

Toward Natural Data Understanding



Axel Polleres

What is "Natural Data"?

"Natural Data" = "Data as it occurs in the wild"

- Heterogenous data assets
- From different sources and origins
- Different Formats
- Different Semantics
- Sparse descriptive metadata
- Raw/not necessarily for human consumption

→ Hard to search and integrate!

Where does "Natural Data" occur?

Example 1: Open (Government) Data

Open Data is a Global Trend!

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



data.gov.at - Open Data Österreich

Startseite Daten Dokumente Anwendungen Infos News

Open Data Österreich

Suchbegriff

24.818	465	60
Datensätze	Anwendungen	Organisationen

Themen durchsuchen

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

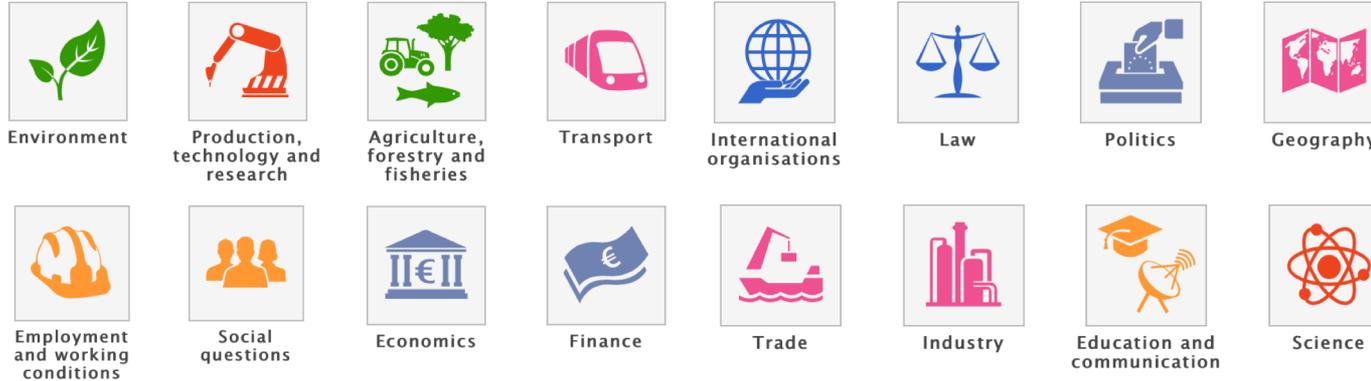
Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED
SEARCH OVER 170,714 DATASETS

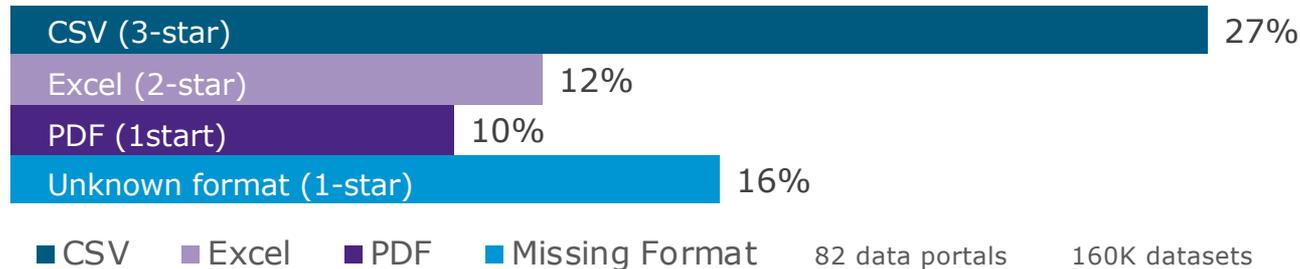
Federal Student Loan Program Data



(Structured) Open Data comes in various ways



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



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...

DATA.GOV DATA TOPICS - IMPACT APPLICATIONS DEVELOPERS CONTACT

DATA CATALOG / Datasets Organizations ?

Home / Department of Housing and ... / US Department of Housing and Urban Development

Housing Affordability Data System (HADS)
Metadata Updated: March 8, 2017

The Housing Affordability Data System (HADS) is a set of files derived from the 1985 and later national American Housing Survey (AHS) and the 2002 and later Metro AHS. This system categorizes housing units by affordability and households by income, with respect to the Adjusted Median Income, Fair Market Rent (FMR), and poverty income. It also includes housing cost burden for owner and renter households. These files have been the basis for the worst case needs tables since 2001. The data files are available for public use, since they were derived from AHS public use files and the published income limits and FMRs. These datasets give the community of housing analysts the opportunity to use a consistent set of affordability measures.

Access & Use Information

- Public:** This dataset is intended for public access and use.
- License:** No license information was provided. If this work was prepared by an officer or employee of the United States government as part of that person's official duties it is considered a U.S. Government Work.

Downloads & Resources

Comma Separated Values File **13730** views
Download

Dates

Metadata Created Date	March 7, 2014
Metadata Updated Date	March 8, 2017

Metadata Source

Data.gov Metadata
Download Metadata

Harvested from HUD JSON

affordability | cost | fmr | households | housing | income | rent | 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

data.gv.at API

Suchbegriff (z.B. Finanzen, Wahlen) Suche starten

Datenkatalog Apps & News Katalog durchstöbern

data.gv.at - offene Daten Österreichs

Startseite Daten Dokumente Anwendungen 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 website with a search bar containing 'Leopoldstadt'. The search results are divided into three sections: 'Suchergebnisse - Daten & Dokumente', 'Suchergebnisse - Anwendungen', and 'Suchergebnisse - News'. The first section shows a result for 'Stadtplan von Anton Behsel 1825' with a date of 29.03.2019. The second and third sections both display 'Keine [Anwendungen/News] gefunden!' (No [applications/news] found!).



Not too much!

Logo for the Open Data Portal Austria (ODP) featuring the letters 'odp' in green and blue, a 'BETA' badge, and the slogan 'All you can Data'. Below the logo, it says 'OPEN DATA PORTAL ÖSTERREICH'. A navigation bar at the bottom includes 'HOME', 'DATEN', 'THEMEN', and 'ANWENDUNGEN'. Below the navigation bar, it says 'OpenDataPortal Österreich > Datenkatalog > Daten'.

Organisationen

Für diese Suche wurden keine Organisationen gefunden.

Gruppen

Für diese Suche wurden keine Gruppen gefunden.

Die Daten in unserem Katalog stehen unter CC BY 3.0 oder CC0 Lizenz frei zur Verfügung. Der Katalog ist unter anderem nach Themenbereichen sortierbar. Die genauen Datenrichtlinien erfahren Sie bei den einzelnen Datensätzen.

Suche: Leopoldstadt

Keine Datensätze gefunden bei der Suche "Leopoldstadt"

Sortieren nach Relevanz

Where does "Natural Data" occur?

Example 2: Large Enterprise Data!?

Probably similar issues:

Handelsblatt 30.10.2006 **Management-Erkenntnisse aus der Bibel**

<https://www.handelsblatt.com/politik/konjunktur/oekonomie/wissenswert/neue-bwl-studie-management-erkenntnisse-aus-der-bibel/2725926.html>

*Nur dann habe die Vision eines Unternehmens, in dem alle Mitarbeiter ihr Wissen teilen, eine Chance. [...] Dass sich das lohnen dürfte, ahnte Heinrich von Pierer schon 1995. „**Wenn Siemens wüsste, was Siemens weiß**“, seufzte der damalige Vorstandschef vor elf Jahren auf der Bilanzpressekonferenz, „**dann wären unsere Zahlen noch besser.**“*

Data Challenges in Siemens?



Call for Ideas: Data Analytics in

(How) can customer data be leveraged for novel, complex analytics, machine learning, and finally decision making?

- **Smart Sales and Customer Analytics:** Customer analytics help optimize sales and marketing by leveraging customer related information, as well as their interactions with different company's functions and IT systems. Customer analytics typically address business questions related to customer churn (e.g. how to increase customer loyalty).
- **Data Privacy and Organizational Aspects:** The increased availability of low-cost data has led to a growing concern about data privacy and security in the past years. The recent GDPR Regulation (GDPR) poses strict limitations and rules on data processing.

(How) can customer data be leveraged in a privacy-aware, legally compliant manner? I.e., how can we enable re-use of "insights" across systems, across projects, etc.)

What is "Natural Data"?

"Natural Data" = "Data as it occurs in the wild"

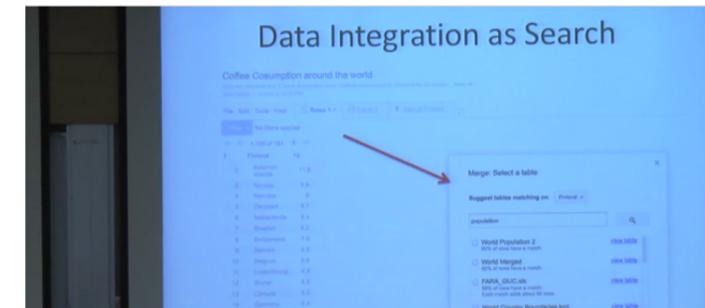
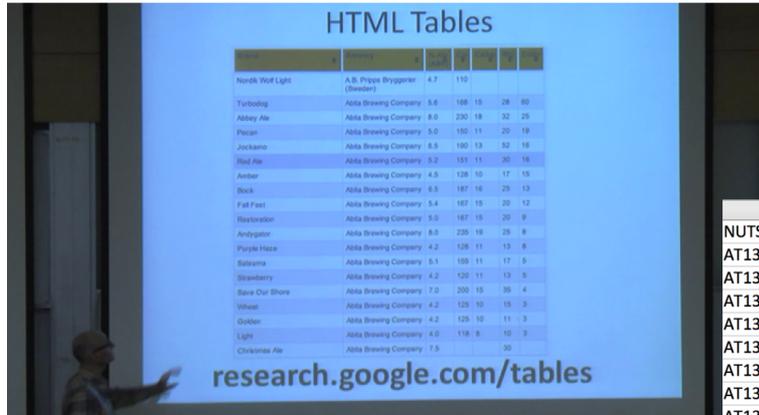
- Heterogenous data assets
- From different sources and origins
- Different Formats
- Different Semantics
- Sparse descriptive metadata
- Raw/not necessarily for human consumption
- Policies forbid to re-share sensitive Data
- Data Catalogs not even existing yet

→ Hard to search and integrate!

Why is Search & Integration of Natural Data a problem?

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

Structured Data in Web Search by Alon Halevy



VS.

Katalog
Bevölkerung in Wien: Bezirk - Geschlecht

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

“Natural Data” Search: How would a human approach this problem?

The screenshot shows the data.gov.at website. The search bar contains 'Leopoldstadt land use'. Below the search bar, there are two tabs: 'Suchergebnisse - Daten & Dokumente' and 'Suchergebnisse - Anwendungen'. The 'Daten & Dokumente' tab is active and shows 'Keine Datensätze gefunden!' (No datasets found!). The 'Anwendungen' tab is also visible and shows 'Keine Anwendungen gefunden!' (No applications found!).

e.g. you wouldn't probably search in data.gov ...



HOME DATEN

OpenDataPortal Österreich

Organisationen

Für diese Suche wurden keine Organisationen gefunden.

Gruppen

Für diese Suche wurden keine Gruppen gefunden.

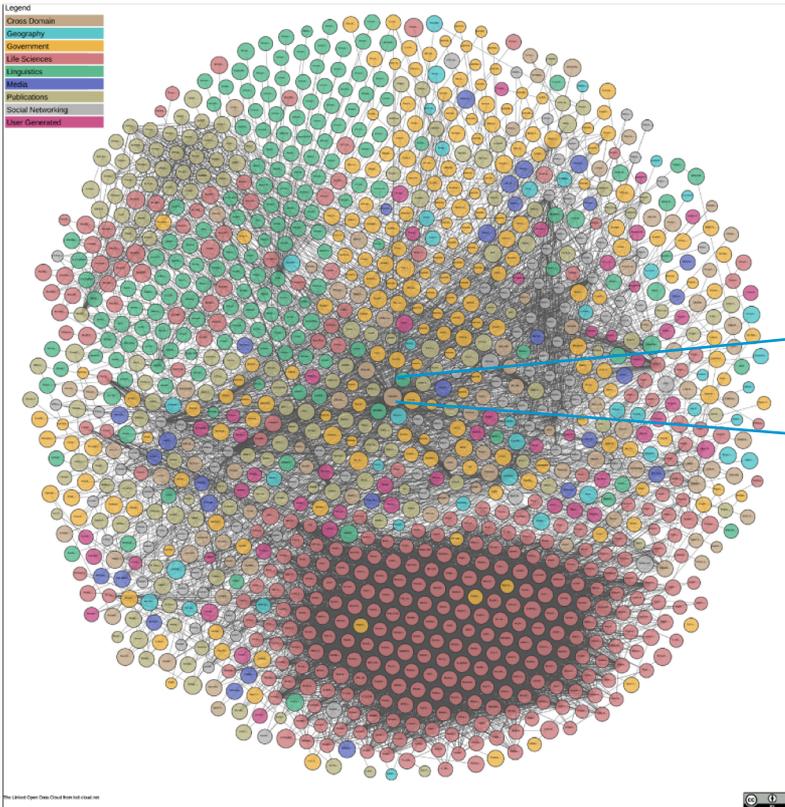
Die Daten in unserem Katalog stehen unter CC BY ist unter anderem nach Themenbereichen sortiert. Die Daten sind in der Liste der einzelnen Datensätzen.

Keine Datensätze gefunden
der Suche "leopoldstadt land use"

The screenshot shows the data.gov website. The search bar contains 'Leopoldstadt land use'. The search results show '48,728 datasets found for "Leopoldstadt land use"'. The first result is '2010 Land Use' by 'City of Austin'. The page also includes a 'Filter by location' section with a search bar for location and a 'Clear' button. The 'Order by' dropdown is set to 'Relevance'.

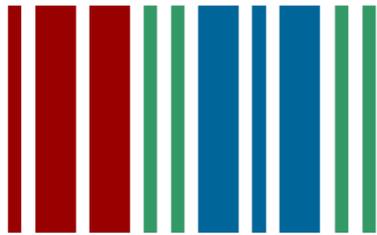


Our research: Knowledge Graphs for Natural Data Search & Integration!



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

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



WIKIDATA

Our research: Knowledge Graphs for Natural Data Search & Integration!

- 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?

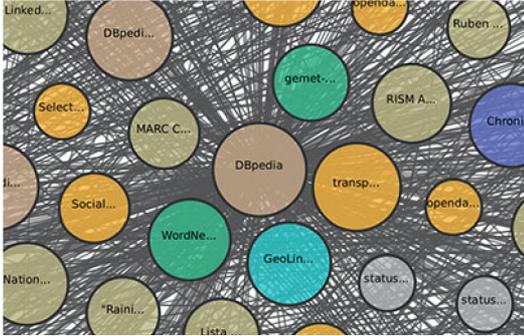
- Identifying the most likely semantic label for a bag of numerical values
- Deliberately ignore surroundings

population (a district) (country Austria)

A blue arrow points from the label box above to the table below.

98157
18763
29730
...

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**

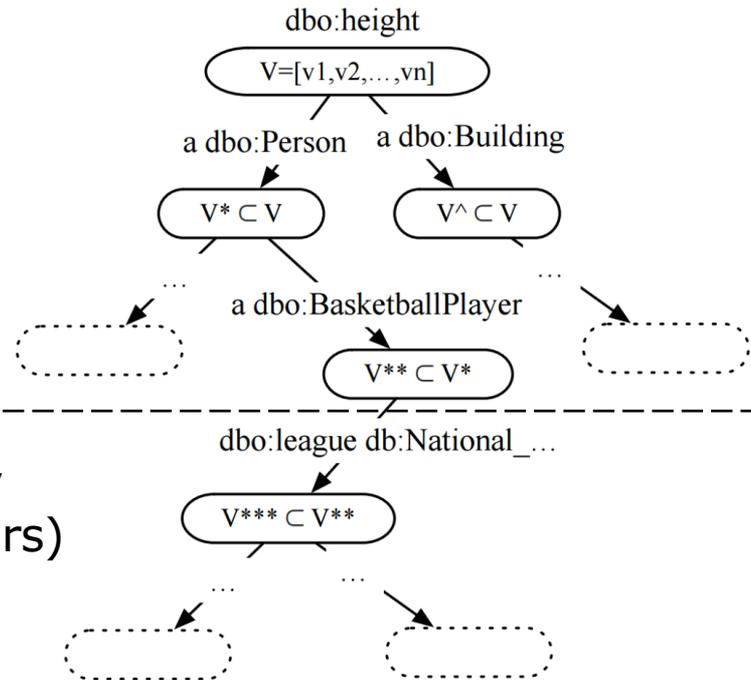
Background Knowledge Graph

- Find properties with **numerical range**
- Hierarchical clustering approach

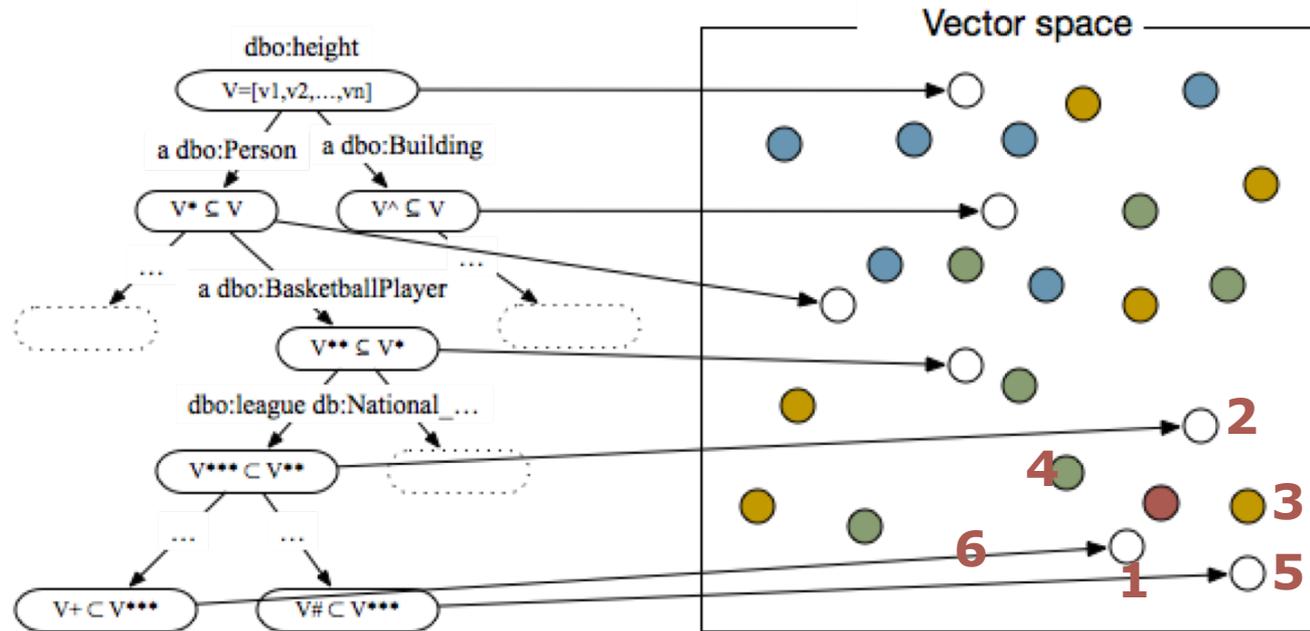
- Two hierarchical layers:

- 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 Background Knowledge Graph**
- *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, maybe 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?

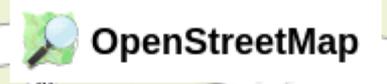
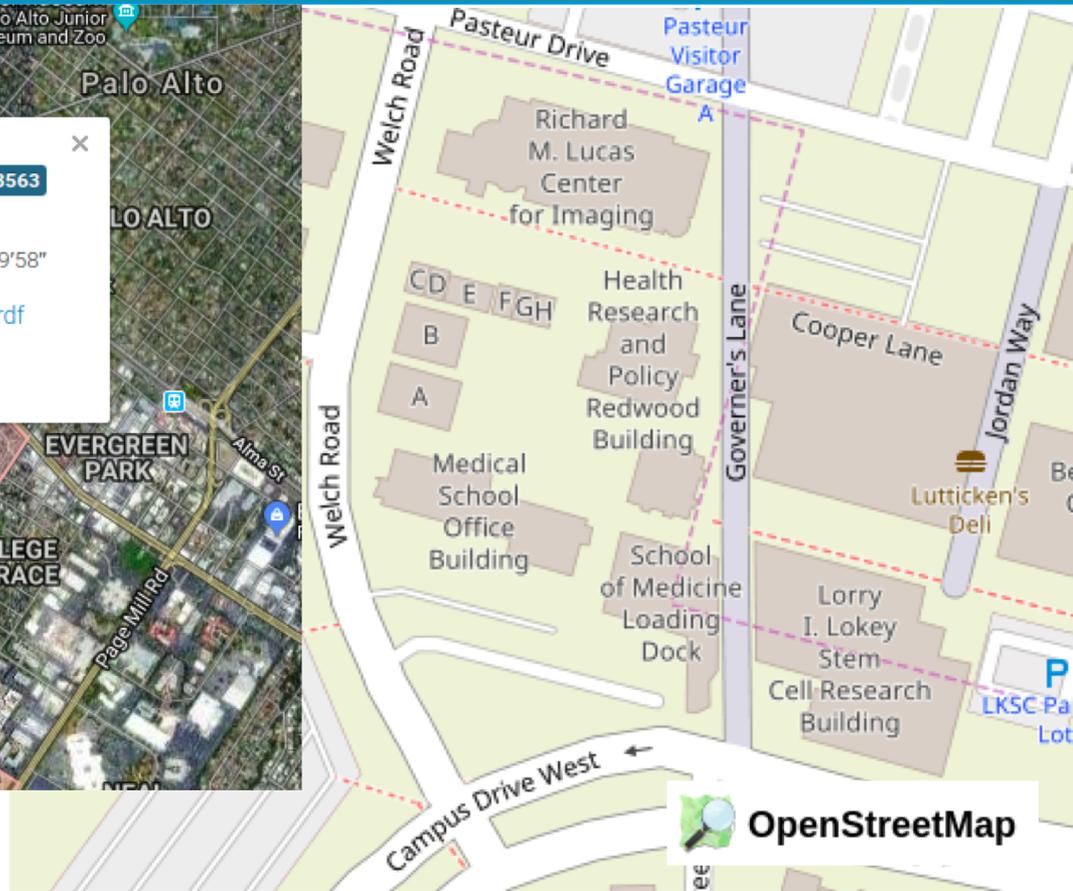
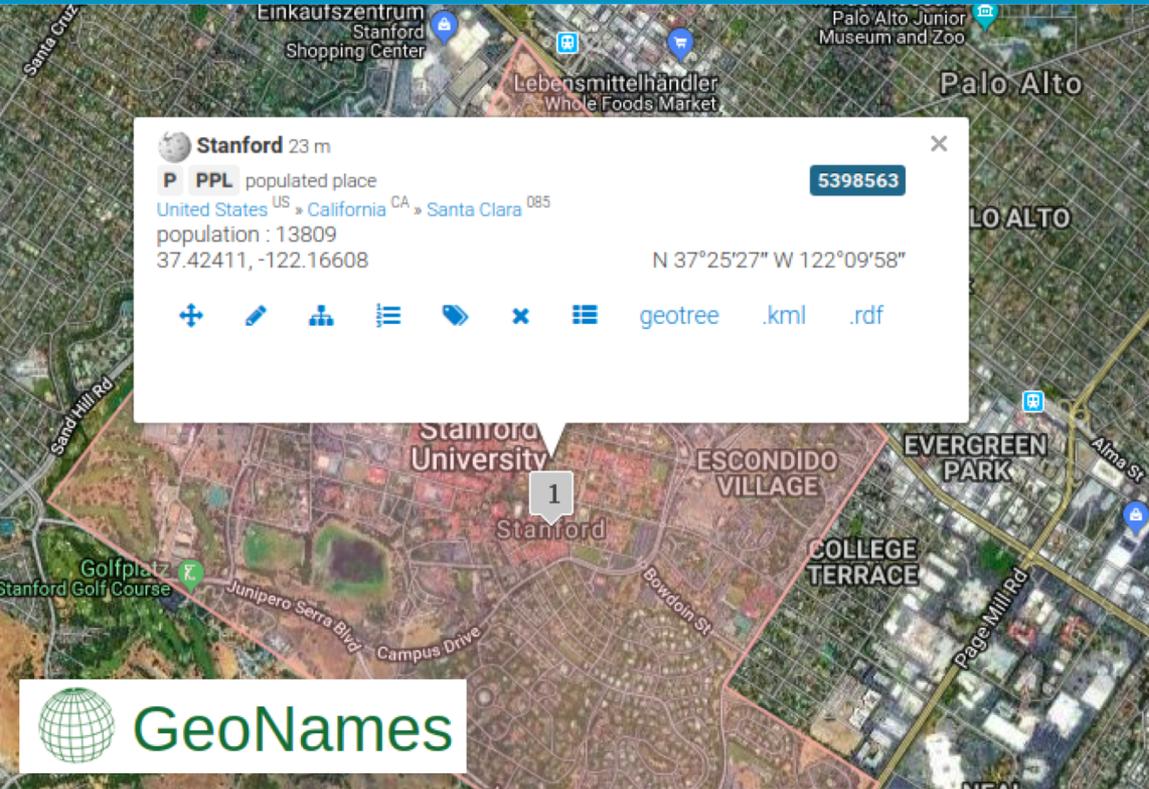
Focus on specific dimensions:

- 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



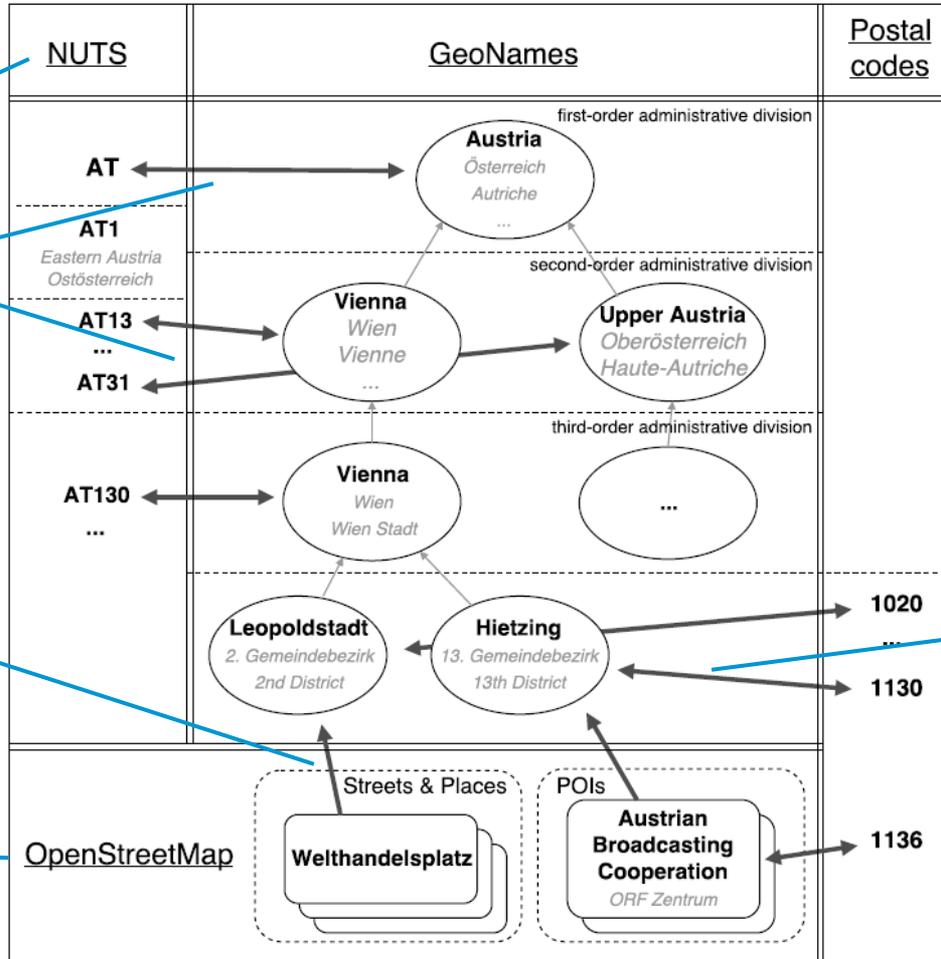
Geo-Knowledge Graph Construction

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


Wikidata Query

Beispiele
Hilfe
Werkzeuge

```

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 ]
    
```



3 Ergebnisse in 193 ms
Code
Herunterladen

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

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



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 => Erhebungsgemeinde; Jahr => Kalenderjahr; Erhebungsgemeinde lt. Tourismus-Statistik-Verordnung 2002 §2 Abs.7: Städte und Gemeinden mit mehr als 1.000 Gästenächtigungen im Kalenderjahr

CSV

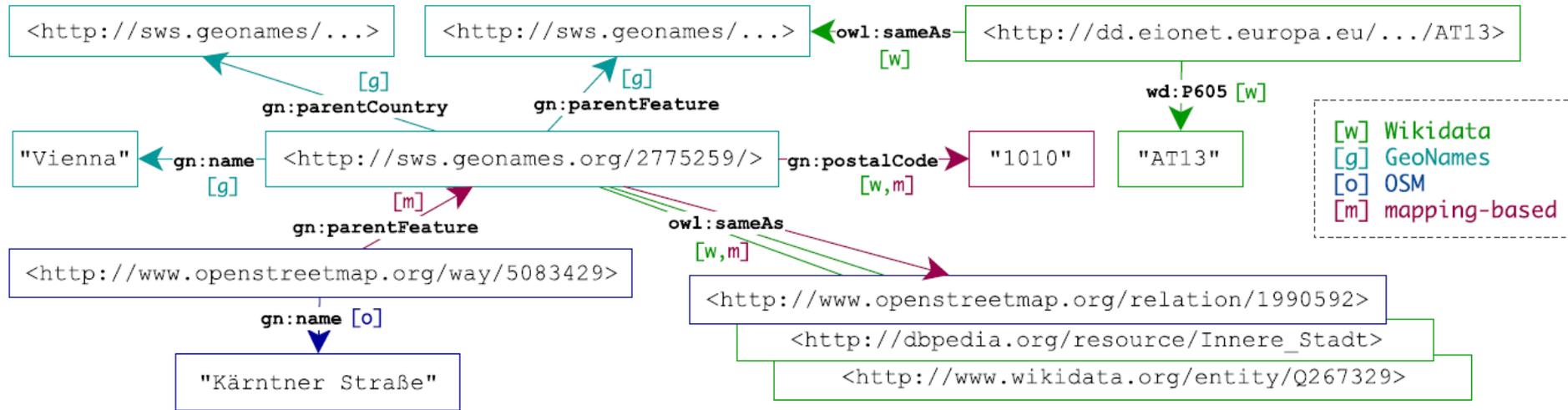
NUTS2	Gemeindefürname	Jahr	Ankuenfte	Naechtigungen
AT31	Linz	2000	340880	579683
AT31	Steyr	2000	38726	78644
AT31	Wels	2000	84370	150417
AT31	Altheim	2000	4989	10744
AT31	Aspach	2000	2637	21316
AT31	Auerbach	2000	484	3541
AT31	Braunau a. Inn	2000	15748	33911

Diagram illustrating disambiguation of the 'Linz' entity:

- Upper Austria (Oberösterreich / Haute-Autriche)
- Linz
- Saxony
- Germany

Disambiguate

RDF Export 1/2: Knowledge Graph

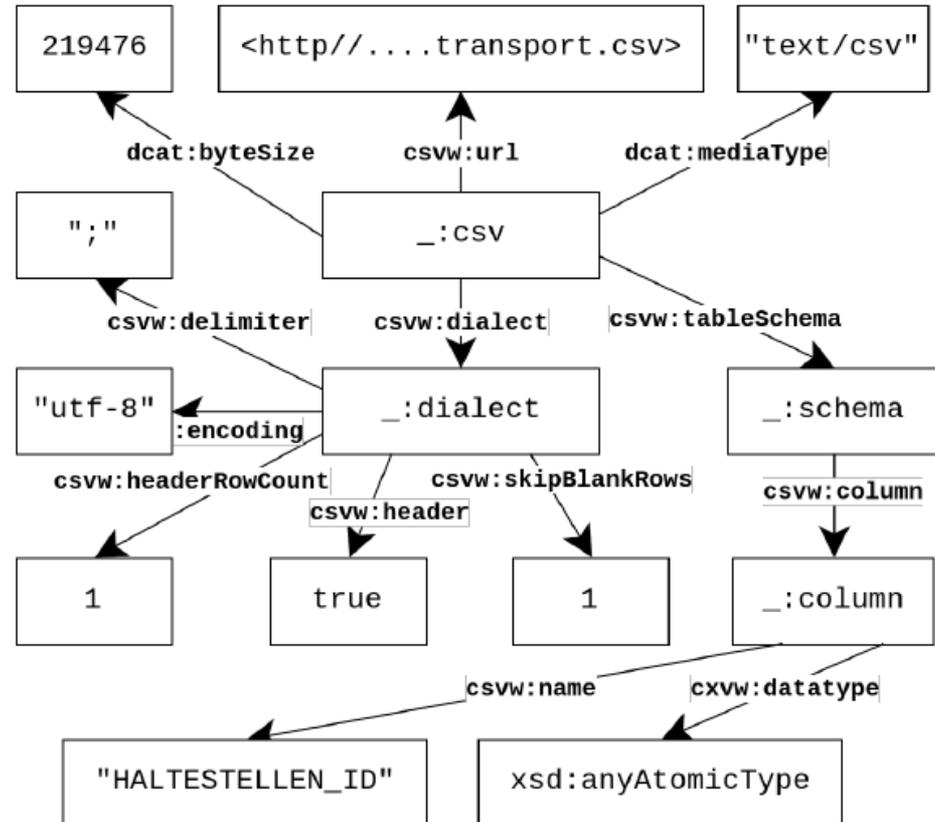


- Spatial and temporal base knowledge graph
- Annotated data points in metadata and CSV cells
- CSV metadata using CSVW vocabulary
 - e.g., delimiter, encoding, header, ...

RDF Export 2/2: Annotate Datasets → 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>



Enable e.g. Search by 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](#)
POI's (Points of Interest) für Hotspot (freies, kostenloses WiFi) in der Stadt Linz. Die Koordinaten sind im im EPSG-Codes WGS84 verfügbar. <http://data.gv.at>

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](#)
Finanzdaten der 444 oberösterreichischen Gemeinden <http://data.gv.at>

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

How well does it work? 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

Lessons learned

- Geospatial and Temporal scope is the most useful search feature for Open Data
 - Respective Hierarchical Knowledge Graphs can be built from existing Linked Data Sources
 - Our algorithms annotate CSV tables **and** their metadata descriptions
- KGs improve search (with some extra work)
- Main Problem still persists: coverage of openly available KGs for Open Data



First Look Journal of Web Semantics

Enabling Spatio-Temporal Search in Open Data

JWS: Information Retrieval

23 Pages • Posted: 20 Dec 2018 • First Look: Accepted

[Sebastian Neumaier](#)

Vienna University of Economics and Business; Vienna University of Technology

[Axel Polleres](#)

Vienna University of Economics and Business; Complexity Science Hub Vienna; Stanford University

Abstract

Intuitively, most datasets found on governmental Open Data portals are organized by spatio-temporal criteria, that is, single datasets provide data for a certain region, valid for a certain time period. Likewise, for many use cases (such as, for instance, data journalism and fact checking) a pre-dominant need is to scope down the relevant datasets to a particular period or region. Rich spatio-temporal annotations are therefore a crucial need to enable semantic search for (and across) Open Data portals along those dimensions, yet - to the best of our knowledge - no working solution exists. To this end, we (i) present a scalable approach to construct a spatio-temporal knowledge graph that hierarchically structures geographical as well as temporal entities, (ii) annotate a large corpus of tabular datasets from open data portals with entities from this knowledge graph, and (iii) enable structured, spatio-temporal search and querying over Open Data catalogs, both via a search interface as well as via a SPARQL endpoint, available at data.wu.ac.at/odgraphsearch/.

Keywords: open data, spatio-temporal labelling, spatio-temporal knowledge graph

Suggested Citation:

Neumaier, Sebastian and Polleres, Axel, Enabling Spatio-Temporal Search in Open Data (December 20, 2018). Available at SSRN: <https://ssrn.com/abstract=3304721> or <http://dx.doi.org/10.2139/ssrn.3304721>

Coverage of openly available KGs for Open Data - example

Item [Discussion](#) [Read](#) [View history](#)

land use (Q1165944)

total of arrangements, activities, and inputs that people undertake in a certain usage of lands

[In more languages](#) [Configure](#)

Language	Label	Description
English	land use	total of arrangements that people undertake type
German	Flächenverbrauch	Art der Nutzung von
Bavarian	No label defined	No description defined
French	Utilisation du sol	No description defined

[All entered languages](#)

Statements

subclass of [spatial planning](#)

[0 references](#)

land use
total of arrangements, activities, and inputs that people un...

Land Use Policy
journal

General plan (*land-use planning*)
regulation for fair and sustainable use of land, largely to pr...

Land use statistics by country
Wikimedia list article

land use map
visual representation of the use of a unit of land at a partic...

Land-use conflict

Land Use in Australia : Past, Present and Future
non-fiction book

more

containing...
land use

occupation du sol
utilisation des sols
occupation des sols

- We need better ontologies...

What's next?

- Which ontologies? What are other **core dimensions** to build knowledge graphs for?
- *How to build Knowledge Graphs from the data itself?*
- *Which Background knowledge do human experts use to understand data?*
 - → *May need qualitative research!*
- *How to combine both scalable data integration and machine learning?*
 - → *How to efficiently find tables that can be joined?*
 - → *How to find patterns in the data?*
- *Can we adopt Research from Natural Language Processing?*

- *How can we enable search, integration and analytics on sensitive Data?*
 - *respecting Data Usage Policies?*
- +
- *developing novel Data Anonymization/Synthetization techniques?*

Which Background knowledge do human experts use to understand data?

A **human** looking at an arbitrary **table** is able to...

- recognize districts of Vienna, although term “Vienna” not mentioned
- understand that the sum of the area sizes relates to Vienna
- understand the different land use (sub-)properties

land use in ha, 2018							
district	total	building land			green	water	transport
		total	residential	public			
Innere Stadt	286,9	141,7	65,8	29,7	27,3	3,1	114,9
Leopoldstadt	1.924,2	437,8	278,7	125,1	674,5	410	402
Landstraße	739,8	412,1	215,9	81,6	110,7	0,5	216,5
Wieden	177,5	114,38	94,88	12,79	17,73	-	45,43
Margareten

Which Background knowledge do human experts use to understand data?

- compare with other datasets, e.g., US cities and their land use, even if it's provided in different units such as square miles
- combine with other related datasets:

NUTS3	DIST	USAGE_CODE	AREA
AT130	11	5	31479,12
AT130	11	27	180207,18
AT130	11	1	5214,41
...

Use of non-human-readable labels and geocode standards (see NUTS identifier) in public datasets. Also, apparently the numbers in the dataset use a different scale than Table 1.²

What's next?



- Which ontologies?
 - What are other **core dimensions** to build knowledge graphs for?
 - *How to build Knowledge Graphs from the data itself?*
 - *Which Background knowledge do human experts use to understand data?*
 - → *May need qualitative research!*
 - *How to combine both scalable data integration and machine learning?*
 - → *How to efficiently find tables that can be joined?*
 - → *How to find patterns in the data? **How can we learn/train patterns that humans "see"?***
 - *Can we adopt Research from Natural Language Processing?*
 - ***How can we use embeddings, "n-Grams" for structured data?***
 - *How can we enable search, integration and analytics on sensitive Data?*
 - *respecting Data Usage Policies?*
- +
- *developing novel Data Anonymization/Synthetization techniques?*





Backup slides:

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



Projects

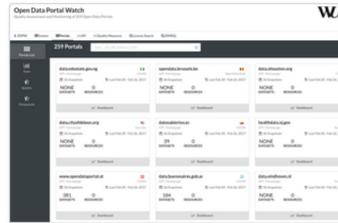


WU Open Data Portal

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

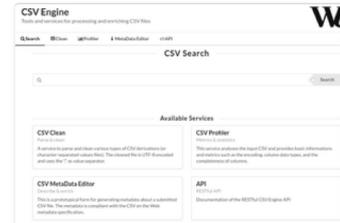


Open Data Portal Watch

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

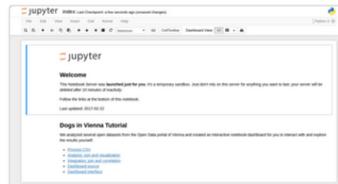
The CSV Engine is a collection of tools and services for processing and enriching CSV files.



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

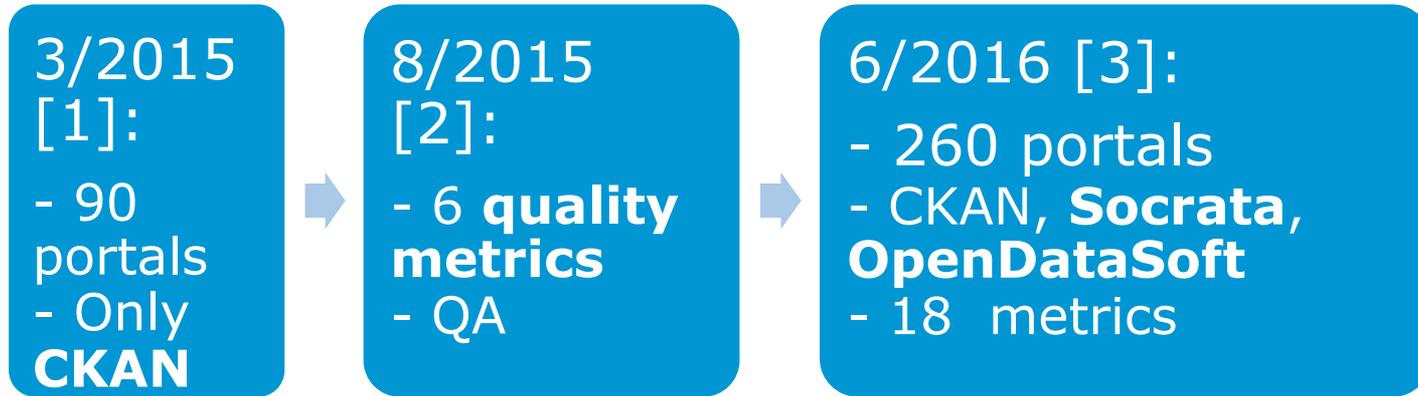
The assistant will help you to explore the content of the austrian open data portals: data.gvat and opendataportal.at.

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

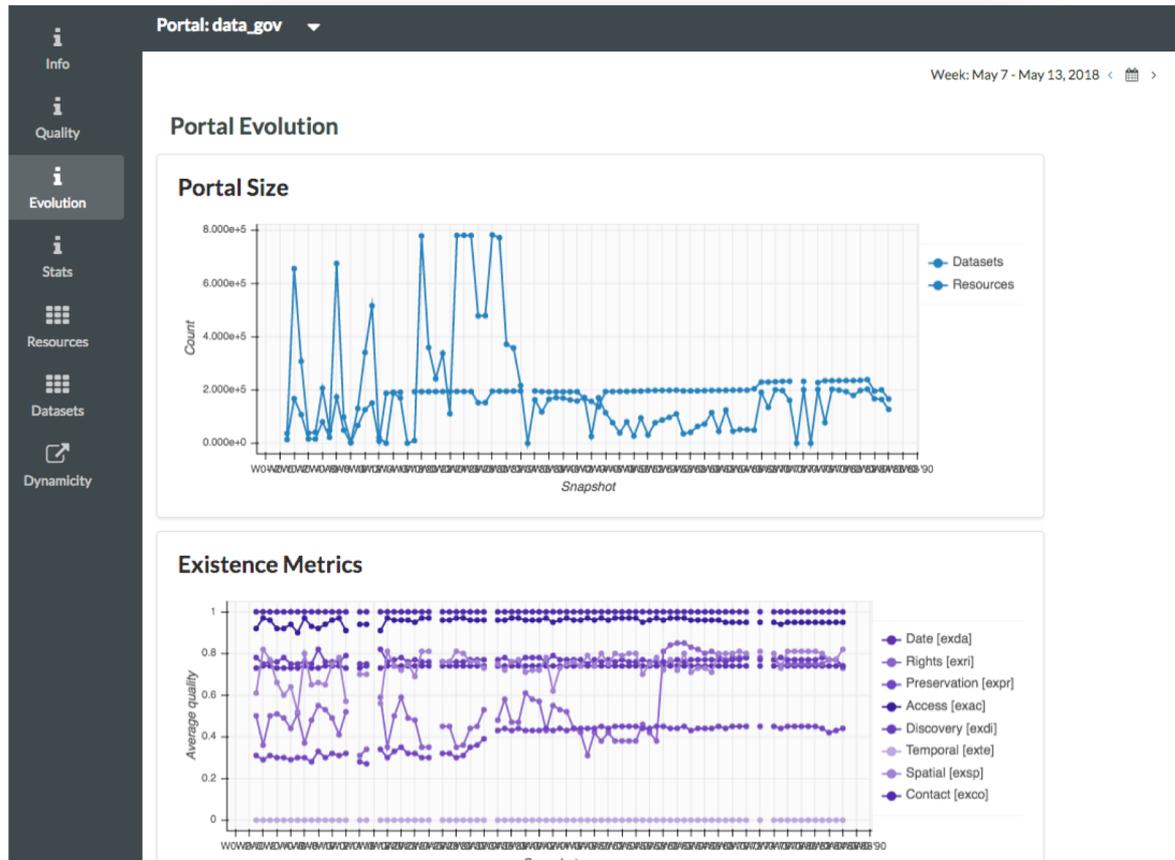


	total	CKAN	Socrata	ODSoft	DCAT
portals	261	149	99	11	2
datasets	854,013	767,364	81,268	3,340	2,041
URLs	2,057,924	1,964,971	104,298	12,398	6,092

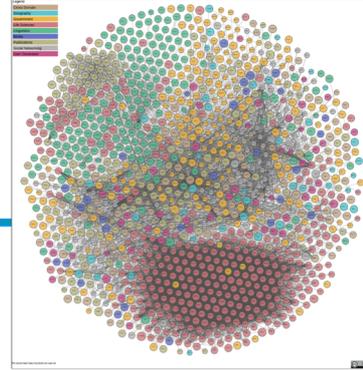
- [1] Towards assessing the quality evolution of open data portals. In ODQ2015: Open Data Quality Workshop, Munich, Germany
- [2] Quality assessment & evolution of open data portals. In: International Conference on Open and Big Data, Rome, Italy (2015)
- [3] Automated quality assessment of metadata across open data portals. ACM Journal of Data and Information Quality (2016)

Demo:

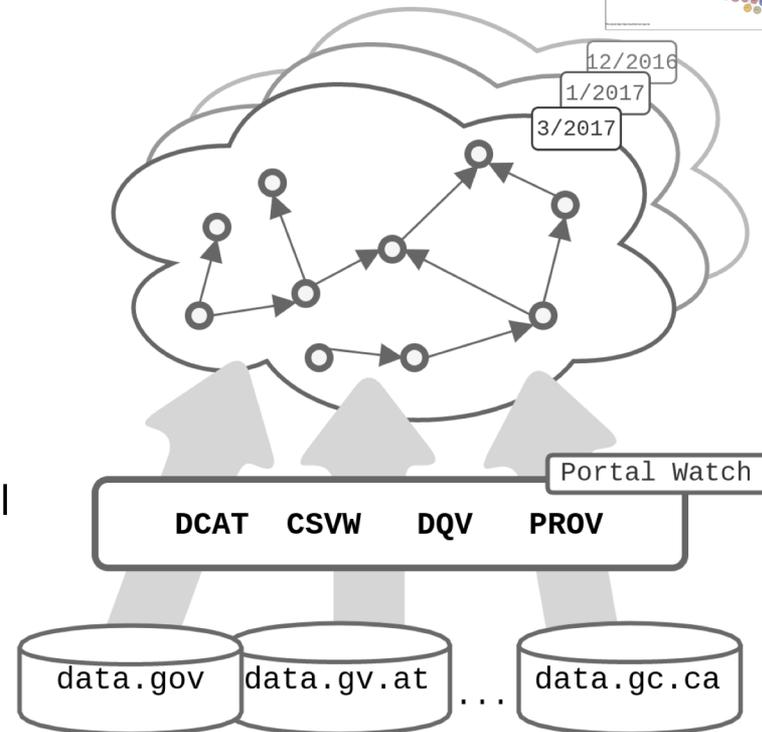
http://data.wu.ac.at/portalwatch/portal/data_gov/1818



2) Mapping to Standard vocabularies & Linked Data



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

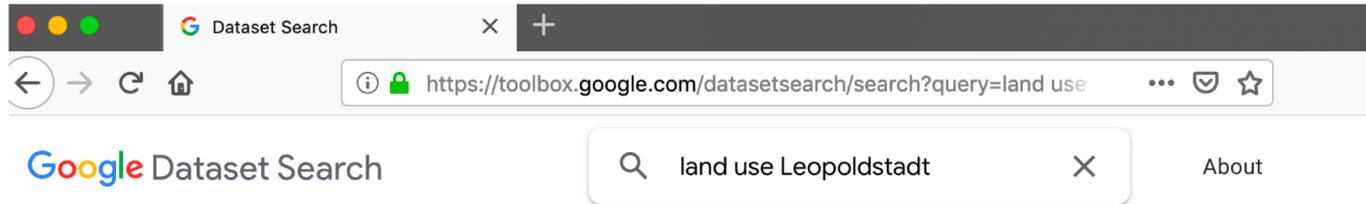


[1] <http://data.wu.ac.at/portalwatch/sparql>

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

Google Dataset Search

<https://toolbox.google.com/datasetsearch/>



Your search - **land use Leopoldstadt** - did not match any datasets.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

[Learn](#) how you can add new datasets to our index.