

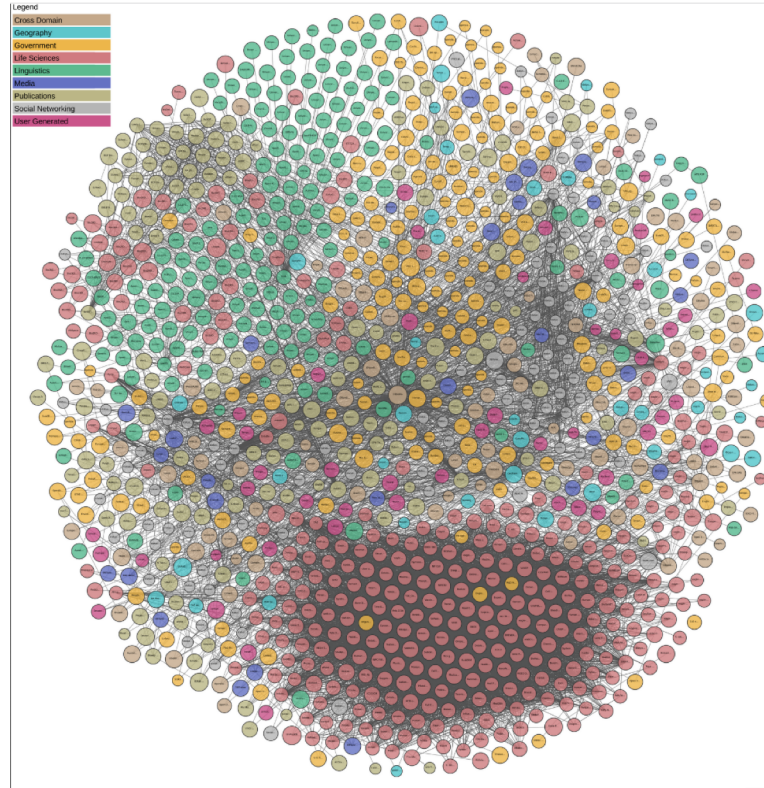
Building an Open Knowledge Graphs *for* and *from* Open Data



Axel Polleres

Joint work with: Sebastian Neumaier, Jürgen Umbrich

When we hear about Open Data and Knowledge Graphs... many think about Linked Open Data...



The Linked Open Data Diagram from lod-cloud.net Latest release 04-30-2018- 1184 Datasets

Open Data is a Global Trend!

- EU & Austria, but also the (previous) US and UK administration are/were pushing Open Data!

THE WORLD BANK
Open Data
wien **Open Government Data**
Offene Daten für Wien

UNdata

Open Data Berlin

Open Government Data Österreich

Opening up Europe's public data

DATA.GOV

DATA TOPICS IMPACT APPLICATIONS DEVELOPERS CONTACT

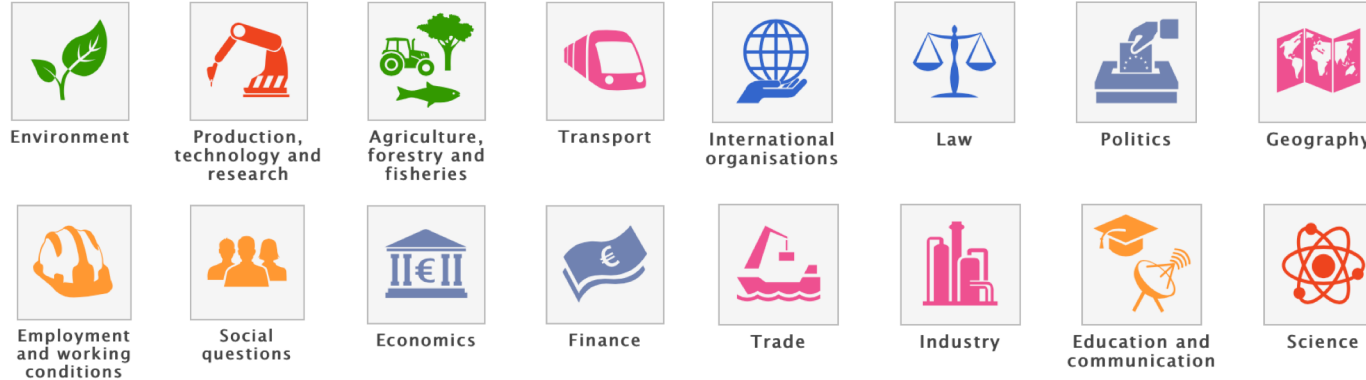
The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

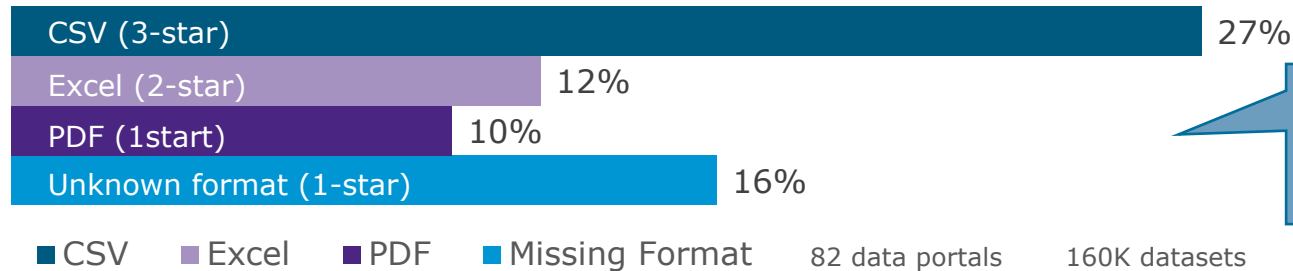
GET STARTED
SEARCH OVER 170,714 DATASETS

Federal Student Loan Program Data

(Structured) Open Data comes in various ways



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



RDF/Linked Data? Not significant

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

Open Data as a Global Trend:

Country	URL	Datasets
United States	data.gov	170.7k
Canada	open.canada.ca	79.1k
UK	data.gov.uk	45.1k
France	www.data.gouv.fr	34.2k
Russia	opengovdata.ru	30.3k
Japan	data.go.jp	21k
Italy	dati.gov.it	20.4k
Germany	govdata.de	19.8k

Data portals of the G8 countries

Different portals...

The screenshot shows the DATA.GOV website. At the top, there is a navigation bar with 'DATA', 'TOPICS', 'IMPACT', 'APPLICATIONS', 'DEVELOPERS', and 'CONTACT'. Below this is the 'DATA CATALOG' header with 'Datasets' and 'Organizations' tabs. The main content area displays the 'Housing Affordability Data System (HADS)' dataset, updated on March 8, 2017. It includes a description of the data, 'Access & Use Information' (Public, License), and 'Downloads & Resources' (Comma Separated Values File, 13730 views, Download button). On the left, there is a sidebar with 'Topics' (BusinessUSA, Consumer, Energy, Finance, Opportunity), 'Publisher' (US Department of Housing and Urban Development), 'Contact' (Shula Markland), 'Share on Social Sites' (Google+, Twitter, Facebook), and 'Dates' (Metadata Created Date: March 7, 2014; Metadata Updated Date: March 8, 2017). At the bottom, there is a 'Metadata Source' section for 'Data.json Metadata' and a tag cloud with terms like 'affordability', 'cost', 'fmr', 'households', 'housing', 'income', 'rent', and 'renter'.

Additional Metadata

Resource Type	Dataset
Metadata Created Date	March 7, 2014
Metadata Updated Date	March 8, 2017
Publisher	US Department of Housing and Urban Development
Unique Identifier	HUD031
Maintainer	Shula Markland
Maintainer Email	Shula.Markland@HUD.gov

The screenshot shows the data.gv.at website. It features a search bar with the text 'Suchbegriff (z.B. Finanzen, Wahlen)' and a 'Suche starten' button. Below the search bar, there are links for 'Datenkatalog', 'Apps & News', and 'Katalog durchstöbern'. The top right corner has an 'API' icon. The main navigation bar includes 'Startseite', 'Daten', 'Dokumente', 'Anwendungen', and 'Infos'.

Katalog Bildungsausgaben

Bildungsausgaben; Regionale Gliederung; Bildungseinrichtung

Daten und Ressourcen

OGD_bildungsausgaben_BILDAUS_1	Entdecke
OGD_bildungsausgaben_BILDAUS_1_HEADER	Entdecke
OGD_bildungsausgaben_BILDAUS_1_C-A10-0	Entdecke
OGD_bildungsausgaben_BILDAUS_1_C-BARG-0	Entdecke
OGD_bildungsausgaben_BILDAUS_1_C-BABEL-0	Entdecke

Titel und Beschreibung Englisch	Educational expenditure
Veröffentlichende Stelle 📍	Statistik Austria
Datenverantwortliche Stelle 📍	Statistik Austria, Guglgasse 13, 1110 Wien, Austria
Kontaktseite der datenverantwortlichen Stelle 📍	http://www.statistik.at/web_de/kontakt
Datenverantwortliche Stelle - E-Mailkontakt 📍	open.data@statistik.gv.at
Lizenz 📍	Creative Commons Attribution License
Lizenz Zitat 📍	Datenquelle: CC-BY-3.0: Statistik Austria - data.statistik.gv.at
Link zur Lizenz 📍	https://creativecommons.org/licenses/by/3.0/

Weiterführende Metadaten - Link [📍](#)

	http://statcube.at/statcube/opendatabase?id=debildungsausgaben;http://www.statistik.at/web_de/statistiken/bildung_und_kultur/formales_bildungswesen/bildungsausgaben/index.html;http://www.statistik.at/web_en/statistics/education_culture/formal_education/educational_expenditure/index.html
	C-A10-0ZeitC-BARG-0Regionale Gliederung; C-BABEL-0Bildungseinrichtung; F-INSG Ausgaben (gesamt); F-TR_PAPersonalaufwand; F-TR_SA Sachaufwand; F-

Veröffentlichende Organisation bzw. Person [📍](#)

Statistik Austria

Kategorie [📍](#)

Bildung und Forschung

Finanzen und Rechnungswesen

Wirtschaft und Tourismus

Schlagworte [📍](#)

Bildungsausgaben

API - Link zu allen Metadaten [📍](#)

/api/3/action/package_show?id=7113735-2c65-328f-b57d-be941ada765e

RSS-Feeds für Statistik Austria

geänderte Datensätze

Letzte Änderung

30.04.2018 00:59:46

What do you find on Open Data Portals?

The screenshot shows the data.gv.at search page. The search bar contains 'Leopoldstadt' and the search button is labeled 'Suche starten'. Below the search bar, there are tabs for 'Daten & Dokumente' and 'Apps & News'. The main content area is titled 'Katalogsuche' and shows a search input field with 'Leopoldstadt' and a 'Suche starten' button. Below this, there is a 'Filter' section with a 'Filter einblenden' button. The search results section shows 'Suchergebnis zu "Leopoldstadt" (0 gefunden)' and 'Seite 1 von 0'. There is also a link 'alle Datensätze anzeigen' and a pagination control showing 'Ergebnisseiten: <- Erste Letzte (0) -> 1 Gehe zu'.




Not too much!

Suchergebnis zu "Leopoldstadt" (0 gefunden)

Seite 1 von 0

[alle Datensätze anzeigen](#)

Ergebnisseiten: <- Erste Letzte (0) -> 1 [Gehe zu](#)

 Suchergebnisse von [opendataportal.at](#) (0 gefunden)

Titel	Veröffentlichende Stelle / Datenverantwortliche	Veröffentlicht auf	Letzte Änderung auf	Format	Lizenz
Stelle	Stelle	opendataportal.at am	opendataportal.at		

Why is Search in Open Data a problem?

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

Structured Data in Web Search by Alon Halevy

HTML Tables

Nordkold Light	A.B. Pilsener Bryggeriet (Sweden)	4.7	110			
Turboled	Alba Brewing Company	3.8	188	15	28	80
Abbey Ale	Alba Brewing Company	8.0	230	18	32	25
Pilsner	Alba Brewing Company	6.0	180	11	20	18
Jockamo	Alba Brewing Company	6.5	180	13	32	18
Red Ale	Alba Brewing Company	5.2	181	15	36	18
Aurora	Alba Brewing Company	4.5	128	10	17	15
Black	Alba Brewing Company	4.5	187	16	25	13
Fat Fat	Alba Brewing Company	5.4	187	15	20	12
Reaktion	Alba Brewing Company	5.0	187	15	20	9
Andromeda	Alba Brewing Company	6.0	235	19	25	9
Purple Haze	Alba Brewing Company	4.2	128	11	13	8
Silberne	Alba Brewing Company	5.1	150	11	17	5
Strawberry	Alba Brewing Company	4.2	128	11	13	5
Black Out Shore	Alba Brewing Company	7.0	200	10	30	4
Wheat	Alba Brewing Company	4.2	128	10	15	3
Golden	Alba Brewing Company	4.2	128	10	11	3
Light	Alba Brewing Company	4.0	118	8	10	3
Christmas Ale	Alba Brewing Company	7.5				30

research.google.com/tables

data.gv.at
Suchbegriff (z.B. Finanzen, Wahlen)
Datenkatalog Apps & News
Startseite Daten Dokumente Linked Data Anwendungen News

Katalog Bevölkerung in Wien: Bezirk - Geschlecht

VS.

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

Open Data Search is hard...

- a) No natural language „cues“ like in Web tables...
- b) Existing knowledge graphs don't cover the domain of "Open Data" well
- c) Open Data is not properly geo-referenced

Data Integration as Search

Coffee Consumption around the world

Merge Select a table

Suggest table matching on:

World Population 2014

Not in this talk: What is a Knowledge Graph?

... good question!



Official Blog
Insights from Googlers into our product
technology, and the Google culture

Introducing the Knowledge Graph: things, not strings
May 16, 2012

Cross-posted on the [Inside Search Blog](#)

Search is a lot about discovery—the basic human need to learn and broaden horizons. But searching still requires a lot of hard work by you, the user. So today I'm really excited to launch the Knowledge Graph, which will help you discover new information quickly and easily.

Take a query like [taj mahal]. For more than four decades, search has essentially been about matching keywords to queries. To a search engine the words [taj mahal] have been just that—two words.

The screenshot shows a Google search for "taj mahal". The search results include Wikipedia entries for the monument and a musician, as well as a listing for the Trump Taj Mahal casino. A Knowledge Graph panel is visible on the right, displaying a map of the Taj Mahal in Agra, India, and a list of related entities: Taj Mahal (monument), Taj Mahal (musician), and Trump Taj Mahal Casino Resort. The Knowledge Graph panel also includes a "See results about" section with icons for each entity.

Says more what a KG **does**
than what it **is**...
"interesting things and
[understanding their] relationships"

In this talk:

2 applications for Knowledge Graphs for Open Data Search!

- What we do: 2 approaches how knowledge graphs could help to solve the Open Data search problem (aside the obvious):
 1. Hierarchical labelling of Labeling of numeric data
 2. Hierarchical labelling of Spatio-Temporal entities

Example Table

<i>federal state</i>	<i>district</i>	<i>year</i>	<i>sex</i>	<i>population</i>
Upper Austria	Linz	2013	male	98157
Upper Austria	Steyr	2013	male	18763
Upper Austria	Wels	2013	male	29730
...

Open Data CSVs look more like this

<i>NUTS2</i>	<i>LAU2_NAME</i>	<i>YEAR</i>	<i>SEX</i>	<i>P_TOTAL</i>
AT31	Linz	2013	1	98157
AT31	Steyr	2013	1	18763
AT31	Wels	2013	1	29730
...

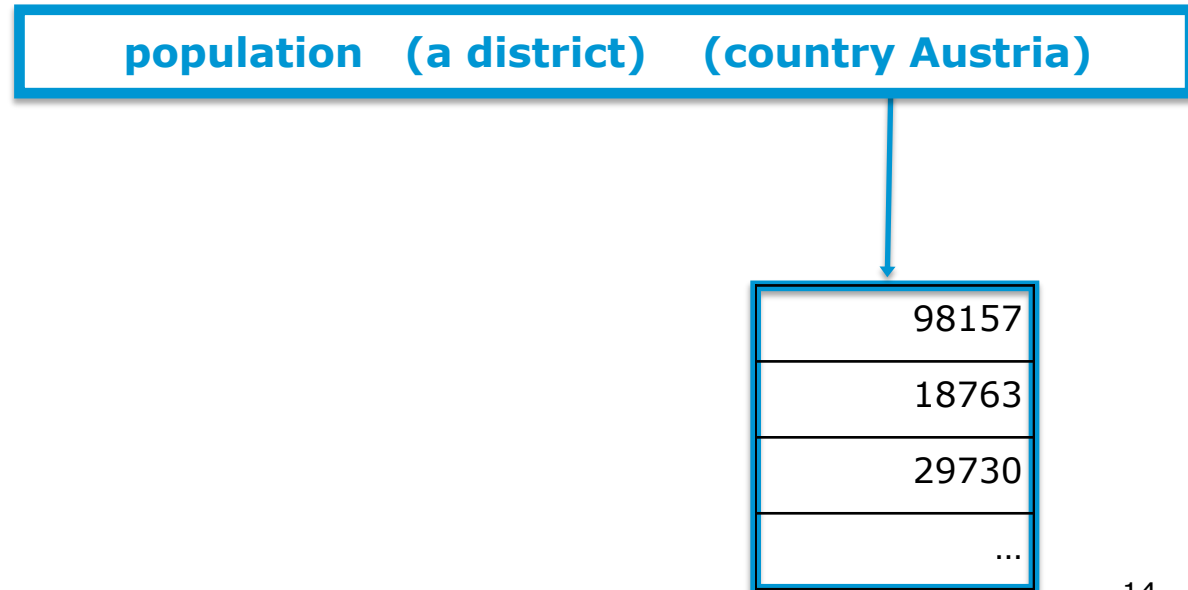
Why not use the numeric values?

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

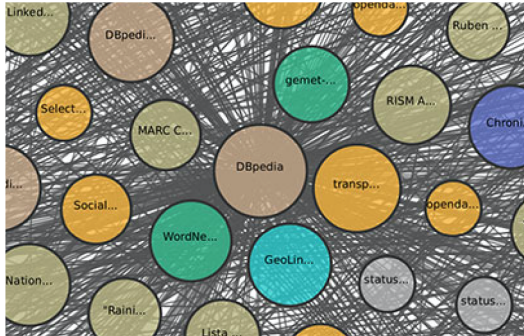
<i>NUTS2</i>	<i>LAU2_NAME</i>	<i>YEAR</i>	<i>SEX</i>	<i>P_TOTAL</i>
AT31	Linz	2013	1	98157
AT31	Steyr	2013	1	18763
AT31	Wels	2013	1	29730
...

Why not use numeric values?

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



Background Knowledge Graph

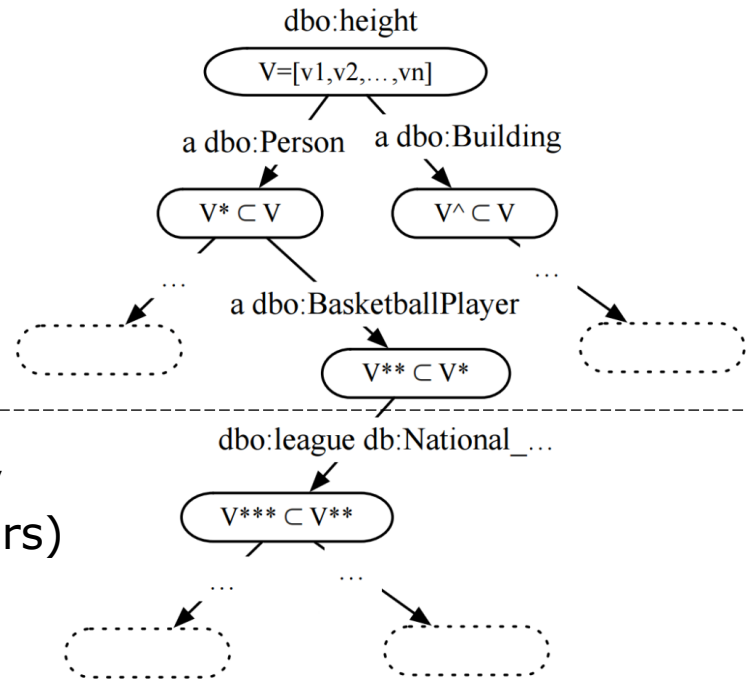


What's in there?

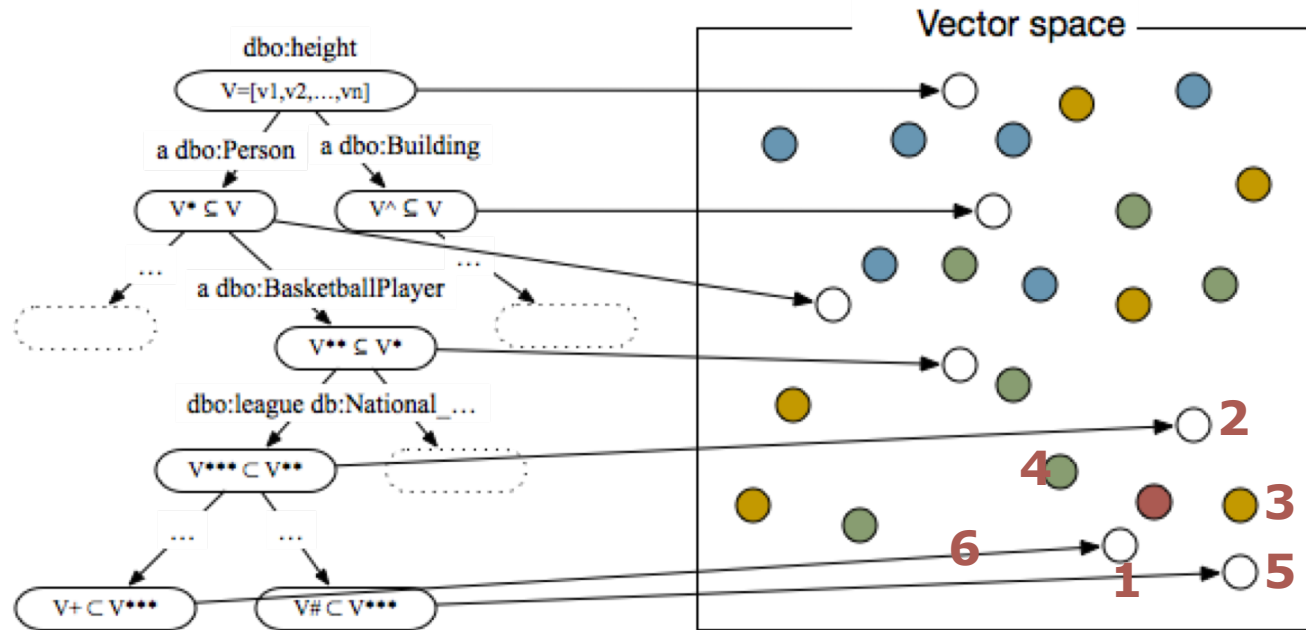
- Cities
 - **Population**
 - **Area**
 - Country
 - Location (**Coordinates**)
 - Economic indicators
 - ...
- Organisations:
 - **Revenues**
 - Board members
 - ...
- Persons (e.g. celebrities, sports)
 - Name
 - Profession
 - **Height**
- Landmarks (e.g. famous buildings)
 - **Country**
 - **Location**
 - **Height**
- Events
 - **Dates**
 - **Location**

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)



Label based on Nearest Neighbors



Example OD Labelling

populationTotal (a Settlement)
populationDensity (a City)

NUTS1	NUTS2	NUTS3	DISTRICT_CODE	T	WV	WK	BZ	SPR	WBER	ABG.	UNG.	OEVP	SPOE	FPOE	GRUE	BZOE	NEOS
AT1	AT13	AT130		1	9	0	0	0	1163061	503284	9386	81974	136391	89963	103249	1516	44891
AT1	AT13	AT130		2	9	1	0	0	111279	52674	774	9344	12395	6482	14154	114	5412
AT1	AT13	AT130		2	9	2	0	0	98379	51785	646	10324	10236	4700	15398	124	6569
AT1	AT13	AT130		2	9	3	0	0	110527	45483	810	5317	13304	7816	10944	115	3613
AT1	AT13	AT130		2	9	4	0	0	229521	84387	1953	10097	27922	21091	11631	256	5299
AT1	AT13	AT130		2	9	5	0	0	212262	97755	1806	18703	25314	16613	19333	324	9175
AT1	AT13	AT130		2	9	6	0	0	175288	82790	1321	17560	19059	11765	18996	242	8389
AT1	AT13	AT130		2	9	7	0	0	225805	88410	2076	10629	28161	21496	12793	341	6434
AT1	AT13	AT130	90301	3	9	1	3	0	57528	27320	412	4938	6586	3567	6969	68	2789
AT1	AT13	AT130	90401	3	9	1	4	0	21000	11027	138	2401	2253	1068	3082	26	1277
AT1	AT13	AT130	90501	3	9	1	5	0	32751	14327	224	2005	3556	1847	4103	20	1346

Lessons learned

- We can assign fine-grained semantic labels
 - **If there is enough evidence in BK**
- *However*: Missing domain knowledge for labelling OD

Future work:

- Complementary to existing approaches (column header labeling, entity linking and relation extraction)
- Combined approaches may improve results
- Focusing on *core dimensions of specific domains* e.g. city data, may be more promising than “general” value labeling.

International Semantic Web conference 2016:

Multi-level semantic labelling of numerical values

Sebastian Neumaier¹, Jürgen Umbrich¹, Josiane Xavier Parreira², and Axel Polleres¹

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

What else can we do/use?

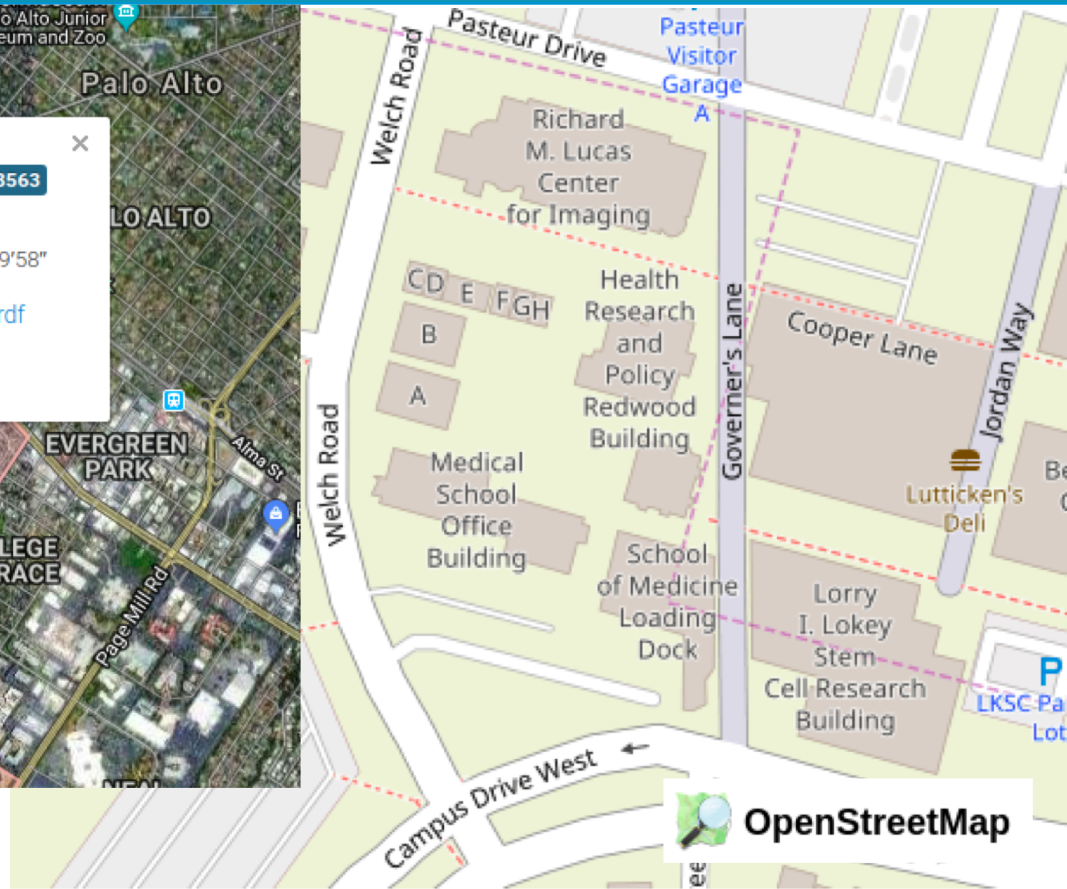
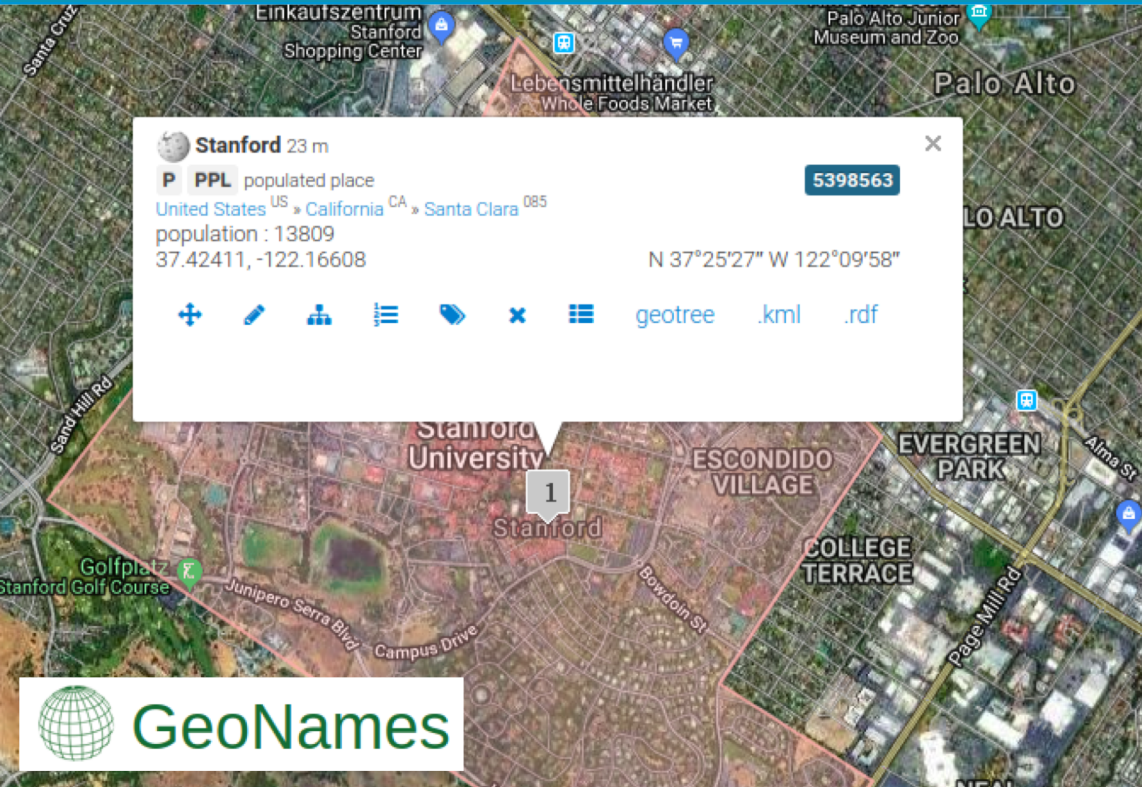
Focus on specific dimensions:

- Particularly **temporal** and **geospatial** queries require better support [2]

<i>NUTS2</i>	<i>LAU2_NAME</i>	<i>YEAR</i>	<i>SEX</i>	<i>AGE_TOTAL</i>
AT31	Linz	2013	1	98157
AT31	Steyr	2013	1	18763
AT31	Wels	2013	1	29730
...

[2] Emilia Kacprzak, et al.: A Query Log Analysis of Dataset Search. International Conference on Web Engineering (2017)

Available Geospatial Knowledge Bases



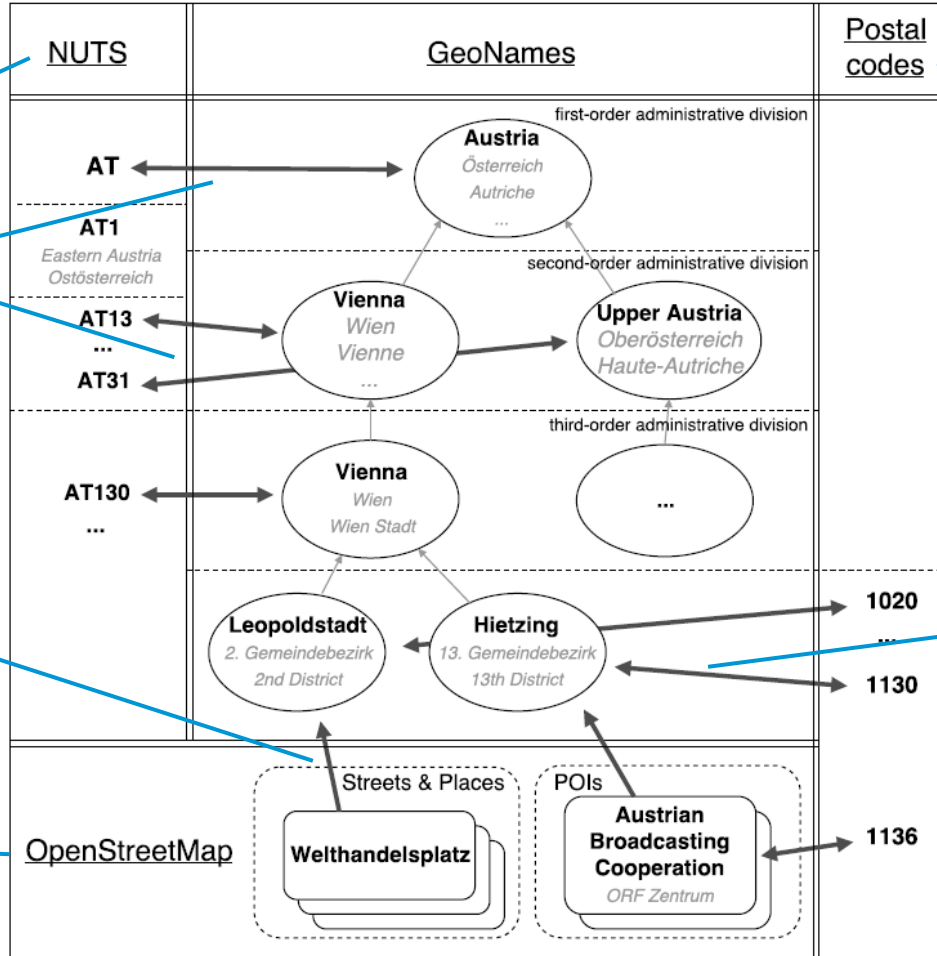
Geo-Knowledge Graph Construction

European Classification of Territorial Units

Wikidata links

Mapping OSM entities to GeoNames regions

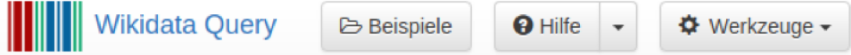
Extracting OSM streets and places



Wikidata, GeoNames


Wikidata links

Available Temporal Knowledge



```

1 SELECT ?itemLabel ?countryLabel ?startLabel ?endLabel
2 WHERE
3 {
4   ?item wdt:P31 wd:Q3558349 ;
5         wdt:P17 ?country ;
6         wdt:P580 ?start ;
7         wdt:P582 ?end .
8   SERVICE wikibase:label { bd:serviceParam wikibase:language "[A
9 }
                
```



Periods

Viewing 4226 - 4250 of 5134

Show periods at a time.

Previous

1 2 ... 169 170 171 ... 205 206

▲ Label	Earliest start	Latest stop
Tairona Period	900	1600
Taisho Era	1912	1926
Taishō period, 1912-1926	1912	1926
Taizong	976	997
Taizong Liao dynasty	926	947

itemLabel	countryLabel	startLabel	endLabel
Kabinett Lincoln	Vereinigte Staaten	1861-03-04T00:00:00Z	1865-04-15T00:00:00Z
Presidency of Cristina Fernández de Kirchner	Argentinien	2007-12-10T00:00:00Z	2015-12-09T00:00:00Z
Presidency of Fidel V. Ramos	Philippinen	1992-06-30T00:00:00Z	1998-06-30T00:00:00Z

Temporal Knowledge Graph Construction

```
CONSTRUCT {
  ?event rdfs:label ?label ; dcterms:isPartOf ?Parent ; dcterms:coverage ?geocoordinates ;
  timex:hasStartTime ?StartDateTime ; timex:hasEndTime ?EndDateTime ; dcterms:spatial ?geoentity .
} WHERE {
  # find events with (for the moment) English, German, or non-language-specific labels:
  ?event wdt:P31/wdt:P279* wd:Q1190554 . ?event rdfs:label ?label .
  FILTER( LANG(?label) = "en" || LANG(?label) = "de" || LANG(?label) = "" ).
  # restrict to certain event categories, e.g. (for the moment) elections and sports events:
  { # elections #sports competitions
    { ?event wdt:P31/wdt:P279* wd:Q40231 } UNION { ?event wdt:P31/wdt:P279* wd:Q13406554 }
  }
  { # with a point in time or start end end date
    { ?event wdt:P585 ?StartDateTime . FILTER ( ?StartDateTime > "1900-01-01T00:00:00"^^xsd:dateTime ) }
    UNION
    { ?event wdt:P580 ?StartDateTime. FILTER ( ?StartDateTime > "1900-01-01T00:00:00"^^xsd:dateTime)
      ?event wdt:P582 ?EndDateT. FILTER ( DATATYPE(?EndDateT) = xsd:dateTime )
    }
  }
  OPTIONAL { ?event wdt:P361 ?Parent }
  # specific spatialCoverage if available
  OPTIONAL { ?event wdt:P276?(/wdt:P17|wdt:P131) ?geoentity }
  OPTIONAL { ?event wdt:P276?/wdt:P625 ?geocoordinates }
  BIND ( if(bound(?EndDateT), ?EndDateT, xsd:dateTime(concat(str(xsd:date(?StartDateTime)),"T23:59:59"))) AS ?EndDateTime )
}
```



- Named events and their labels
- Links to parent periods

```
CONSTRUCT {
  ?P rdfs:label ?label ; dcterms:isPartOf ?Parent ; dcterms:spatial ?geo ;
  timex:hasStartTime ?StartDateTime ; timex:hasEndTime ?EndDateTime .
} WHERE {
  {
    { ?P skos:prefLabel ?label } UNION { ?P skos:altLabel ?label } UNION { ?P rdfs:label ?label }
  }
  ?P time:intervalFinishedBy ?End ; time:intervalStartedBy ?Start.
  OPTIONAL { ?P periodo:spatialCoverage ?geo }
  OPTIONAL { ?P dcterms:spatial ?geo }
  OPTIONAL { ?P dcterms:isPartOf ?Parent. }
  OPTIONAL{ ?End time:hasDateTimeDescription ?EndTime .
    OPTIONAL{ ?EndTime time:year ?EndYear }
    OPTIONAL{ ?EndTime periodo:latestYear ?EndYear }
  }
  OPTIONAL{ ?Start time:hasDateTimeDescription ?StartTime .
    OPTIONAL{ ?StartTime time:year ?StartYear }
    OPTIONAL{ ?StartTime periodo:earliestYear ?StartYear }
  }
  OPTIONAL{ ?Start (!periodo:aux)+ ?StartYear. FILTER (isLiteral(?StartYear)) }
  OPTIONAL{ ?End (!periodo:aux)+ ?EndYear. FILTER (isLiteral(?StartYear)) }
}
FILTER( ?StartYear >= "1900"^^xsd:gYear || xsd:integer(?StartYear) >= 1900 ||
  ?EndYear >= "1900"^^xsd:gYear || xsd:integer(?EndYear) >= 1900 )

BIND( xsd:dateTime(concat(str(?StartYear),"-01-01T00:00:00")) as ?StartDateTime )
BIND( xsd:dateTime(concat(str(?EndYear),"-12-31T23:59:59")) as ?EndDateTime ) }
```



- Temporal extent: a single beginning and end date
- Links to the spatial coverage

Dataset Labelling

Metadata descriptions

- Geo-entities in titles, descriptions, organizations
- Restricted to „origin“ country of the dataset (from portal)
- Temporal tagging using HeideTime framework [3]

CSV cell value disambiguation

- Row context:
 - Filter candidates by potential parents (if available)
- Column context:
 - Least common ancestor of the spatial entities

Metadata

Tourismus - Ankünfte und Nächtigungen in Oberösterreich

Ankünfte und Nächtigungen in den oberösterreichischen Meldegemeinden ab dem Jahr 2000

Daten und Ressourcen

Ankünfte und Nächtigungen in OÖ seit dem Jahr 2000 Entdecke

Veröffentlichende Stelle: Land Oberösterreich

Datenverantwortliche Stelle: Land Oberösterreich, Abteilung Statistik

Lizenz: Creative Commons Namensnennung 3.0 Österreich

Link zur Lizenz: <https://creativecommons.org/licenses/by/3.0/at/deed.de>

Attributbeschreibung: NUTS2 => Bundesland Oberösterreich; Gemeindefürnummer bzw. Gemeindefürname => Erhebungsgemeindefür; Jahr => Kalenderjahr; Erhebungsgemeindefür lt. Tourismus-Statistik-Verordnung 2002 §2 Abs.7: Städte und Gemeinden mit mehr als 1.000 Gästenächtigungen im Kalenderjahr

CSV

NUTS2	Gemeindefürname	Jahr	Ankuenfte	Naechtigungen
AT31	Linz	2000	340880	579683
AT31	Steyr	2000	38726	78644
AT31	Wels	2000	84370	150417
AT31	Altheim	2000	4989	10744
AT31	Aspach	2000	2637	21316
AT31	Auerbach	2000	484	3541
AT31	Braunau a. Inn	2000	15748	33911

Diagram illustrating disambiguation of the 'Linz' entity:

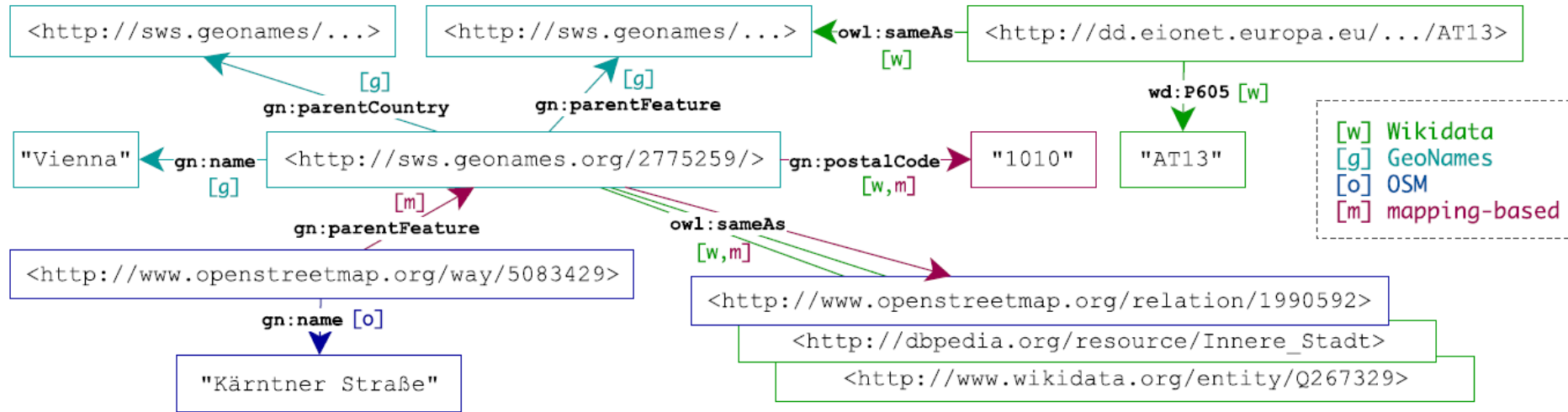
- Upper Austria (Oberösterreich / Haute-Autriche) is the parent of Linz.
- Linz is disambiguated from other potential locations (e.g., Saxony, Germany).

[3] Strötgen, Gertz: Multilingual and Cross-domain Temporal Tagging. Language Resources and Evaluation, 2013.

Indexed Datasets

<u>portal</u>	<u>datasets</u>	<u>CSVs</u>	<u>indexed</u>
<i>total</i>			15728
govdata.de	19464	10006	5646
data.gv.at	20799	18283	2791
offenedaten.de	28372	4961	2530
datos.gob.es	17132	8809	1275
data.gov.ie	6215	1194	884
data.overheid.nl	12283	1603	828
data.gov.uk	44513	7814	594
data.gov.gr	6648	414	496
data.gov.sk	1402	877	384
www.data.gouv.fr	28401	6038	258
opingogn.is	54	49	41

RDF Export 1/2: Knowledge Graph

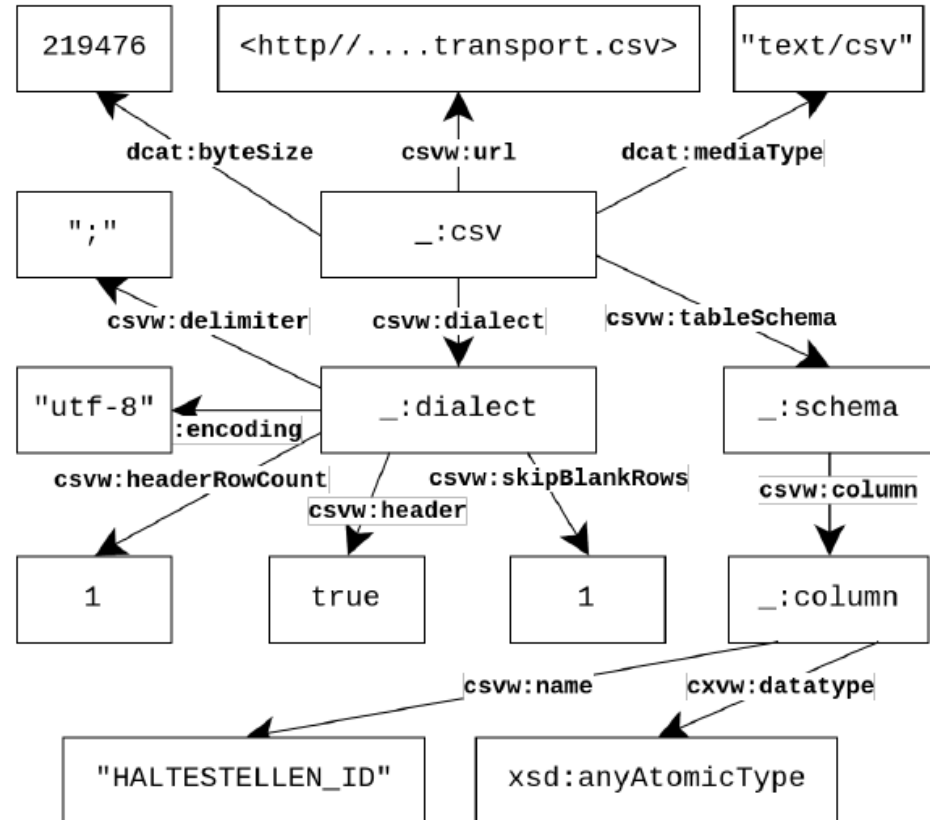


- Spatial and temporal base knowledge graph
- Annotated data points in metadata and CSV cells
- CSV metadata using CSVW vocabulary
 - e.g., delimiter, encoding, header, ...

RDF Export 2/2: CSV on the Web Metadata [4]

- Note: no real cell level annotations, we needed to add those!
- E.g.:
 - **csvwx:cell**
 - **csvwx:hasTime**
 - **csvw:refersToEntity**
 - ...

Details: cf.:
<http://data.wu.ac.at/ns/csvwx>



SPARQL Endpoint (1)

- Find datasets within time-range and referring to geospatial entity:

```
SELECT ?d ?url WHERE {  
  # select the dates of the past two election in Austria  
  wd:Q1386143 timex:hasStartTime ?t1 .  
  wd:Q19311231 timex:hasStartTime ?t2 .  
  
  # select the min and max date values of a dataset  
  ?d dcat:distribution [  
    dcat:accessURL ?url ;  
    timex:hasStartTime ?start ;  
    timex:hasEndTime ?end  
  ] .  
  # select only datasets about Vienna  
  ?d csvwx:refersToEntity <http://sws.geonames.org/2761369/> .  
  
  FILTER((?start >= ?t1) && (?end <= ?t2))  
}
```

SPARQL Endpoint (2)

- Text search for a time period and its temporal and spatial coverage
- Query for cells within time period and referring to geo-entity

```
SELECT ?d ?url ?rownum WHERE {
  # get the "Anschluss movement"
  ?p rdfs:label ?L.
  FILTER (CONTAINS(?L, "Anschluss movement") ) .
  ?p timex:hasStartTime ?start ; timex:hasEndTime ?end ; dcterms:spatial ?sp .
  # find the GeoNames entities
  ?spatial owl:sameAs ?sp .
  ?d dcat:distribution [ dcat:accessURL ?url ] .
  [] csvw:url ?url ; csvw:tableSchema ?s .
  # find a cell where date falls in the range of the found period
  ?s csvw:column ?col1 .
  ?col1 csvwx:cell [
    csvw:rownum ?rownum ;
    csvwx:hasTime ?cTime
  ]
  FILTER((?cTime >= ?start) && (?cTime <= ?end))
  # find another cell in the same row where the geo-entity has the
  # spatial coverage area of the found period as the parent country
  ?s csvw:column ?col2 .
  ?col2 csvwx:cell [
    csvw:rownum ?rownum ;
    csvwx:refersToEntity [ gn:parentCountry ?spatial ]
  ]
}
```

GeoSPARQL Queries

- Standard for representation and querying of geospatial linked data
- (Almost) no complete implementations of GeoSPARQL

```
SELECT ?d ?url ?rownum WHERE {  
  # get the geometry of the Viennese district "Leopoldstadt"  
  <http://sws.geonames.org/2772614/> geosparql:hasGeometry ?polygon .  
  
  ?d dcat:distribution [ dcat:accessURL ?url ] .  
  [ csvw:url ?url ; csvw:tableSchema ?s ].  
  # select the geometries of any annotated cells  
  ?s csvw:column ?col .  
  ?col csvwx:cell [ csvw:rownum ?rownum ; csvwx:refersToEntity [ geosparql:hasGeometry ?g ]  
  
  # filter all annotated data points within the polygon of Leopoldstadt  
  FILTER(geof:sfWithin(?g, ?polygon))  
}
```

Search Interface

Faceted query interface:

- Timespan
- Time pattern
- Geo-entities
- Full-text queries

Back end:

- **MongoDB** for efficient key look-ups
- **ElasticSearch** for indexing and full-text queries
- **Virtuoso** as a triple store

▼ Temporal filters

Filter results by timespan: Off Title & description CSV columns

1/2010 1/2020

Filter pattern

Apply Filter

Linz

Republic of Austria > Oberösterreich > Linz Stadt > Linz

Spatial entity or Full-text results

Hotspot - Standorte - [Hotspot Standorte](#) Stadt Linz <http://data.gv.at>

POI's (Points of Interest) für Hotspot (freies, kostenloses WiFi) in der Stadt Linz. Die Koordinaten sind im im EPSG-Codes WGS84 verfügbar.

Nummer	Latitude	Longitude	Name	Kurztext	Start im Jahr	Ende im Jahr	Stadt	Postleitzahl
4007	48,304793	14,299414	Hotspot Linz - Rotes...	Hier ist nur einer v...	2013	0	Linz	4020

Finanzgebarung der Gemeinden in Oberösterreich - [Oö. Gemeinde-Finanzgebarung 2015](#) Land Oberösterreich <http://data.gv.at>

Finanzdaten der 444 oberösterreichischen Gemeinden

Jahr	NUTS2	Gemeindenummer	Gemeindename	Ordentliche Einnahme...	Ordentliche Ausgaben	Außero Einnahr
2015	AT31	40101	Linz	628704196,3	718773006,9	131859

- Open (Structured) Data is a rich source of Knowledge worthwhile to tap into
- Most of it is not (yet) Linked Data.

What we did:

- Hierarchical knowledge graph of spatial and temporal entities
- Algorithms to annotate CSV tables and their metadata descriptions

What's next:

- Enable GeoSPARQL (or an alternative geospatial-query language)
- Parsing coordinates in datasets
- Parse other file formats, e.g., XML, PDF, ...
- Test other domains such as tweets or web pages (e.g., newspaper articles)

Other Ongoing Projects (data.wu.ac.at)



Projects

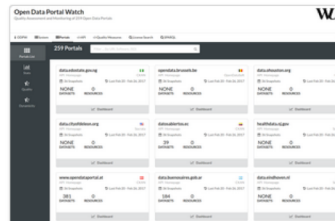


WU Open Data Portal

WU lectures, rooms and organizations

data.wu.ac.at is an Open Data portal where you can find data about lectures, rooms and organizations at WU.

121 datasets

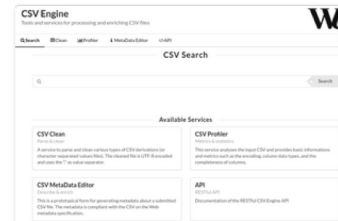


Open Data Portal Watch

Monitoring & exposing portals' metadata

Open Data Portal Watch assesses the evolution of the (meta) data quality of about 260 Open Data portals over since September 2014.

259 portals



CSV Engine

Search & enrich CSVs

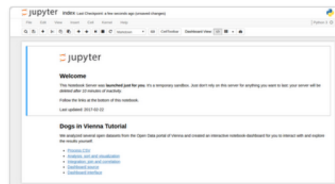
The CSV Engine is a collection of tools and services for processing and enriching CSV files.



DBpedia Wayback Machine

Extract past DBpedia versions

The DBpedia Wayback Machine aims at providing the wayback functionality for DBpedia based on the revisions of their Wikipedia article.

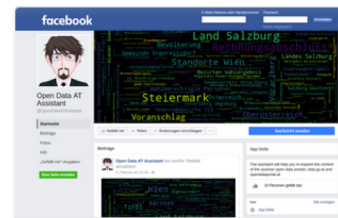


Jupyter Notebook Server

Programming & Documentation

Notebook documents are documents which contain both computer code (e.g. python) and human-readable rich text elements.

<> Only available within local WU Vienna network



Open Data AT Assistant

Search chatbot for Austrian datasets

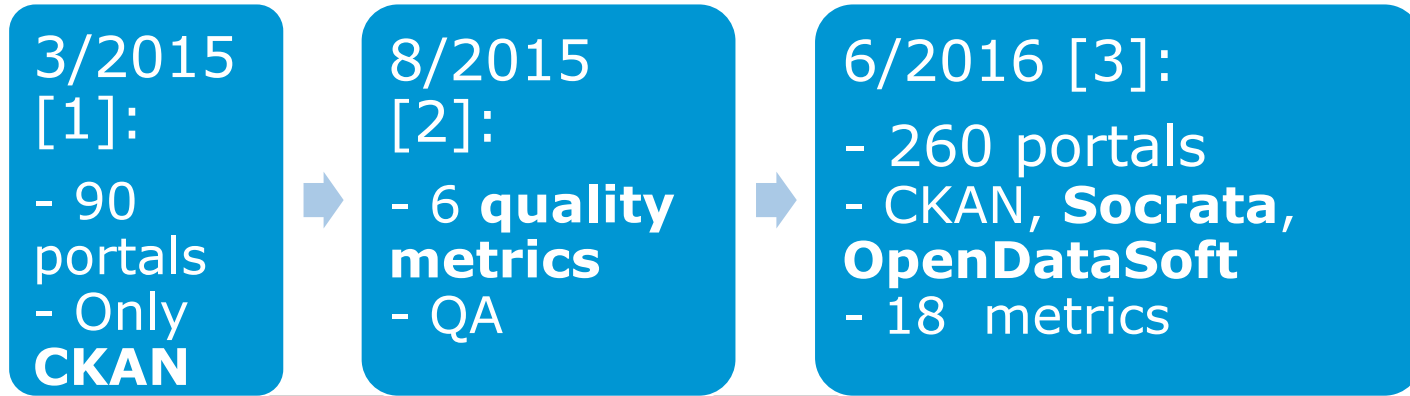
The assistant will help you to explore the content of the austrian open data portals: data.gv.at and opendataportal.at.

f

What else are we working on?

- Open Data Portalwatch
 - 1) Monitoring Metadata quality
 - 2) Mapping to standard vocabularies
 - 3) Enriching Metadata to improve search (*talked about that already*)

1) Monitoring and QA over evolving data portals

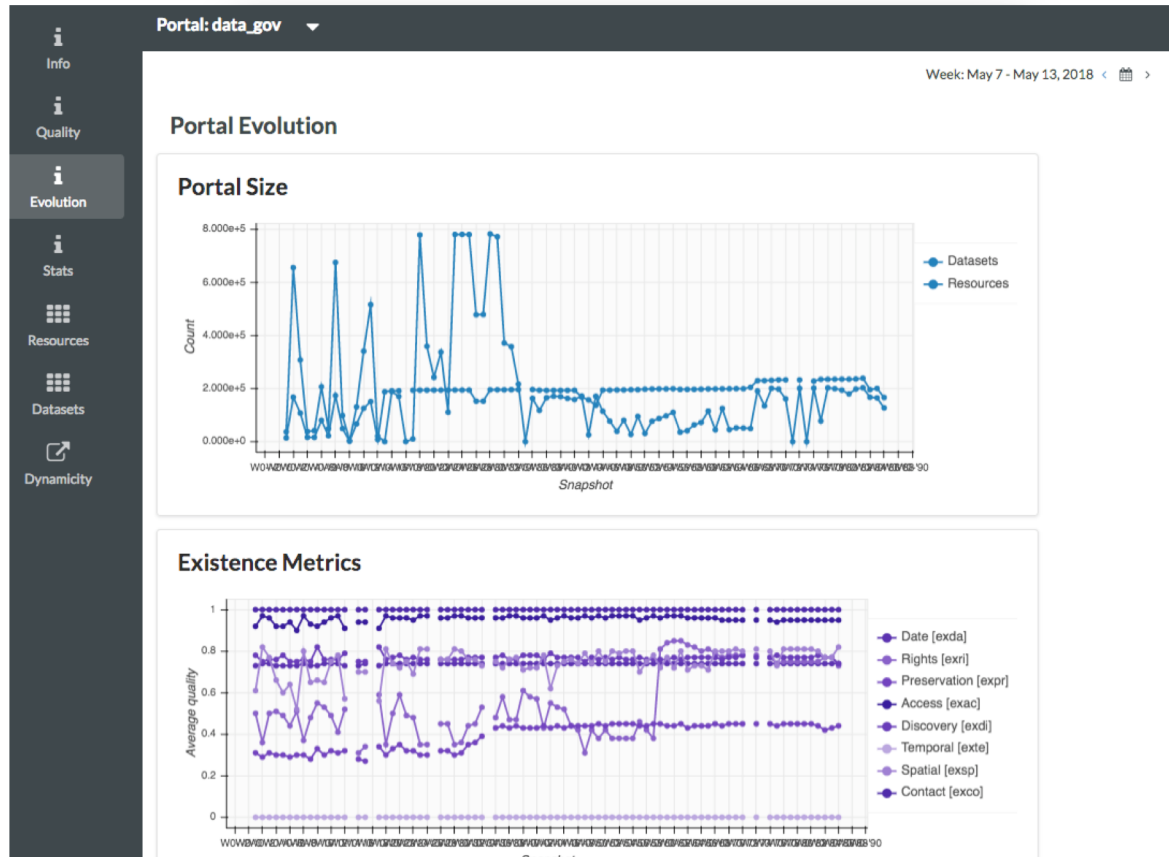


	total	CKAN	Socrata	ODSoft	DCAT
portals	261	149	99	11	2
datasets	854,013	767,364	81,268	3,340	2,041
URLs	2,057,924	1,964,971	104,298	12,398	6,092

- [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)

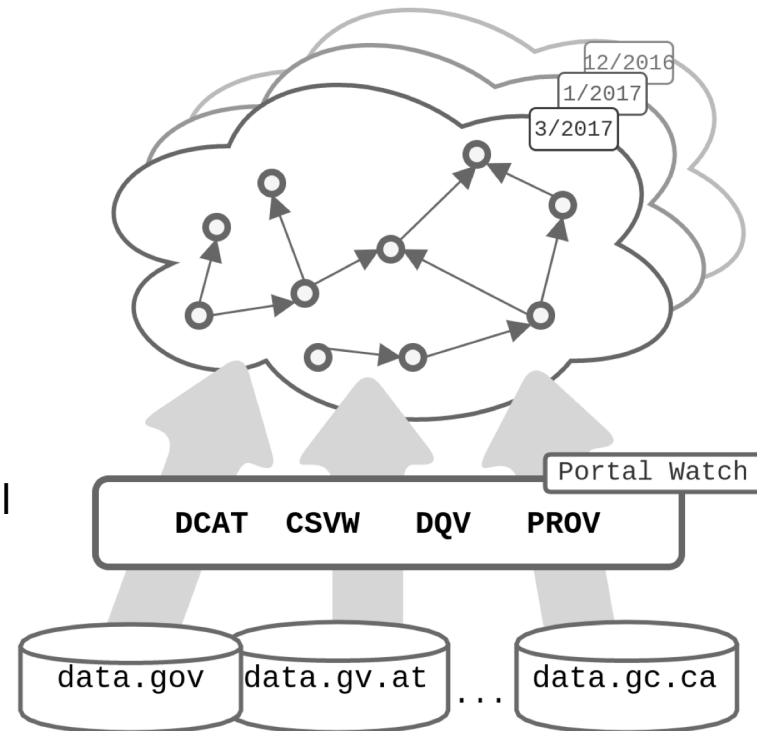
Demo:

http://data.wu.ac.at/portalwatch/portal/data_gov/1818



2) Mapping to Standard vocabularies & Linked Data

- Mapping & Heuristic Enrichment
 - DCAT
 - PROV
 - CSVW
 - Schema.org
- Enable uniform access:
 - SPARQL endpoint
 - Linked Data & Memento Protocol



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

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

Backup Slides

Spatio-temporal labelling – Evaluation:

Total numbers of spatial and temporal annotations of metadata descriptions and columns:

<u>Columns</u>	<i>Spatial</i> <u>Metadata</u>	<i>Temporal</i> <u>Columns</u>	<u>Metadata</u>
3518	11231	2822	9112

10 random CSV datasets per portal (11 portals), 10 random rows per dataset:

- In total inspected 101 datasets 1010 rows
- 87 Correctly assigned labels at the dataset level
- 37 CSV datasets that contain potentially missing annotations (e.g. text that would need to be parsed first, or malformed CSVs, etc.)
- 9 Incorrect links to GeoNames
- 9 Incorrect links to OSM