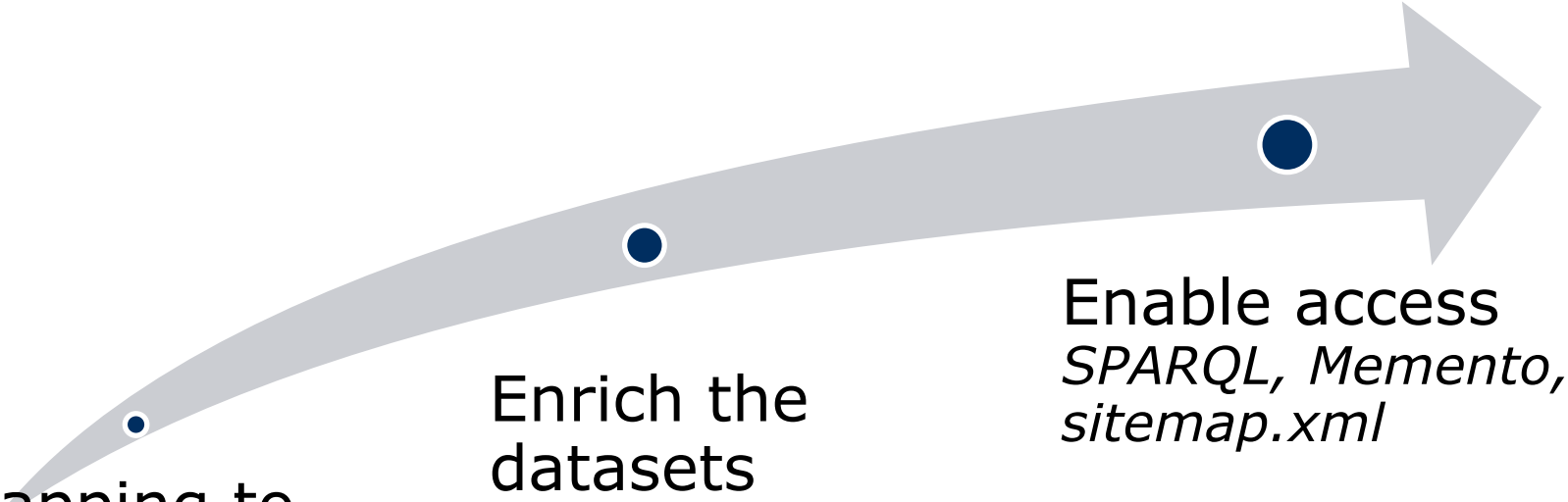
# Identified challenges

- Metadata is **heterogeneous** and (partially) messy
  - Software-specific metadata (CKAN vs Socrata vs …)
  - Portal-specific metadata
  - Missing metadata (file formats, API descriptions, …)
- Metadata not available as Linked Data
  - Only partially in DCAT vocabulary
  - **No mappings** for additional metadata fields
- Poor **discoverability** of datasets
  - No content information in metadata (e.g., CSV headers)
  - Datasets' metadata not optimized for search engines

# Lifting Data Portals to the Web of Data



# Approach



Mapping to standard vocabularies
*DCAT, Schema.org*

Enrich the datasets
*DQV, CSVW, PROV*

Enable access
*SPARQL, Memento, sitemap.xml*

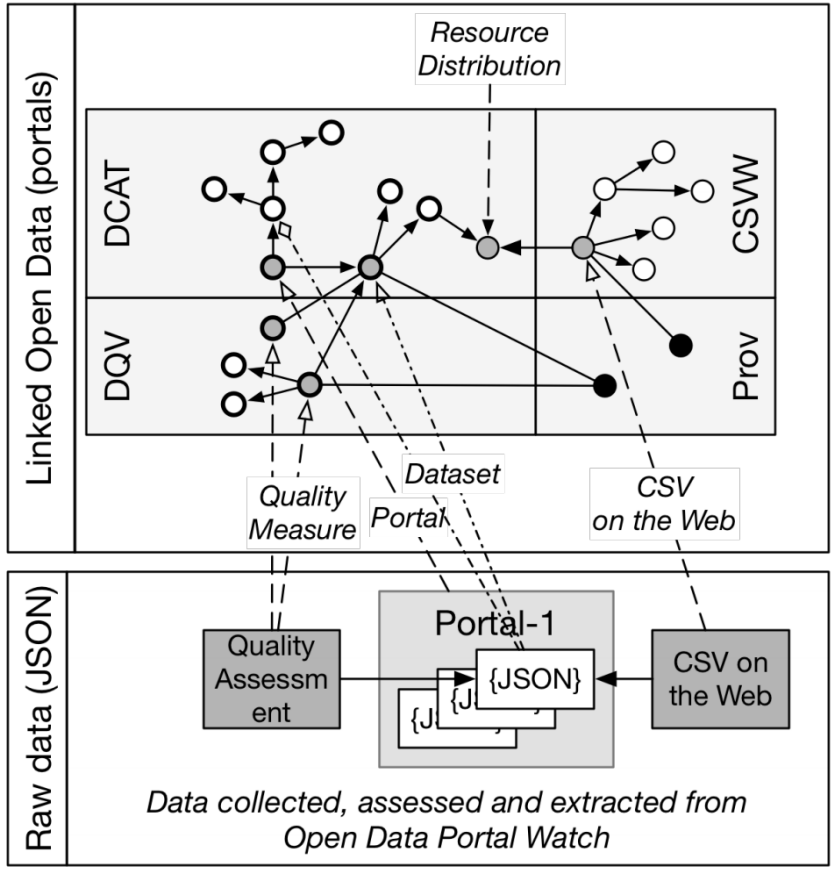
# Mapping to standard vocabularies

- **DCAT** export of metadata from CKAN, Socrata and OpenDataSoft portals
  - Mapping of most frequent (portal/domain specific) extra-metadata fields

- Mapping and publishing of **Schema.org**:
  - Mapping of DCAT to Schema.org‘s dataset vocabulary
  - Enabling integration into knowledge graphs of major search engines

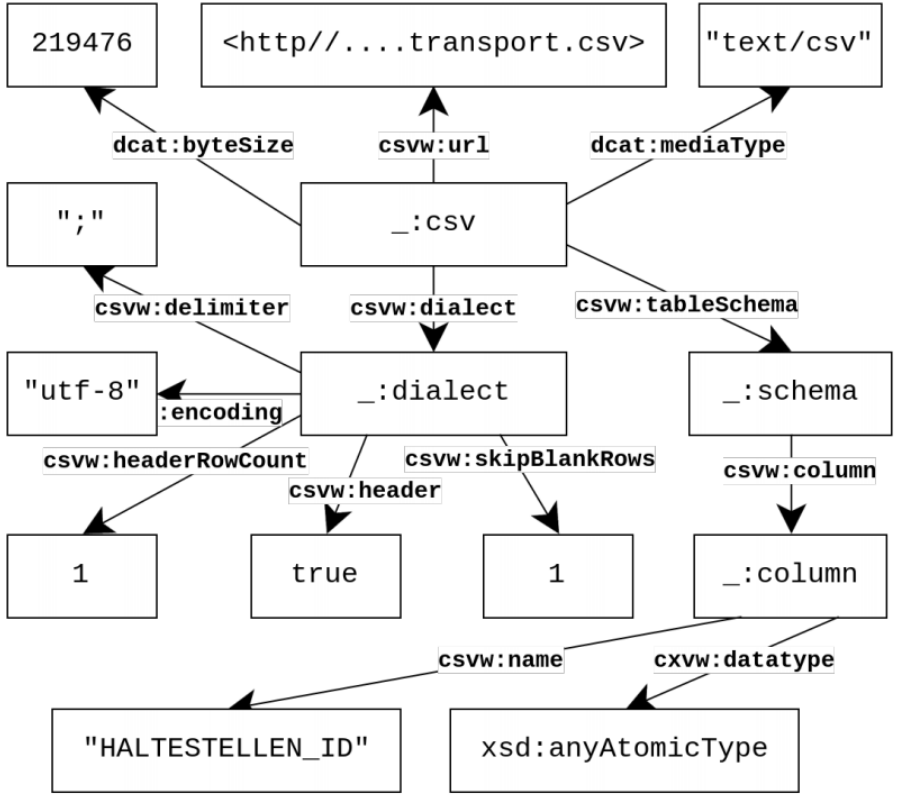
# Enrich the datasets

- *Portal Watch quality dimensions:*
  - **Data Quality vocabulary**
- *Metadata for tables:*
  - **CSV on the Web** vocabulary
- *Record provenance:*
  - **PROV ontology**



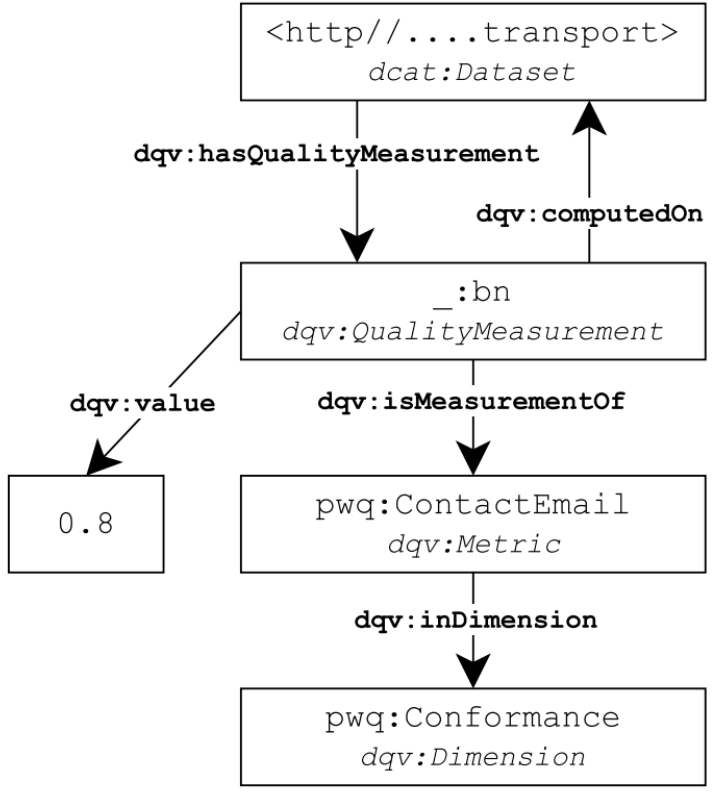
# *CSV on the Web* metadata

- Dialect properties:
  - *HTTP Content-Type*
    → `dcat:mediaType`
  - *Encoding detection*
    → `csvw:encoding`
  - *Delimiter detection*
    → `csvw:delimiter`
- Schema properties:
  - *CSV header line*
    → `csvw:name`
  - *Column type detection*
    → `xsd:datatype`



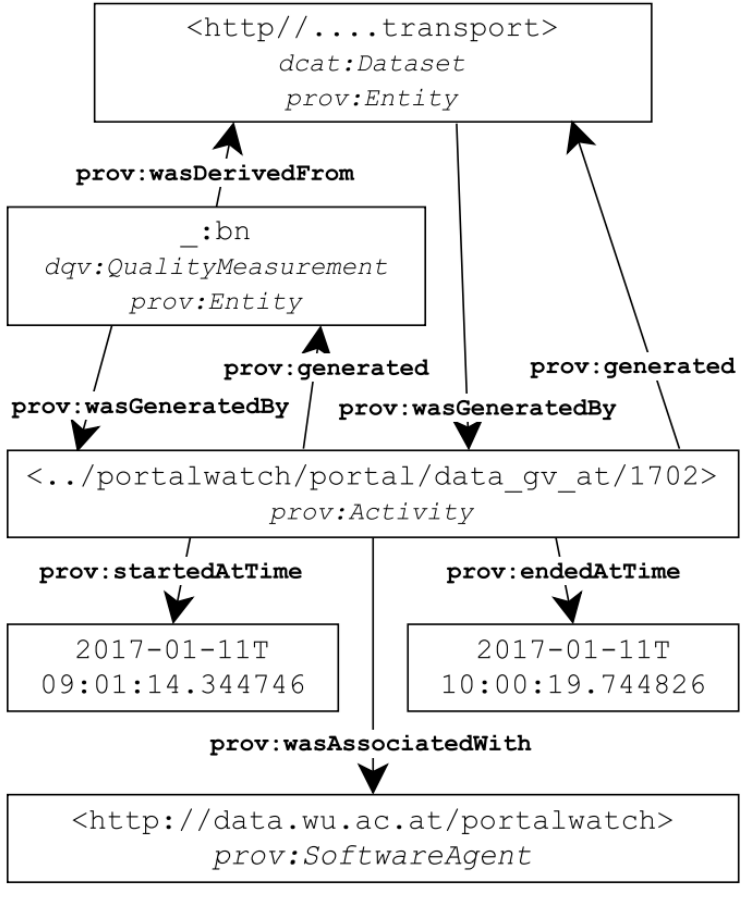
# Portal Watch quality measures

- Existence
  - Is certain metadata available
- Conformance
  - Valid emails/URLs/dates
- Open Data
  - Open format
  - Machine-readable
  - Open license

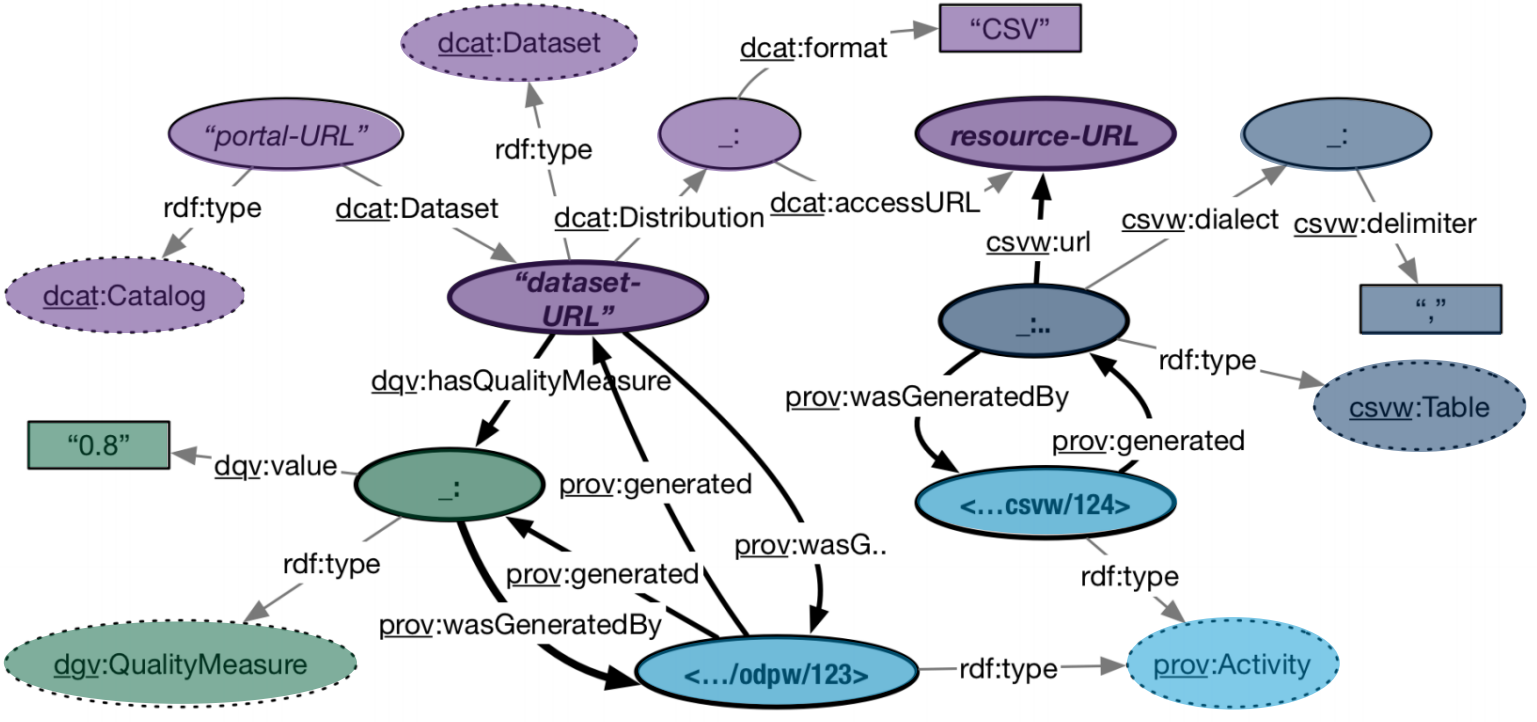


# Add provenance information

- We annotate:
  - Mapped DCAT description
  - Quality measurements
  - CSVW metadata
- PROV-Activities:
  - „fetch“-activity
  - „csvw“-activity



# The big picture



# Please use it!
# You can Access portalwatch via APIs!

- SPARQL endpoint [1]:
  - Three versions in RDF triple store (as named graphs)
  - 120 million triples each
- Archival information to provide access to weekly sneapshots via the Memento framework:
  - Datetime negotiation on “Accept-Datetime” HTTP header
  - Access to original metadata, DCAT, and DQV measures
- Schema.org via sitemap.xml [2]:
  - Publishing of all 850k datasets as HTML-embedded Schema.org

[1] http://data.wu.ac.at/portalwatch/sparql

[2] http://data.wu.ac.at/odso/

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

# Searchability and Semantic Annotation

EFMD EQUIS ACCREDITED

# Semantic Search History…

- A bit of history on Semantic Search…
  - Various research prototypes of Search Engines specifically for RDF and Linked Data…

# Semantic Search history 1/4

- Swoogle (2004-2007): http://swoogle.umbc.edu/


- *Li Ding et al., "Finding and Ranking Knowledge on the Semantic Web", Proceedings of the 4th International Semantic Web Conference, November 2005.*

# Semantic Search history 2/4

- Watson … no, not THAT Watson…


- http://watson.kmi.open.ac.uk/WatsonWUI/
- (2007 - ??? … also "frozen" since a while)

  - Mathieu d'Aquin, Enrico Motta:
    Watson, more than a Semantic Web search engine. Semantic Web 2(1): 55-63 (2011)

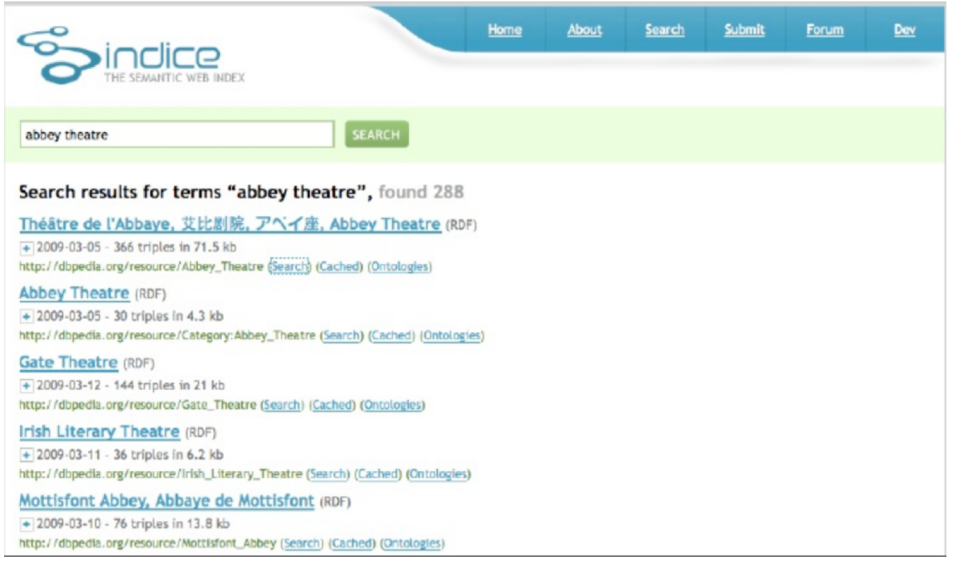
# Semantic Search history 3/4

- SWSE (2007-2011?
- VisiNav (2010-2011?)




- Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, and Stefan Decker. Searching and browsing linked data with SWSE: The semantic web search engine. *Journal of Web Semantics (JWS)*, 9(4):365--401, 2011.

# Semantic Search history 4/4




2007-2014, now offline

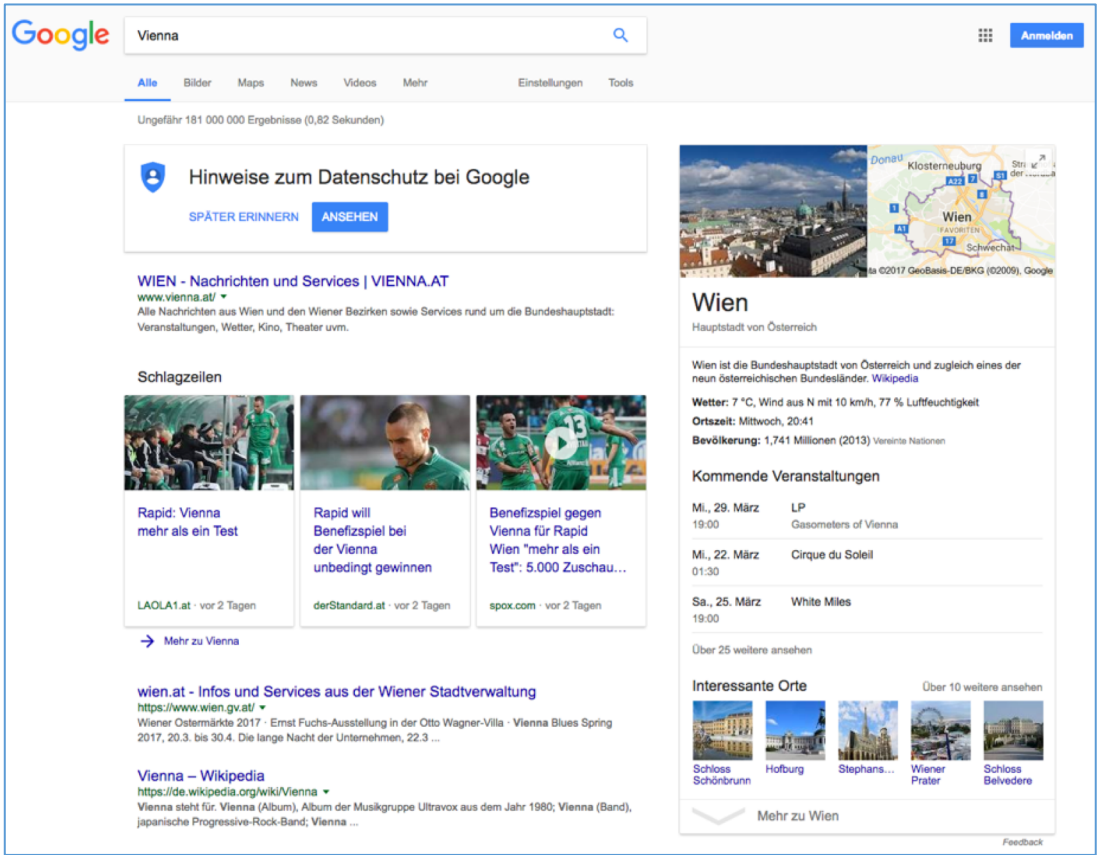
- Eyal Oren, Renaud Delbru, Michele Catasta, Richard Cyganiak, Holger Stenzhorn, Giovanni Tummarello: Sindice.com: a document-oriented lookup index for open linked data. IJMSO 3(1): 37-52 (2008)

# Semantic Search History…

- A bit of history on Semantic Search…
  - Various research prototypes of Search Engines specifically for RDF and Linked Data:
    - "entity search" through labels and "graph neighbourhoods"
    - Specific "snippet" representation of common entities and properties based on entity types
    - *BTW: Both SWSE and Sindice used (tailored, LIGHTweight)* ***RDFS*** *and* ***OWL*** *reasoning → cf. our RW2013 lecture notes for details!!!*
  - So, what happened to Semantic Search? … it has become a commodity!

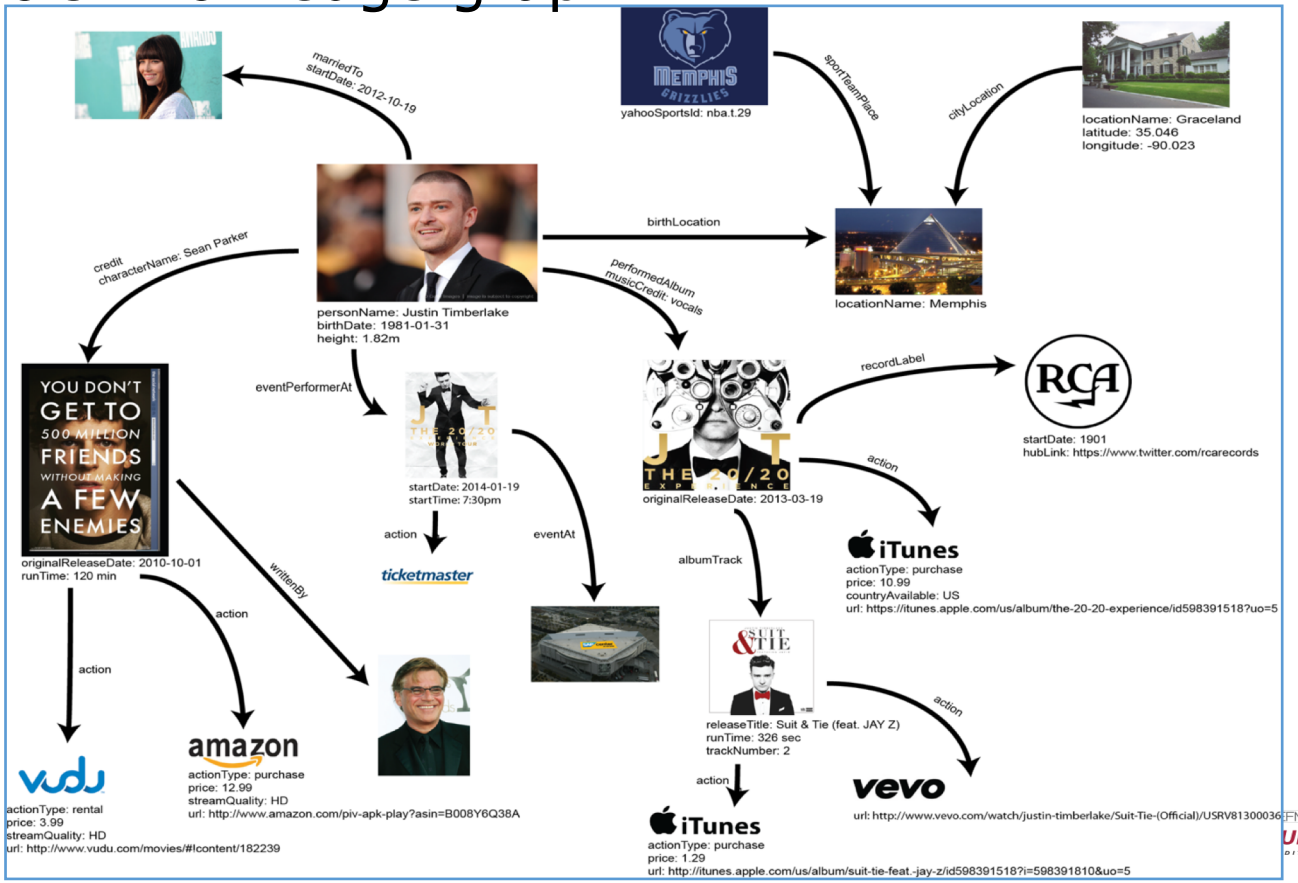
# Semantic Search History

- Semantic Search: Google's knowledge graph & rich snippets



# Semantic Search History

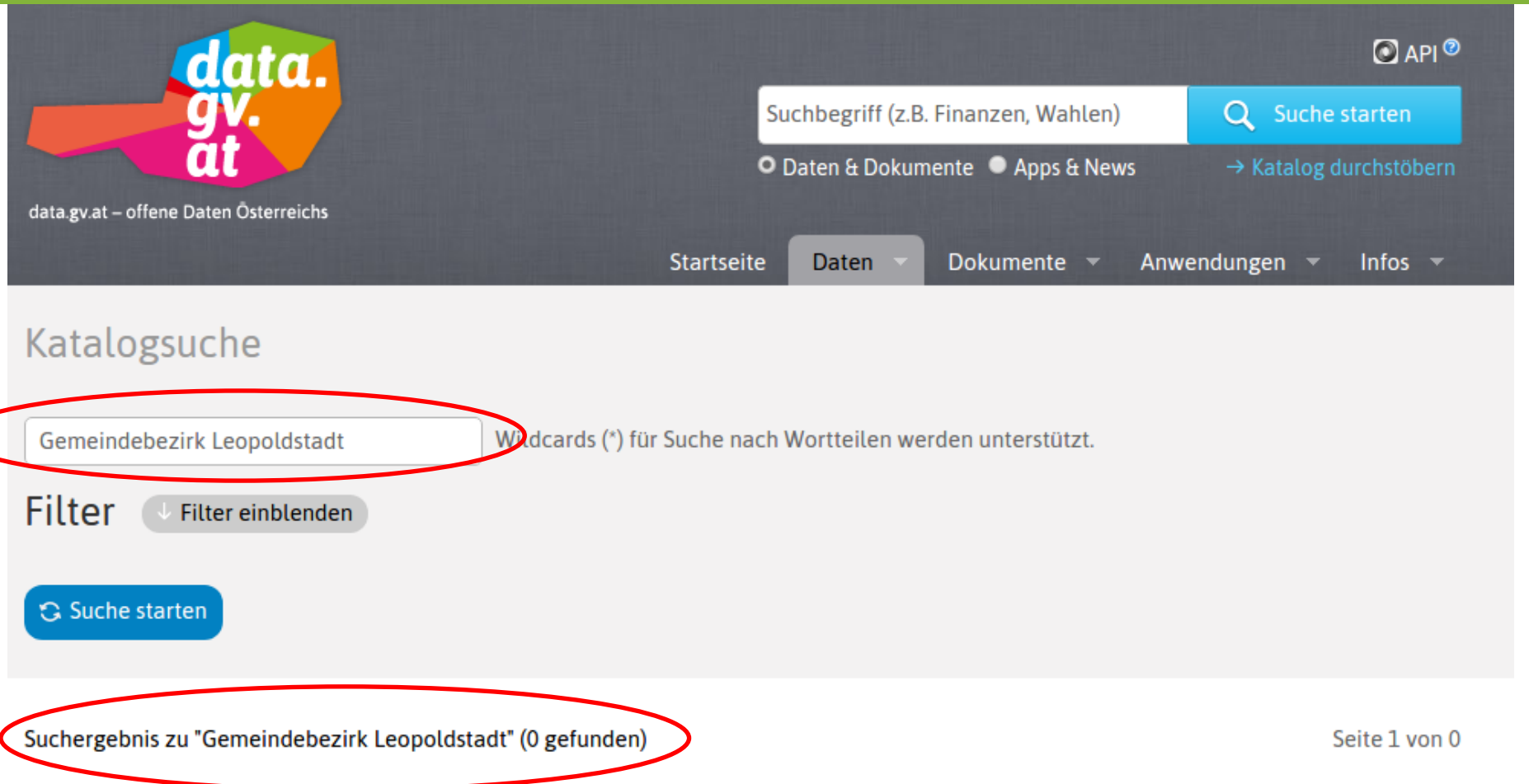
- Semantic Search: Yahoo's knowledge graph…




SOURCE: WHAT HAPPENED TO THE SEMANTIC WEB? PETER MIKA, KEYNOTE AT ACM HYPERTEXT, JULY 5, 2017
HTTPS://WWW.SLIDESHARE.NET/PMIKA/WHAT-HAPPENED-TO-THE-SEMANTIC-WEB

- So, how about search in Open Data?

# Limited Search on Open Data portals (CKAN)



# Limited Search on Open Data portals (CKAN)

- Search only in metadata descriptions
- No search in data (e.g., CSVs)

- Quick fix:
  - Full-text search index over datasets
  - Reproducing Google?

# How to achieve Semantic annotation of tabular data?

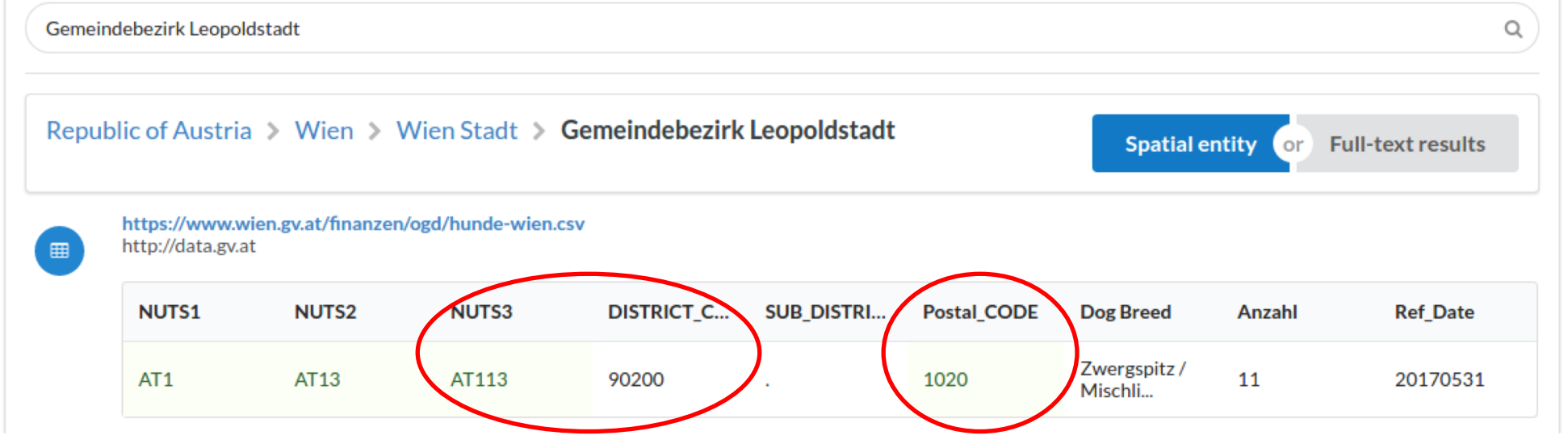
Web/HTML tables differ from typical Open Data tables:

- **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 |

# Approach 1: use „hard-wired“ heuristics
# Our own work in progress: Location Search

- There are all kind of location labels in Open Data…
  - … but no semantics ☹
    - … let’s fix it by rules/heuristics to recognize location labels!



# Approach 2: Automatically linking Open Data Tables to a Knowledge Graph?

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

*International Semantic Web conference 2016:*

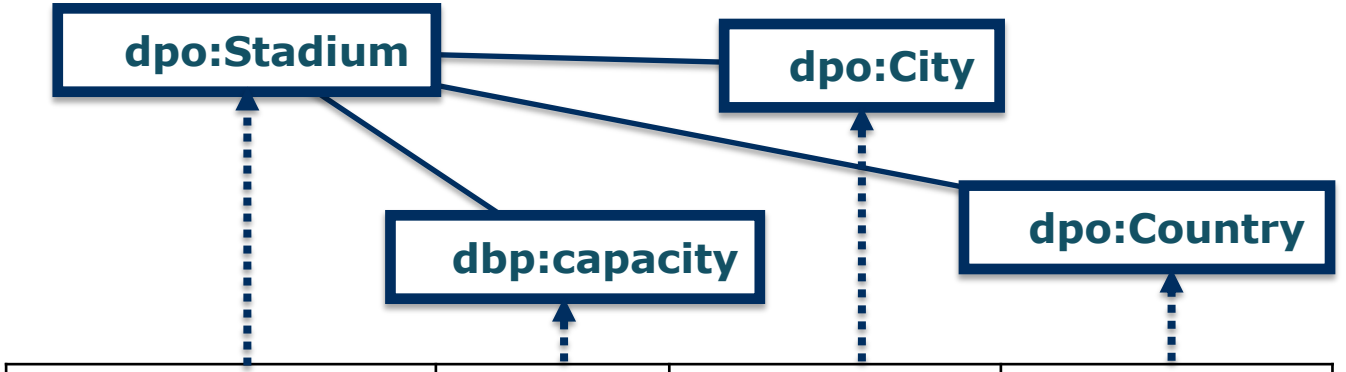
**Multi-level semantic labelling of numerical values**

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.

# Web table example




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

# Web table example

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

# Open Data tables

| | *TOTAL* | *DISTRICT_CODE* | *ISO_2* |
|---|---|---|---|
| Emirates Stadium | 60361 | N7 7AJ | GB |
| Villa Park | 42785 | B6 6HE | GB |
| Ewood Park | 31154 | BB2 4JF | GB |
| … | … | … | … |

# Use numeric values for labelling

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

| | *TOTAL* | *DISTRICT_CODE* | *ISO_2* |
|---|---|---|---|
| Emirates Stadium | 60361 | N7 7AJ | GB |
| Villa Park | 42785 | B6 6HE | GB |
| Ewood Park | 31154 | BB2 4JF | GB |
| … | … | … | … |

# Use numeric values for labelling

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

| | | | |
|---|---|---|---|
| Emirates Stadium | 60361 | N7 7AJ | GB |
| Villa Park | 42785 | B6 6HE | GB |
| Ewood Park | 31154 | BB2 4JF | GB |
| … | … | … | … |

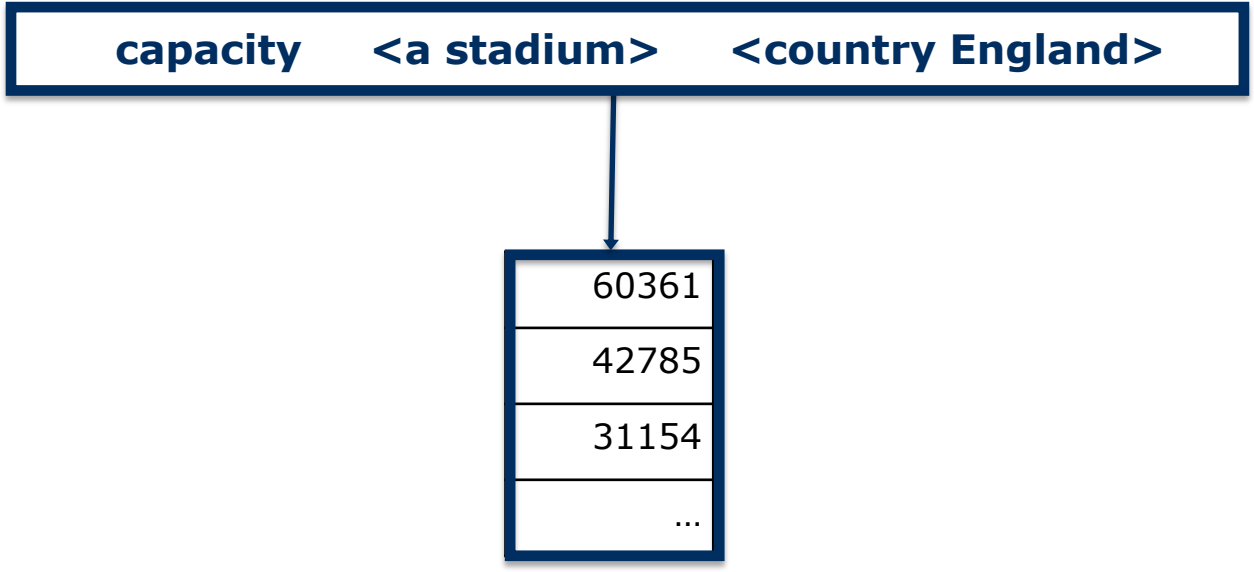
# Use numeric values for labelling

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

| |
|---|
| 60361 |
| 42785 |
| 31154 |
| … |

# Use numeric values for labelling

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

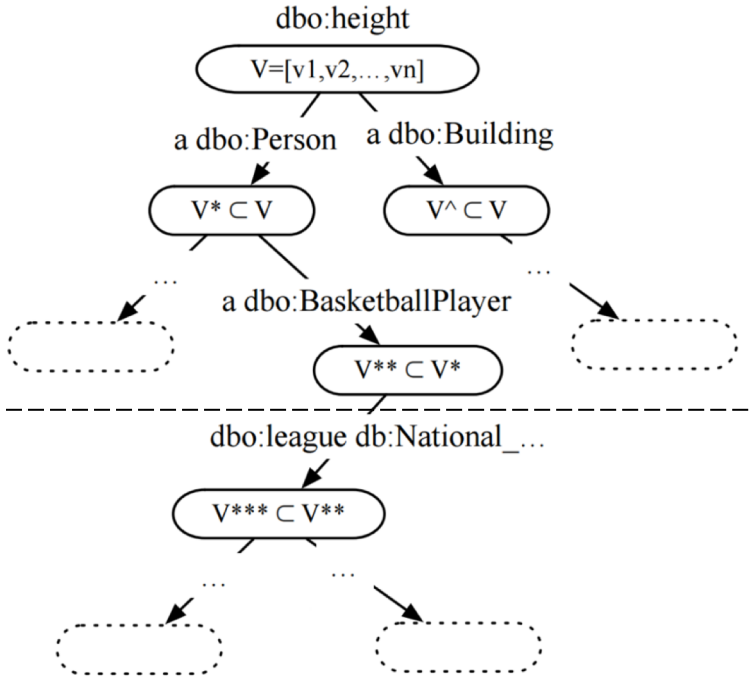


# Our Approach

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

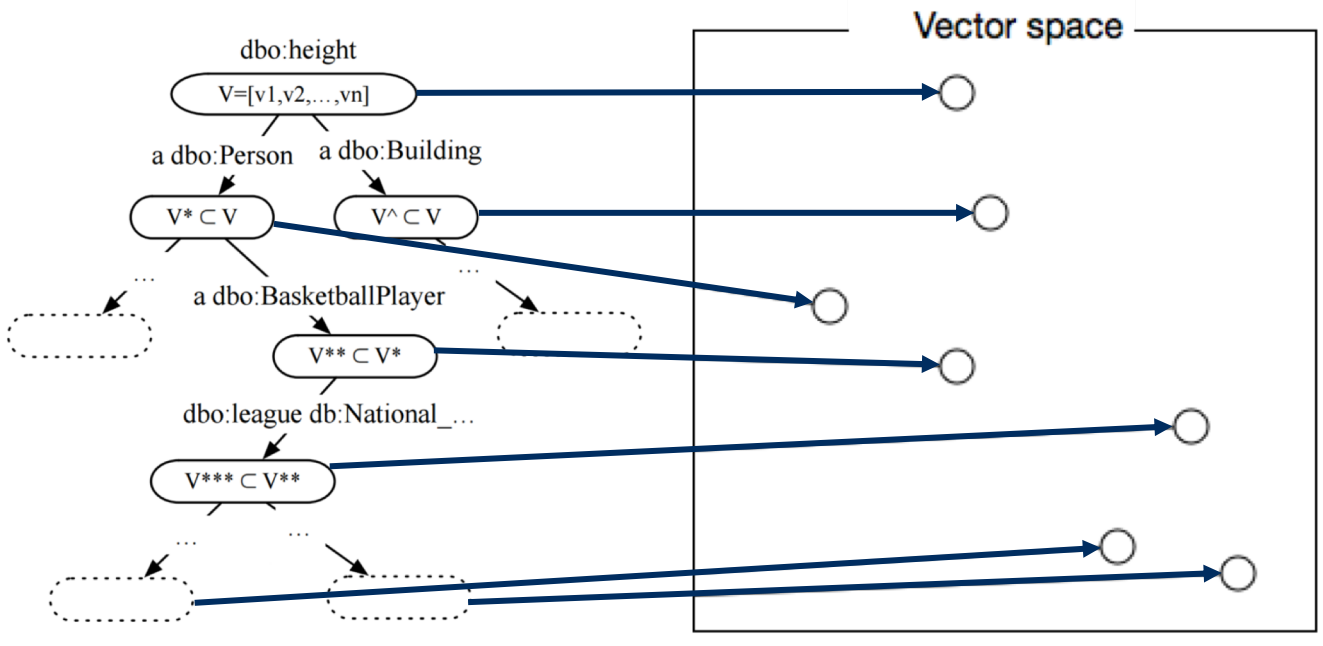
# 1. Background Knowledge Graph

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



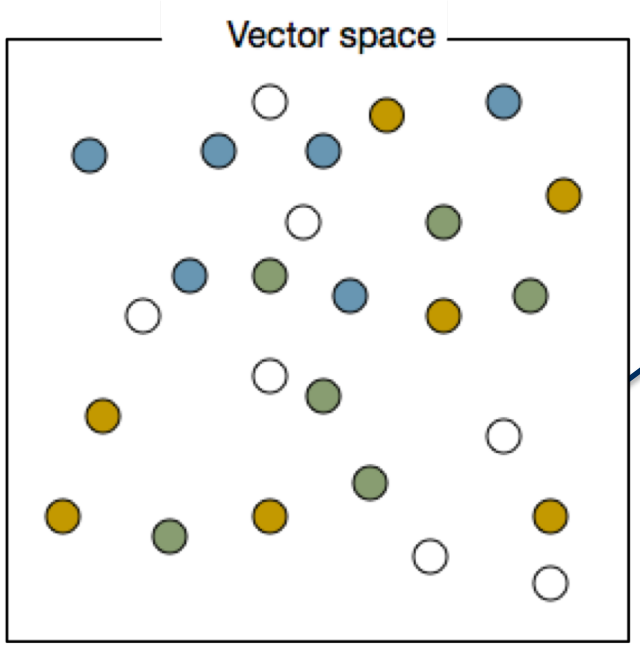
# 2. *k*-Nearest neighbor search

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



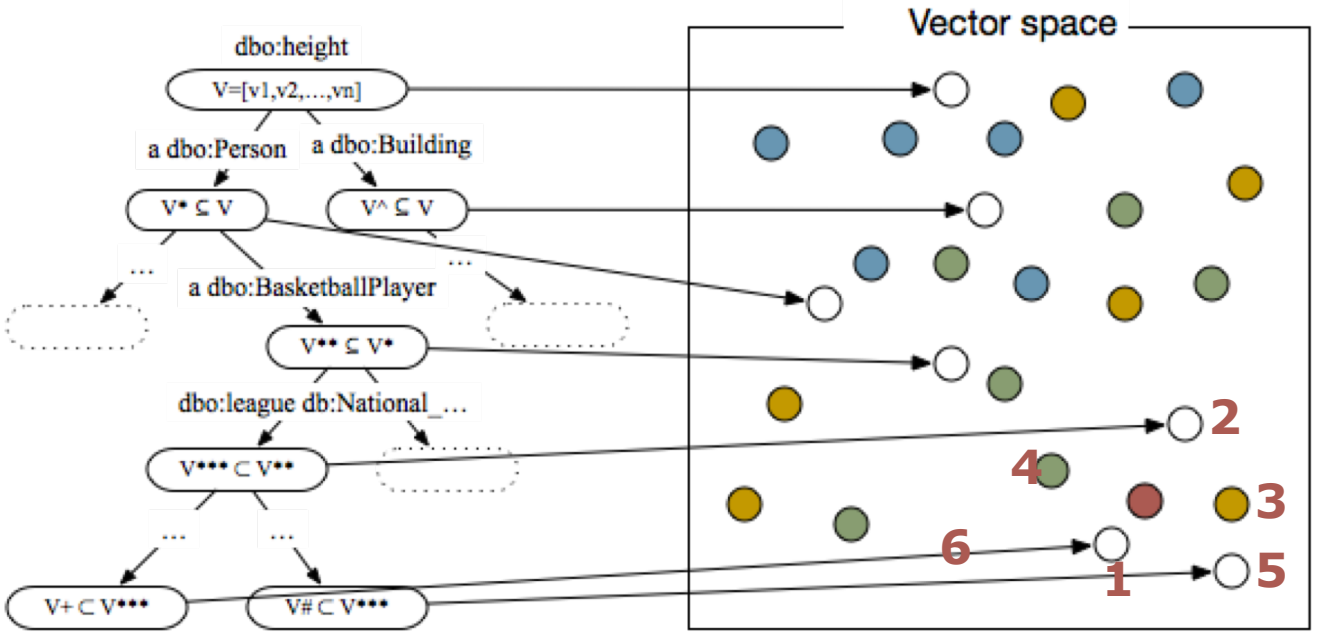
# 2. *k*-Nearest neighbor search

Compute & rank k-nearest neighbours for input values




1) input: [ 187, 201, 199, 198, 195, 199, 203, … ]
2) mapping:
3) compute distance to neighbours
4) select K nearest

# 3. Result Aggregation



# Experimental OD Column labelling

- Works well for “dbpedia-like” data
- Data from two selected Open Data portals
  - 1170 CSV tables
- Manual inspection of  top 100 tables

- **Lessons learned**:
  - Missing domain knowledge
  - Timeline data
  - Combine with (existing) complementary approaches

- Part I:
  - Where to find Open Data?
  - Dealing with “Low-level” data heterogeneity – Which formats are there on the Web?
  - Licenses and Provenance
  - Quality Issues in Open Data
  - How to find Open data: Search over Open Data
- **Part II:**
  - How does reasoning help? A motivating Use Case.
  - Let’s discuss how Rules & Reasoning can help? Group work!

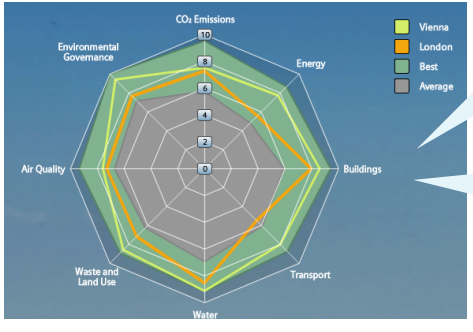
# A motivating use case…

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


… however, these are often **outdated** before

→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 (started 2012)

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

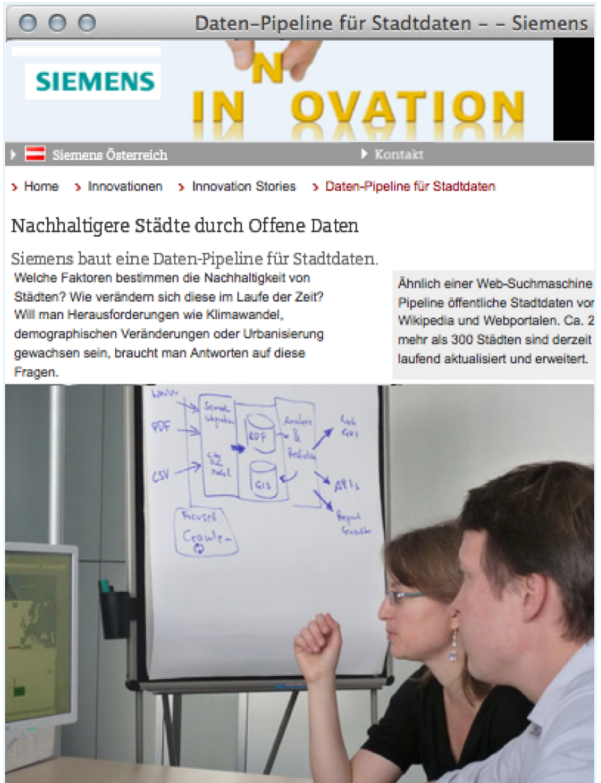
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

SIEMENS

## Open City Data Pipeline

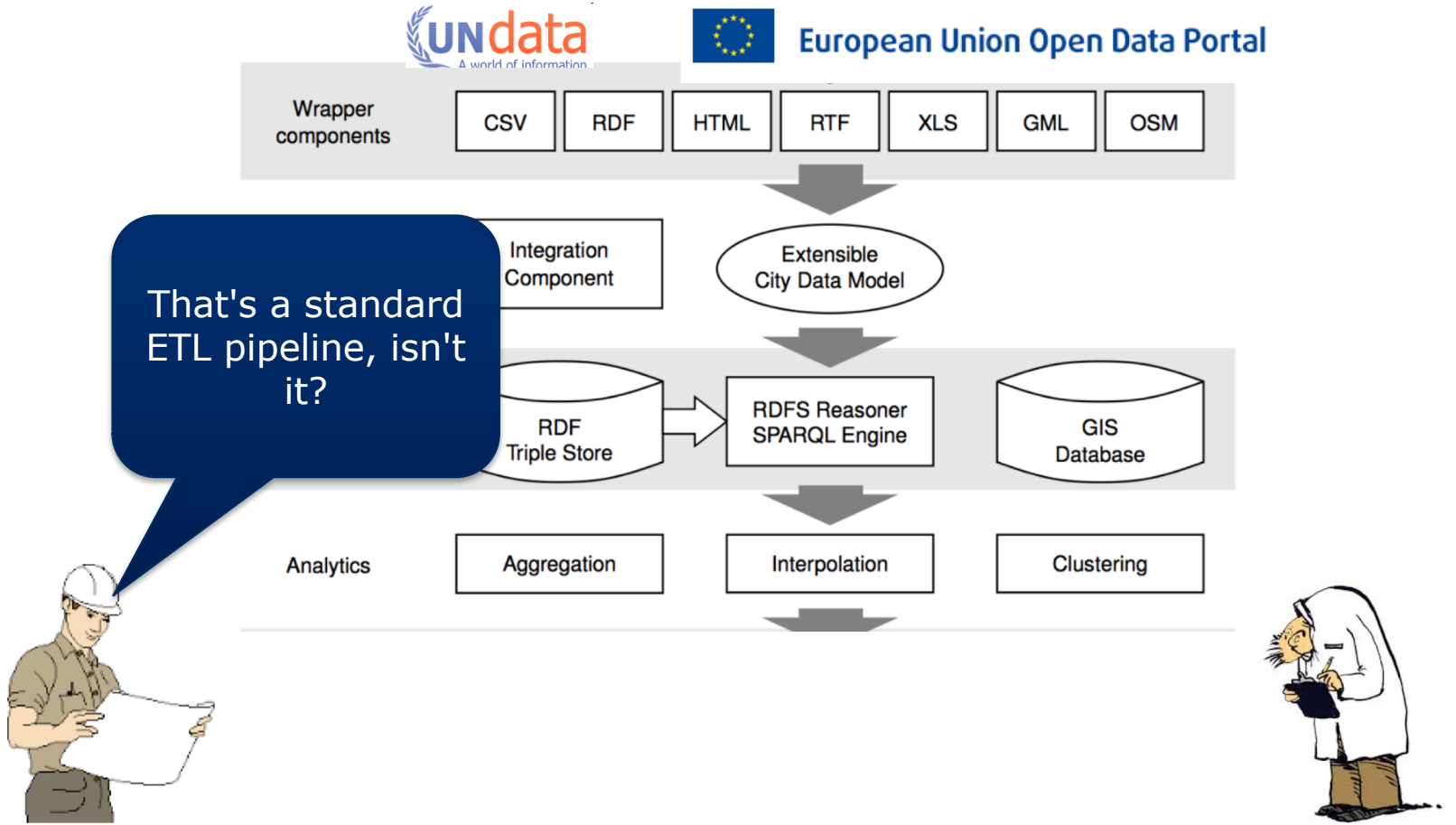
We present the City Data Pipeline – a system for gathering city performance indicators published as Open Data in order to ease the compilation of studies and reports used within Siemens. Under the assumption that Open Data provides means to automatise tedious data research tasks, we have built a system 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**



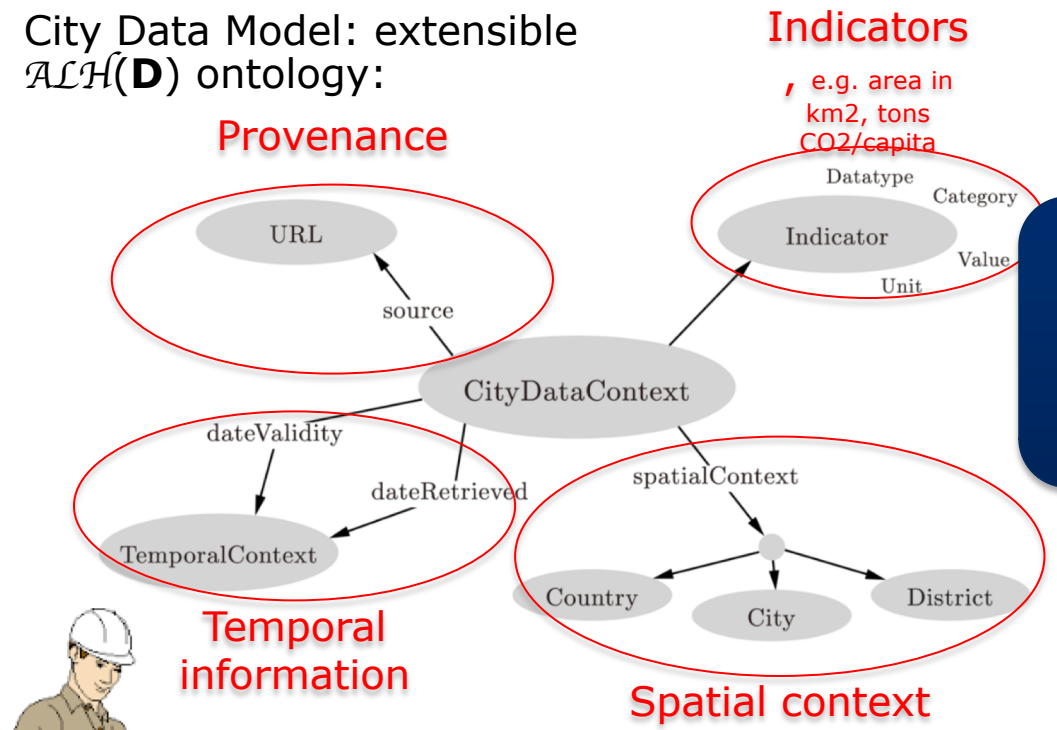

EFMD EQUIS ACCREDITED

# A concrete use case: The "City Data Pipeline"



# A concrete use case: The "City Data Pipeline"

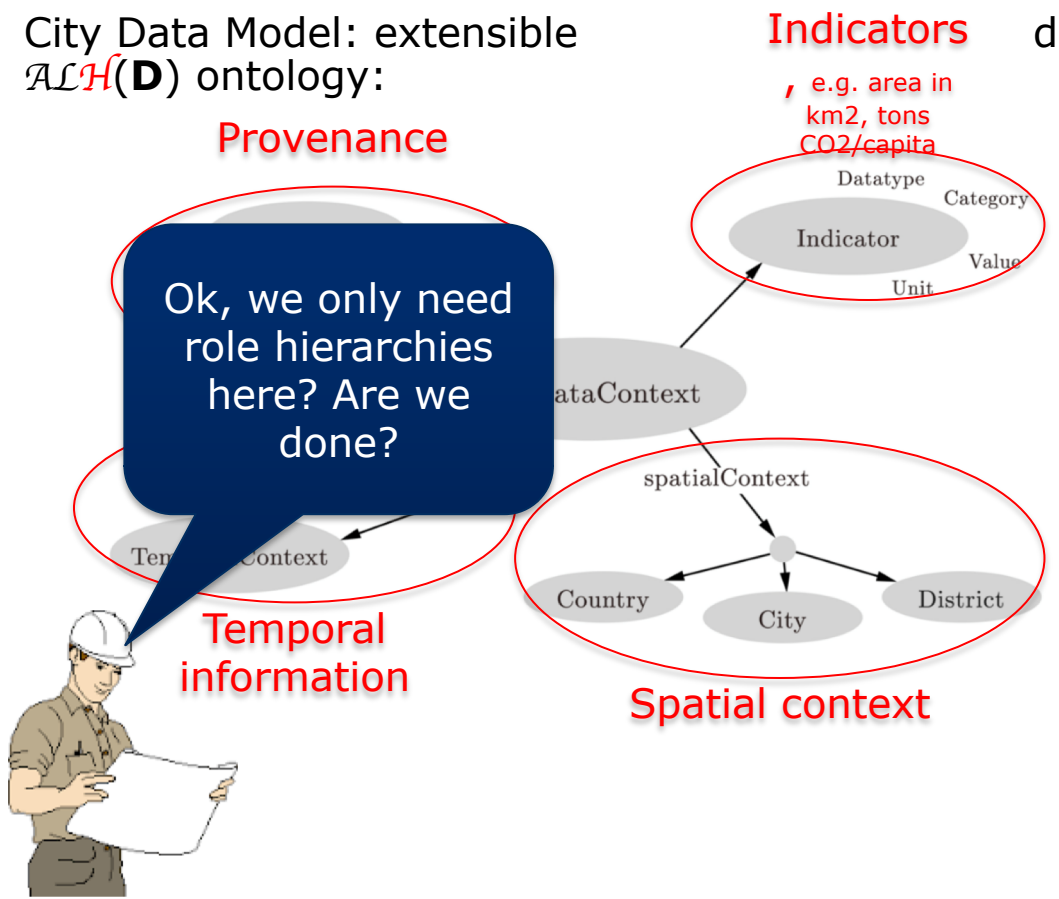
City Data Model: extensible $\mathcal{ALH}(\mathbf{D})$ ontology:




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

# A concrete use case: The "City Data Pipeline"

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

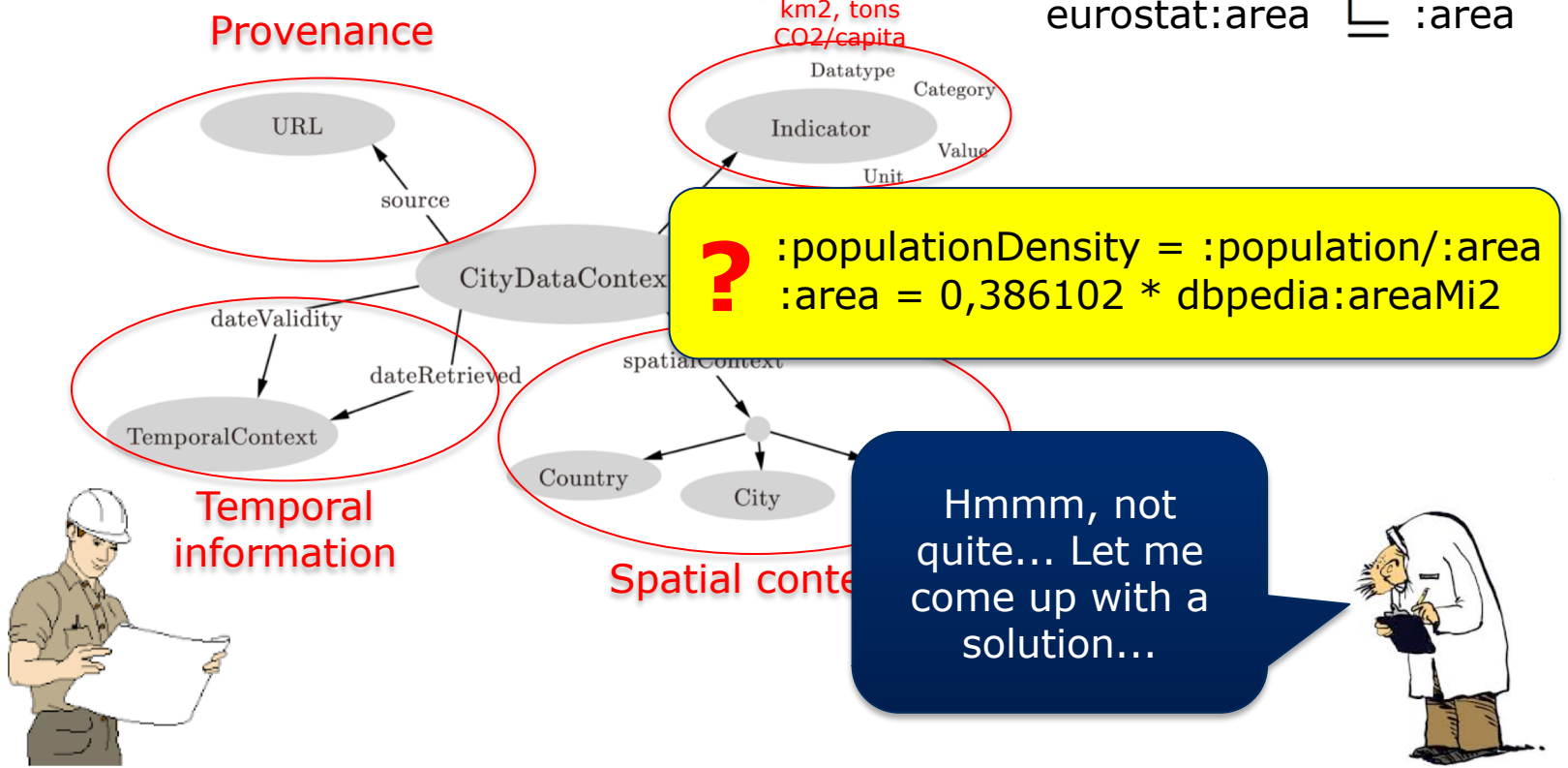



dbpedia:areakm $\sqsubseteq$ :area

eurostat:area $\sqsubseteq$ :area

# A concrete use case: The "City Data Pipeline"

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



# Can equational knowledge co-exist with OWL?

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

## 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

EFMD EQUIS ACCREDITED

# *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

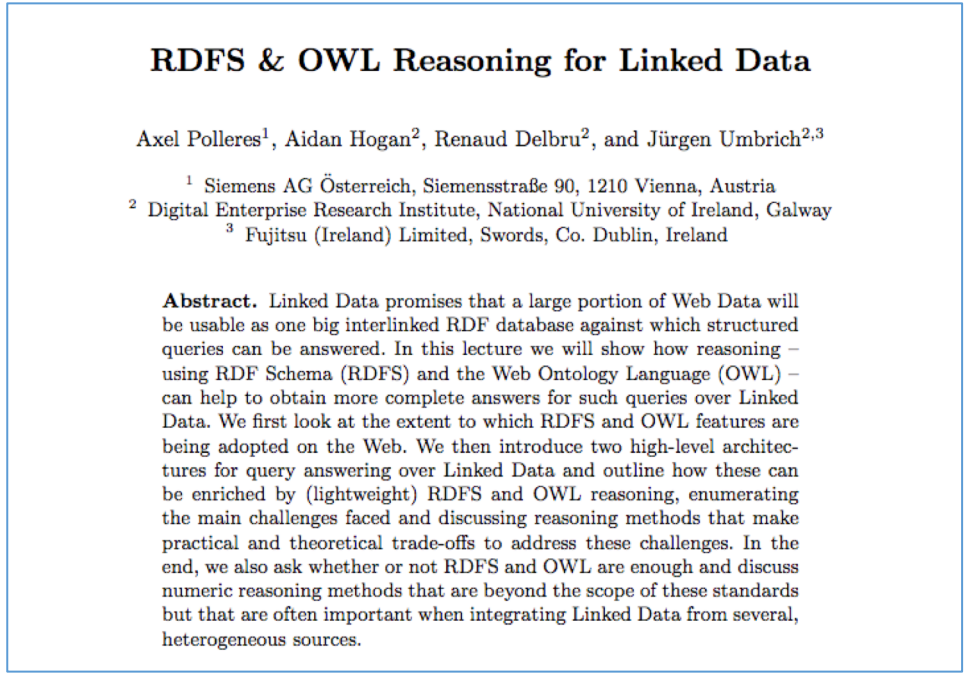
:populationDensity **:defineByEquation** “population/:area” .
:area **:defineByEquation** “areaMi2 * 0,386102 ” .
dbPedia:populationTotal **:rdfs:subPropertyOf** :population.

- 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

# This is more or less where our RW2013 lecture ends...



**RDFS & OWL Reasoning for Linked Data**

Axel Polleres[1], Aidan Hogan[2], Renaud Delbru[2], and Jürgen Umbrich[2,3]

[1] Siemens AG Österreich, Siemensstraße 90, 1210 Vienna, Austria
[2] Digital Enterprise Research Institute, National University of Ireland, Galway
[3] Fujitsu (Ireland) Limited, Swords, Co. Dublin, Ireland

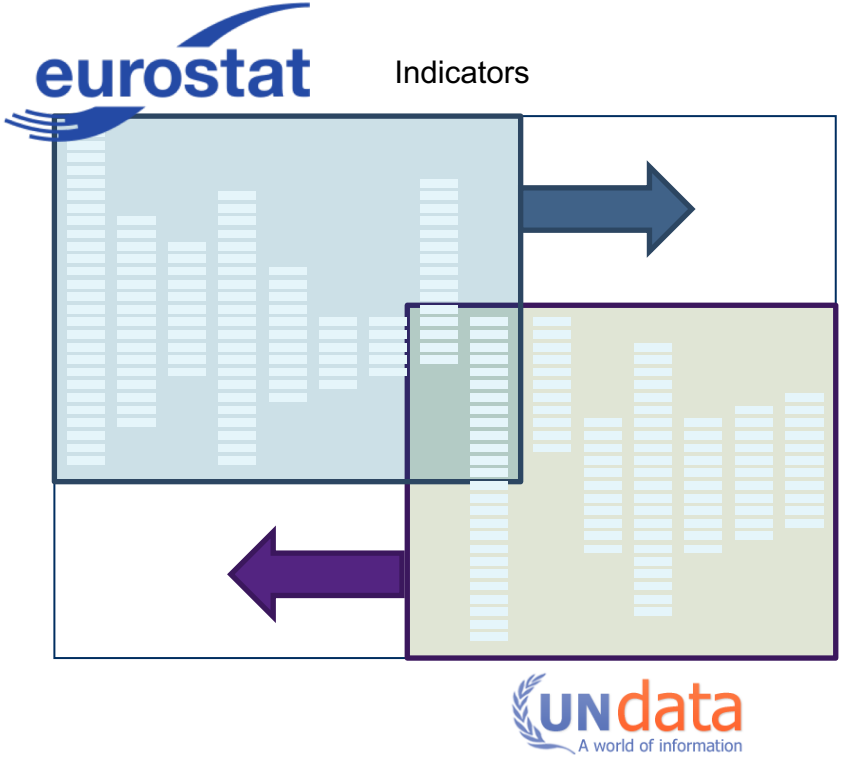
**Abstract.** Linked Data promises that a large portion of Web Data will be usable as one big interlinked RDF database against which structured queries can be answered. In this lecture we will show how reasoning – using RDF Schema (RDFS) and the Web Ontology Language (OWL) – can help to obtain more complete answers for such queries over Linked Data. We first look at the extent to which RDFS and OWL features are being adopted on the Web. We then introduce two high-level architectures for query answering over Linked Data and outline how these can be enriched by (lightweight) RDFS and OWL reasoning, enumerating the main challenges faced and discussing reasoning methods that make practical and theoretical trade-offs to address these challenges. In the end, we also ask whether or not RDFS and OWL are enough and discuss numeric reasoning methods that are beyond the scope of these standards but that are often important when integrating Linked Data from several, heterogeneous sources.

Axel Polleres, Aidan Hogan, Renaud Delbru, and Jürgen Umbrich. RDFS & OWL reasoning for linked data. In Sebastian Rudolph, Georg Gottlob, Ian Horrocks, and Frank van Harmelen, editors, *Reasoning Web. Semantic Technologies for Intelligent Data Access (Reasoning Web 2013)*, volume 8067 of *Lecture Notes in Computer Science (LNCS)*, pages 91--149. Springer, Mannheim, Germany, July 2013.

# Next Challenge – Too many Missing values

Goal: equational knowledge is not enough…

Idea: using both rules and “second wave AI” (machine learning, statistical inference…) 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

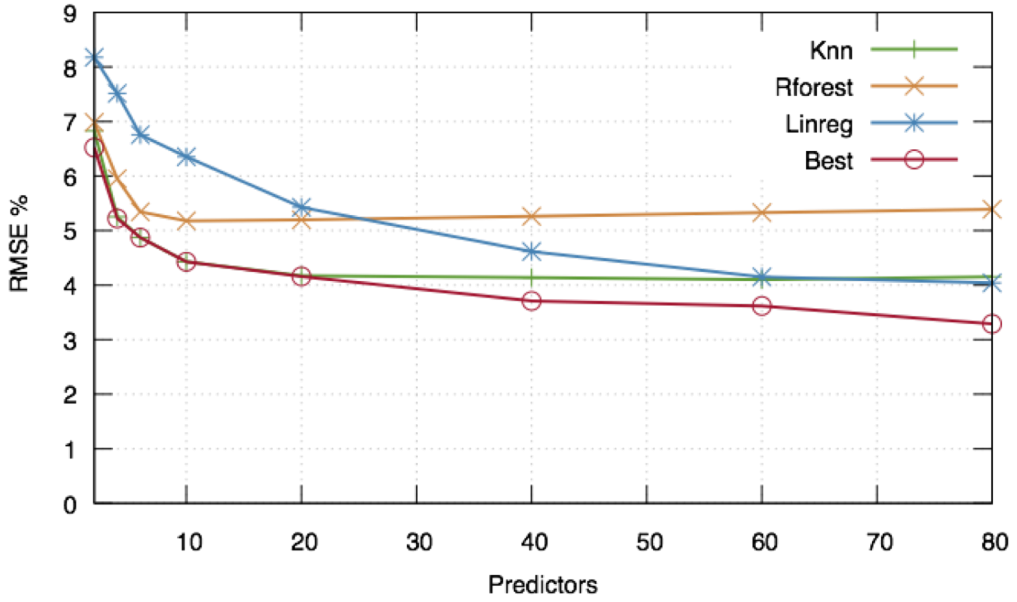
| | Vienna | Augsburg | Valletta | Marbella | Stockholm | Funchal |
|---|---|---|---|---|---|---|
| Cars | 655806 | 111561 | 95858 | | | |
| Nationals | 1342704 | 216289 | 203657 | | | |
| Women per 1000 Men | 109.8 | 108.7 | 101.9 | | | |
| Available Beds per 1000 | | | | 138.3 | 14969 | 166.1 |
| Average 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)
  -
  -

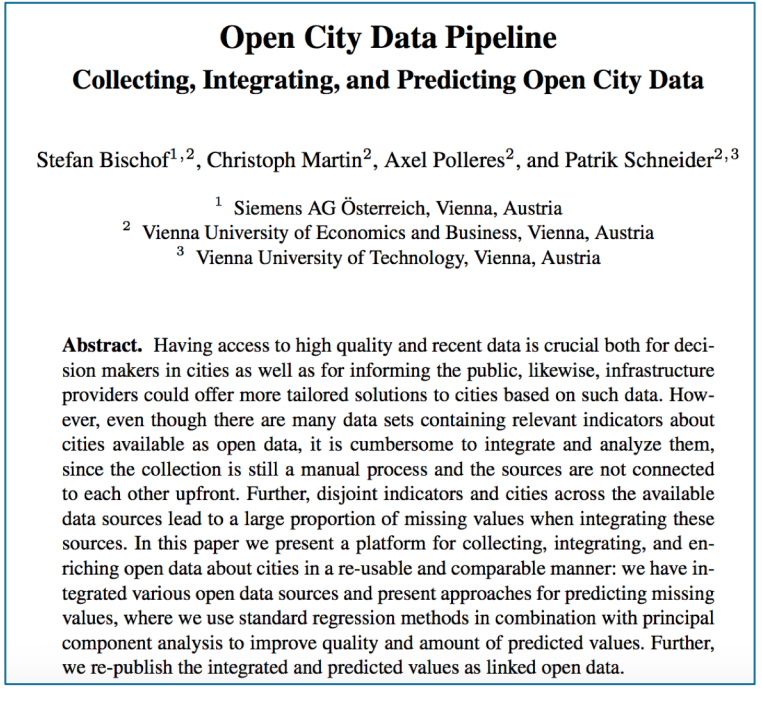
# Missing Values – Hybrid approach choose best prediction method per indicator:

- Instead of using indicators directly we use Principle Components, 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. 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

**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.

Next step:

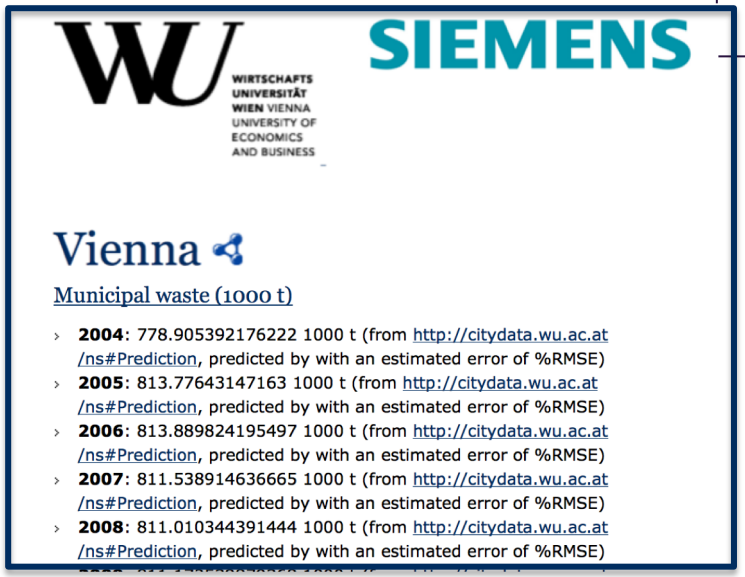
Combine ML and equations "iteratively" (under submission)

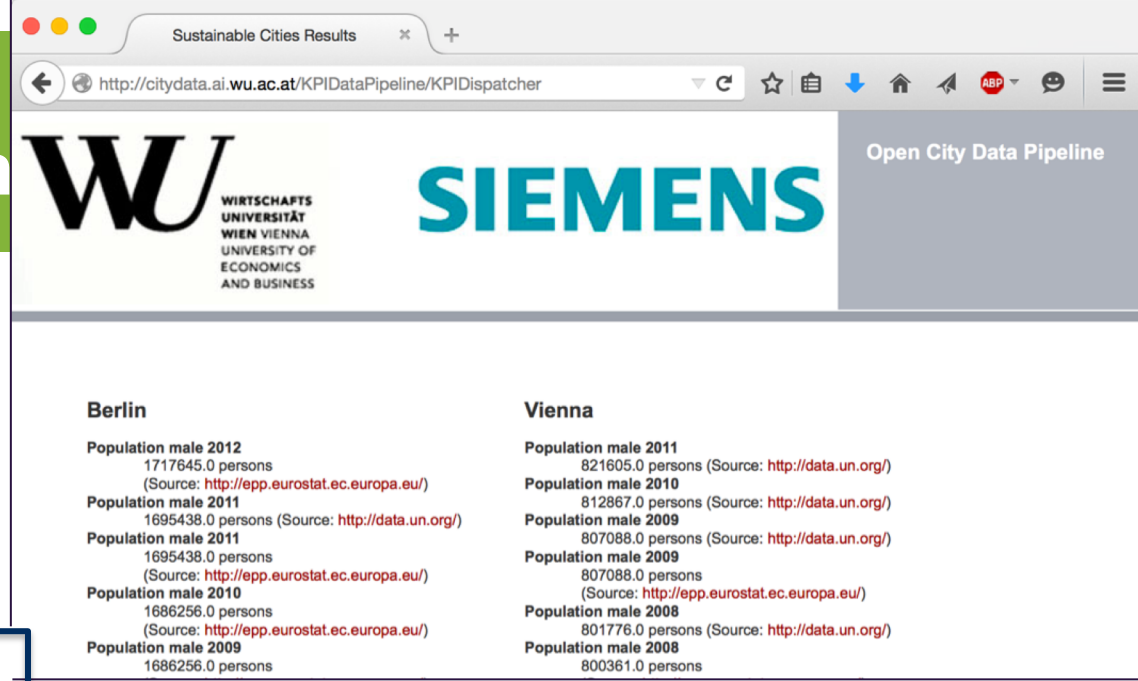
http://epub.wu.ac.at/5438/

# City Data Pipelin

## 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...




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

# A good start, but… many open questions:

## (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 automate **finding relevant open data?**
- What is Open Data? i.e., which data am I **allowed to use** freely?
- How can I **assess the quality** of Open Data?
- Eventually:
  - How can I build a **scalable repository** of Open Data? (search engine, archive, etc.)

# BTW, hope to see many of you in Vienna!

- ISWC2017 **21-25 October**
- Workshop paper submissions **July 21, 2017**

- ***Maybe you jointly develop a nice idea for a workshop paper still here!?***

- **We're here to help ☺**

- *Plus: **we're hiring** a PhD student on an exciting Enterprise Linked Data Integration Project starting in autumn;-)*
  *Check Twitter: @AxelPolleres or e-mail me: axel.polleres@wu.ac.at*


https://iswc2017.semanticweb.org/

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

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

# Challenges and Open Problems – Group Work!

Axel Polleres, Sebastian Neumaier 13th Reasoning Web Summer School, 2017

# What did we hear today?

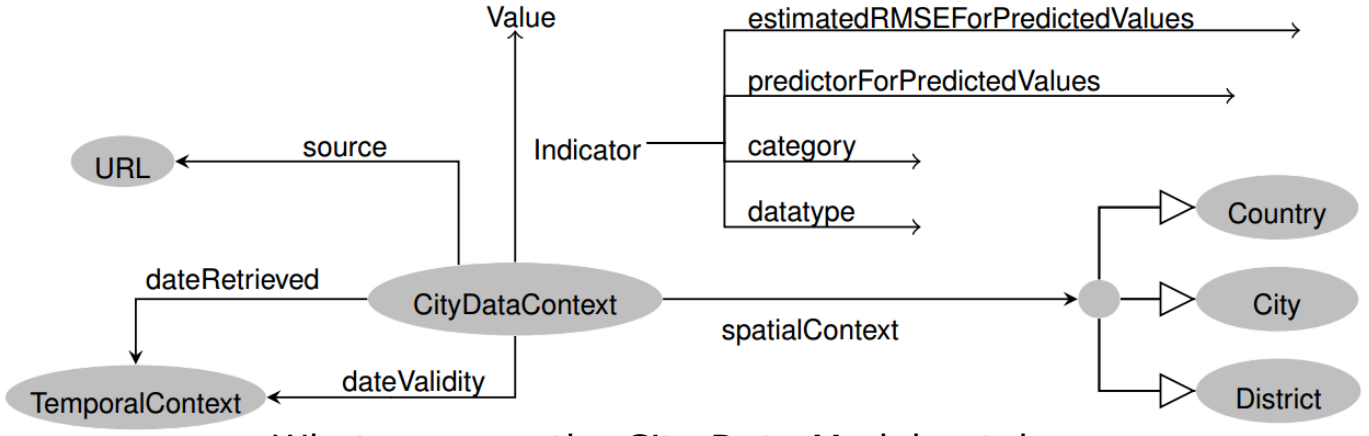
- What is Open Data
- No Linked Data available on (governmental) data portals
- Issues relating to licensing and provenance of data
- High heterogeneity and quality issues
- Attempts to bring structure to the “Web of Open Data”

# Group Work

- **4-6 students per group**
- Brainstorm on one of the presented problems
- How can you (your *work/ideas*) help to overcome the problems

- Outcome:
  - How would you tackle it, relevant literature, evaluations…
  - No in-depth solutions
  - Discuss and hear about the work of your colleagues
  - Get out of your comfort zone: work on topic not related to your area
  - Present discussion and findings ether in a slide (e.g. Google slides) or ideally present **a workshop paper abstract** (for ISWC ;))
    **(2-5 minutes)**

# P1: How to build and formalize an ontology for open data?




What we saw: the City Data Model ontology

**Your task:**

Come up with an approach to increase interoperability of Open Data portals:

- The above model is clearly incomplete if we take all Open Data
- Does it make sense to come up with an general ontology?

# P2: Apply information extraction?

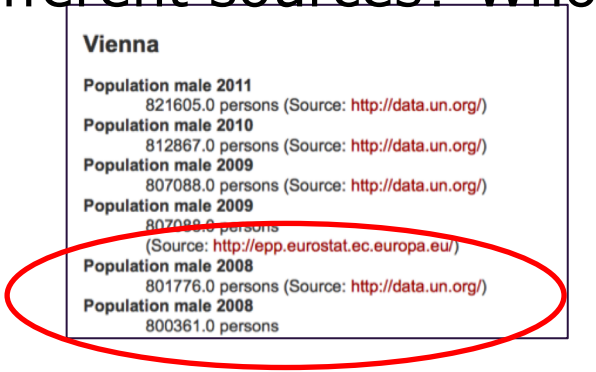
**Your task:**
Extract important dimensions for governmental Open Data

- *Spatial*
  - Location labels, identifiers (postal codes, NUTS), coordinates, …
- *Taxonomic*


- *Temporal*
  - date/time information, regular updates to datasets

# P3: How to assess provenance and trustworthy of a data source?

- Populations from different sources? Who to trust?




  - Can we come up with a confidentiality measure for a data source?
    - Weighted aggregation or choose one source?

- Can (description?) logics or rules help?
  - E.g. inconcistency detection/resolution?

# P4: Can you relate?

- Bridge the gap between what you saw in the lecture and **your PhD/research topic**!
- Can you (your *work/ideas*) help to overcome a certain problem discussed in the lecture?

# Acknowledgements


- This work has been partially funded
  - *by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under the program "ICT of the Future"*

Federal Ministry
for Transport,
Innovation and Technology

  - *by the European Commission under the H2020 Programme*

SPECIAL

European Commission

Horizon 2020
European Union funding
for Research & Innovation

# Let‘s get started!

goo.gl/YhR5cQ