Toward Natural Data Understanding

Axel Polleres

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

OPENDATA data.wu.ac.at

Institute for Information Business. data.wu.ac.at

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!

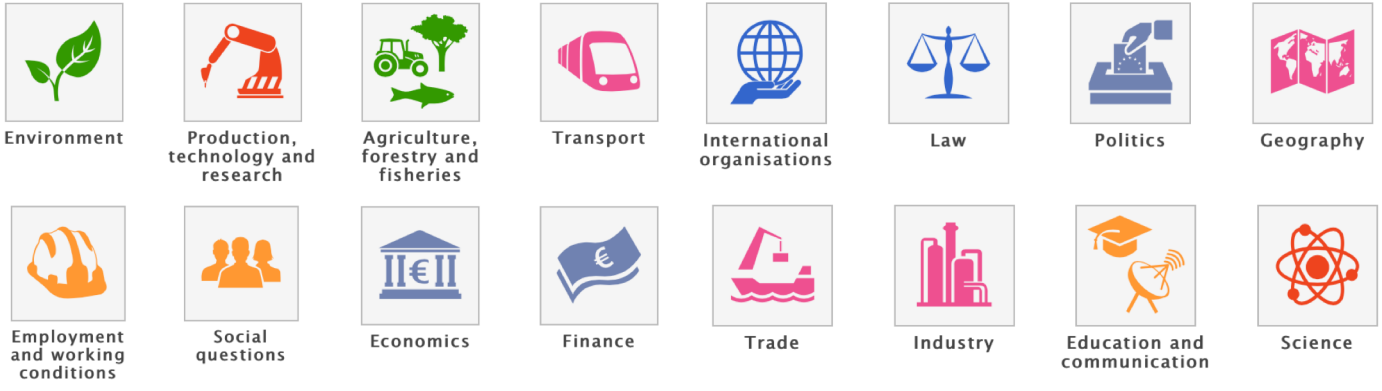
Where does ”Natural Data” occur?

Example 1: Open (Government) Data

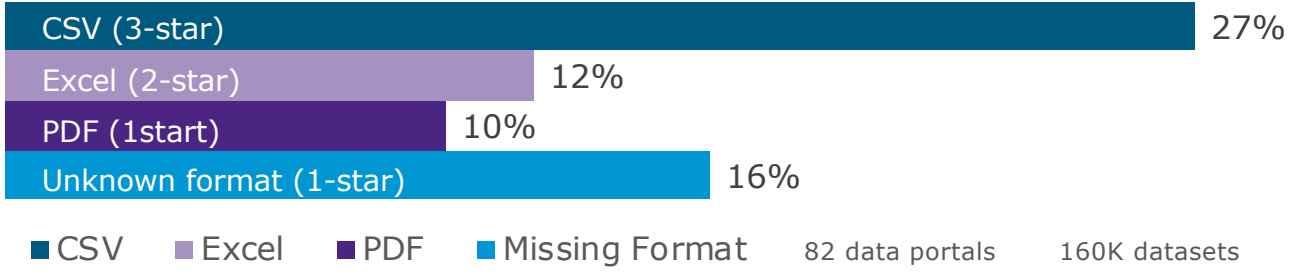
Open Data is a Global Trend!

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

(Structured) Open Data comes in various ways



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



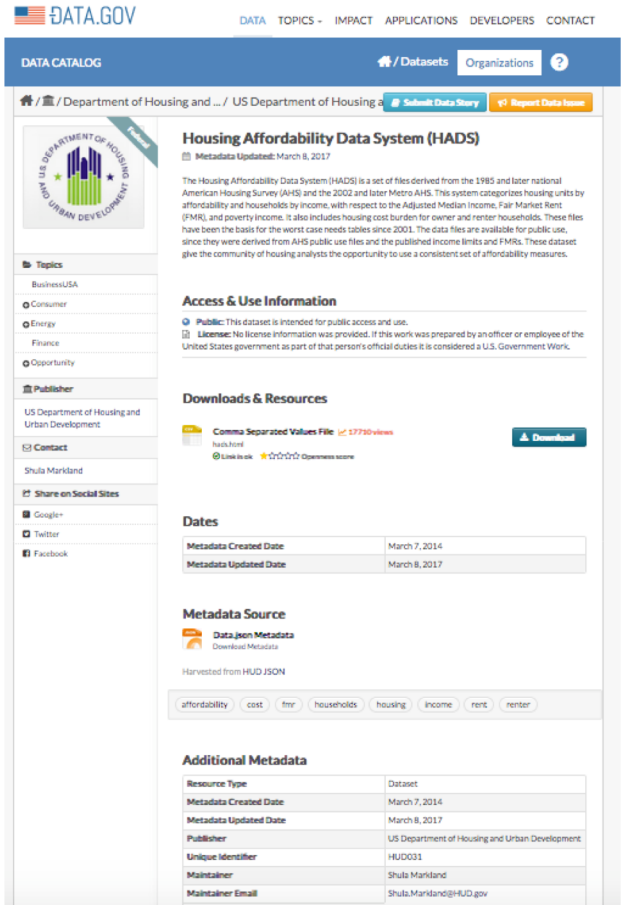
[1] Umbrich, J., Neumaier, S., Polleres, A.: Quality assessment & evolution of open data portals. International Conference on Open and Big Data (2015)

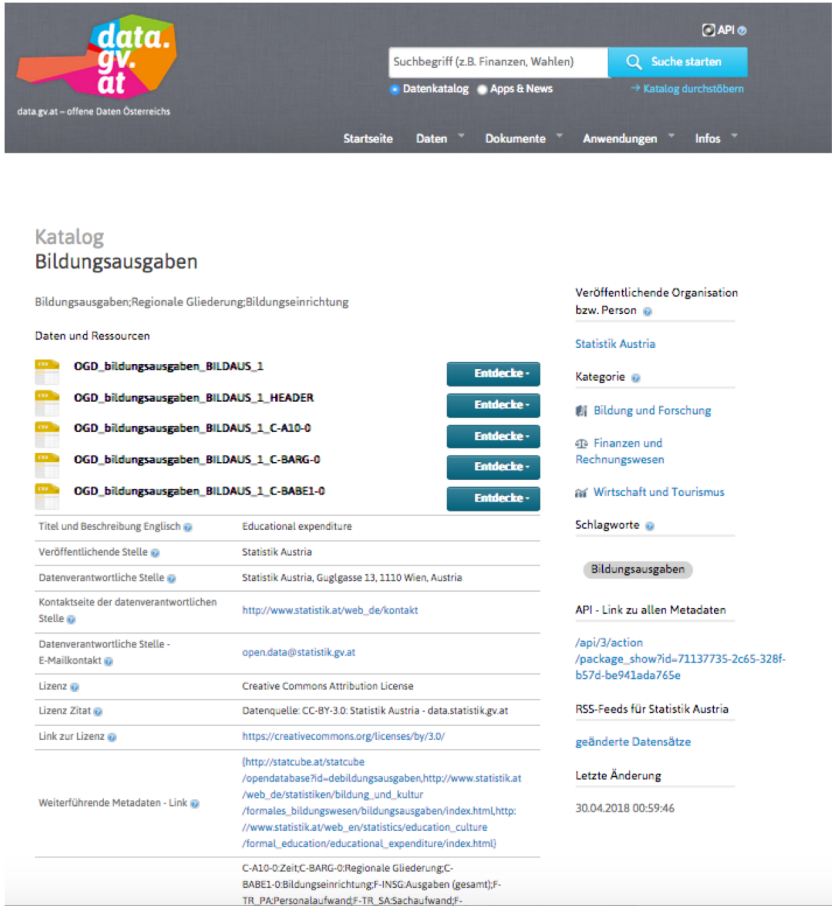
Open Data as a Global Trend:

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

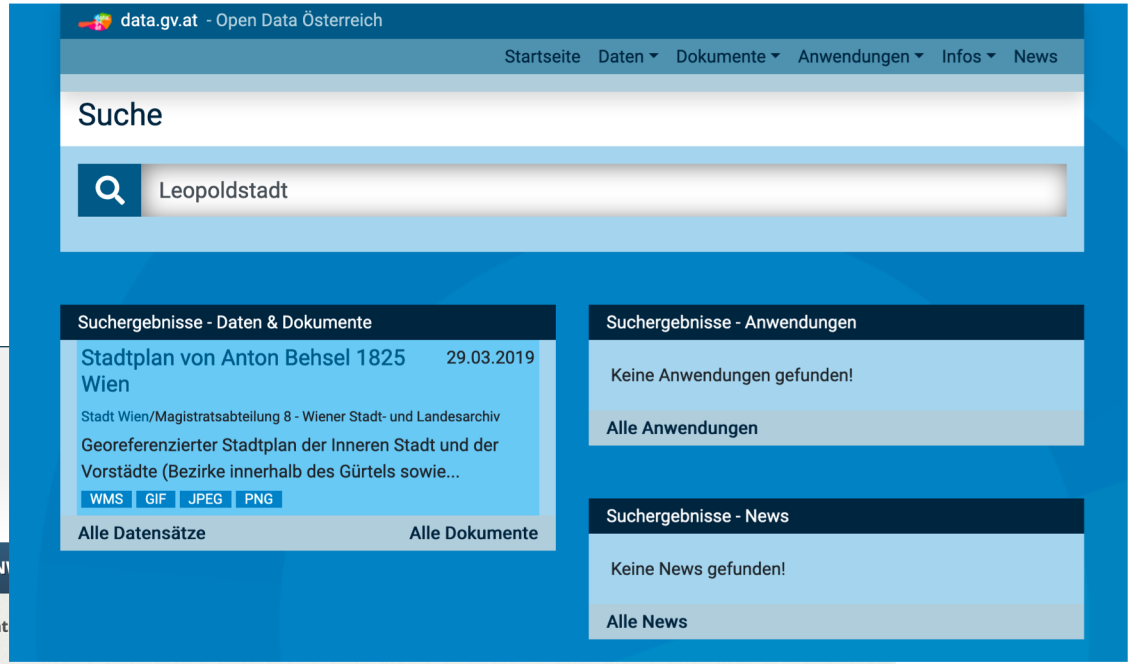
Data portals of the G8 countries

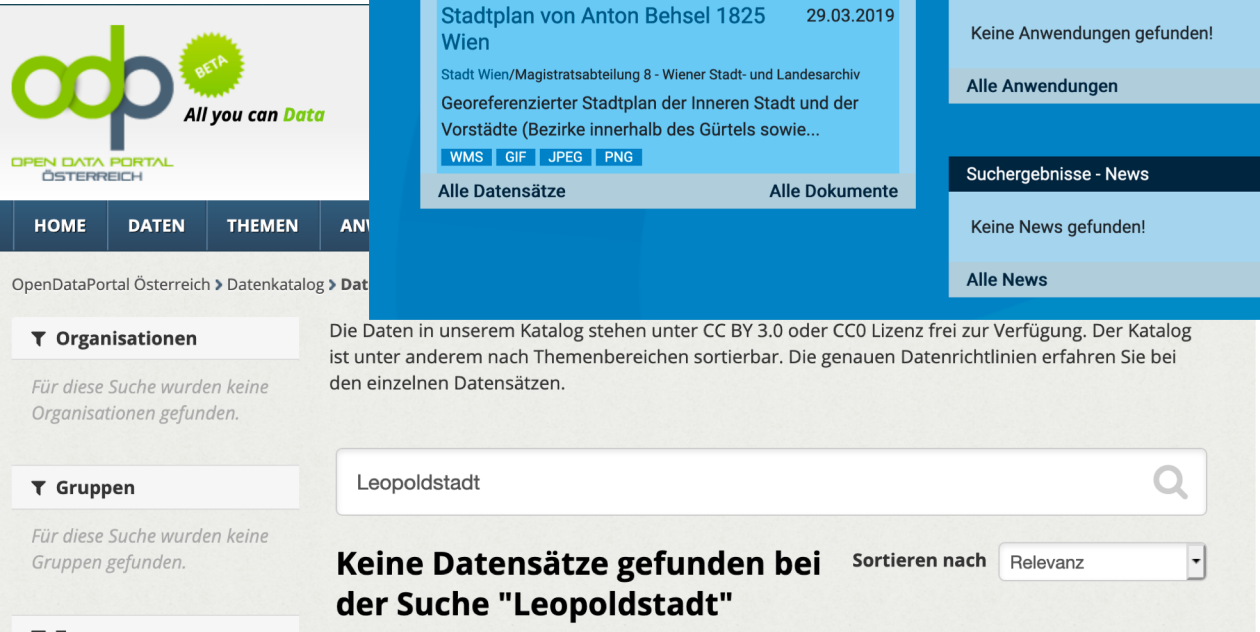
Different portals…





What do you find on Open Data Portals?





Not too much!

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?

- **Smart Sales and Customer Analytics:** Customer analyti… sales and marketing by leveraging customer related information, as well … interactions with different company's functions and IT systems. Customer analytics typically address …

- **Data Privacy and Organizational Aspects:** The … the increased availability of low-… privacy and security in the past years. The recent pu… Regulation (GDPR) poses strict limitations and rules on …

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

(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”

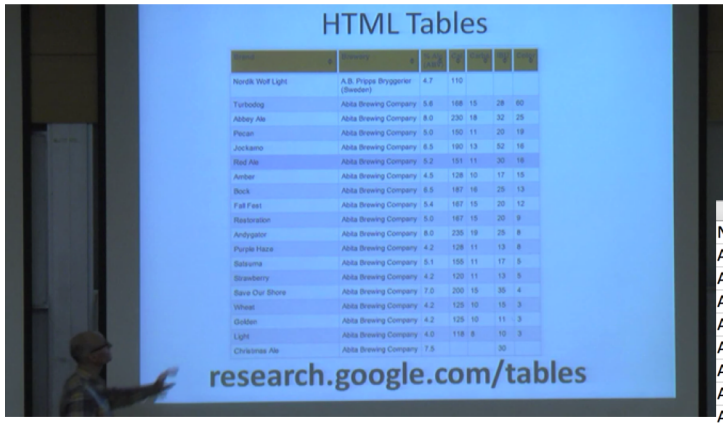
- Heterogenous data assets
- From different sources and origins
- Different Formats
- Different Semantics
- Sparse descriptive metadata
- Raw/not necessarily for human consumption
- Policies forbid to re-share sensitive Data
- Data Catalogs not even existing yet

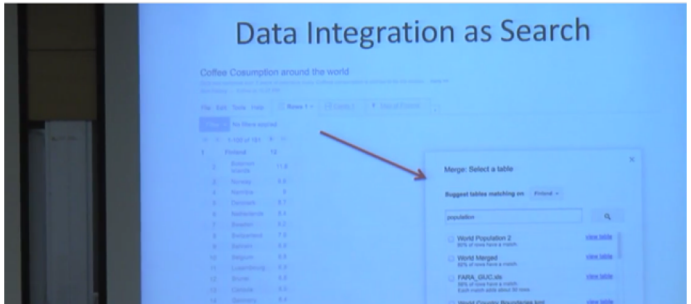
→ Hard to search and integrate!

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.

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…

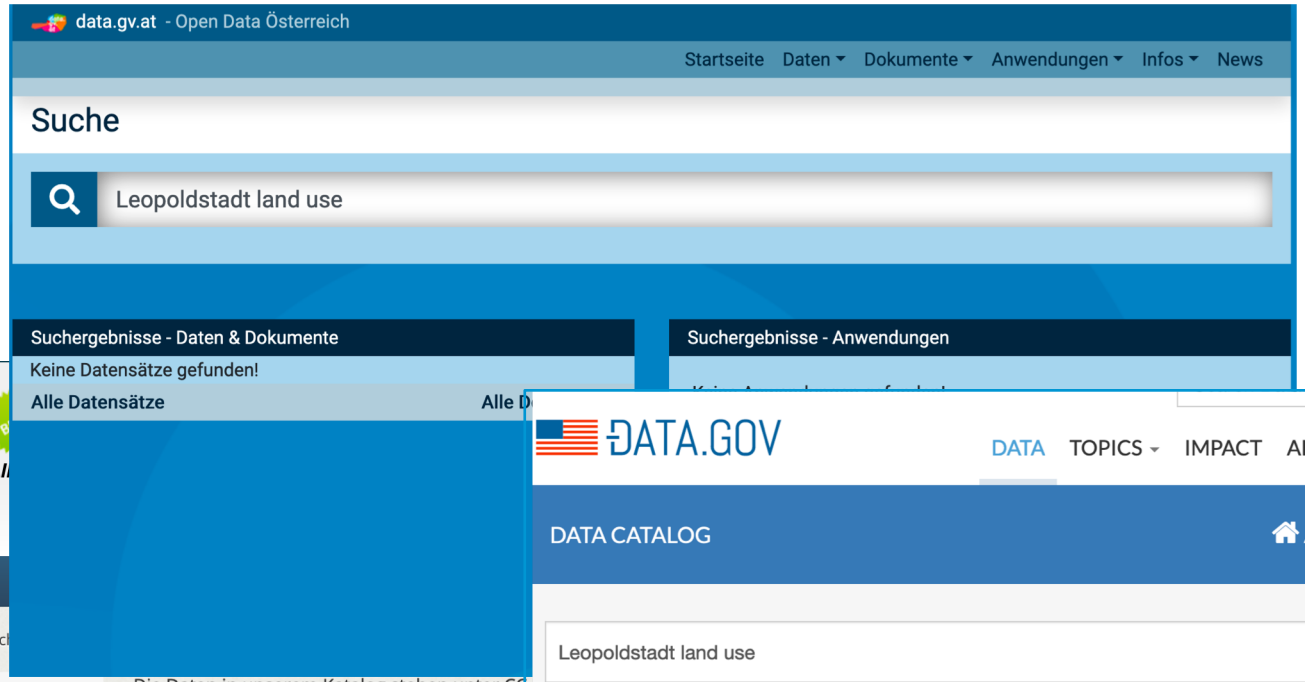
a) No natural language „cues" like in Web tables...

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

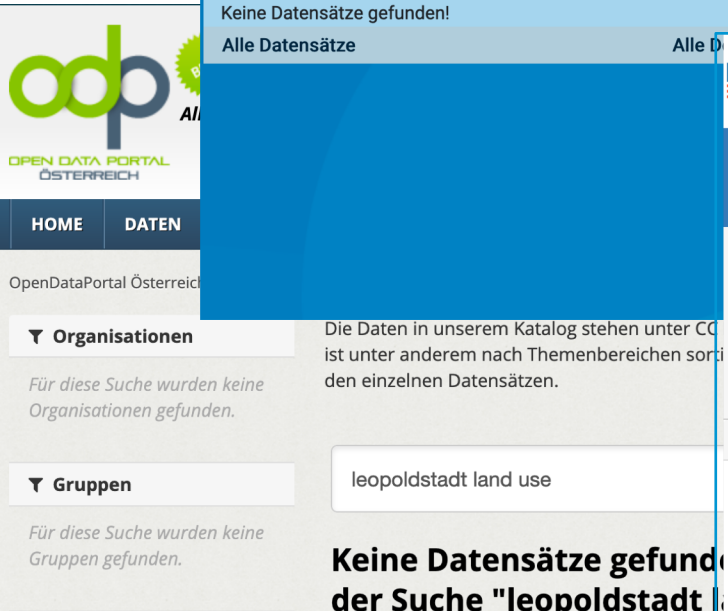
c) Open Data is not properly geo-referenced

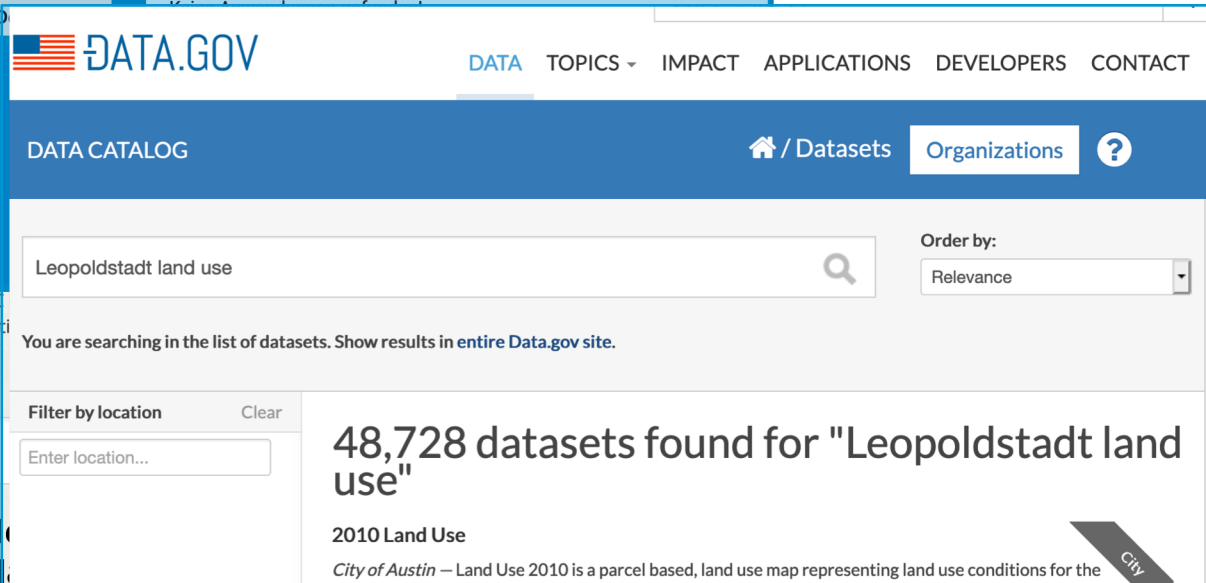
“Natural Data” Search: How would a human approach this problem?

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS



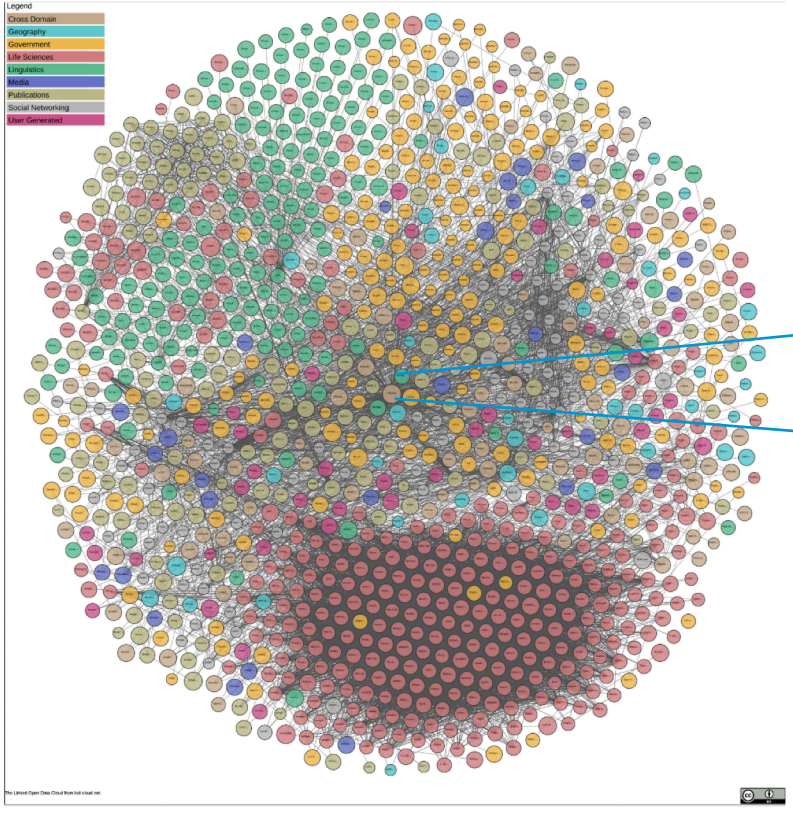
e.g. you wouldn’t probably search in data.gov …





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

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS



DBpedia

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!

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- 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

federal state	*district*	*year*	*sex*	*population*
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

NUTS2	*LAU2_NAME*	*YEAR*	*SEX*	*P_TOTAL*
AT31	Linz	2013	1	98157
AT31	Steyr	2013	1	18763
AT31	Wels	2013	1	29730
…	…		…	…

Source: https://www.data.gv.at/katalog/dataset/e108dcc3-1304-4076-8619-f2185c37ef81

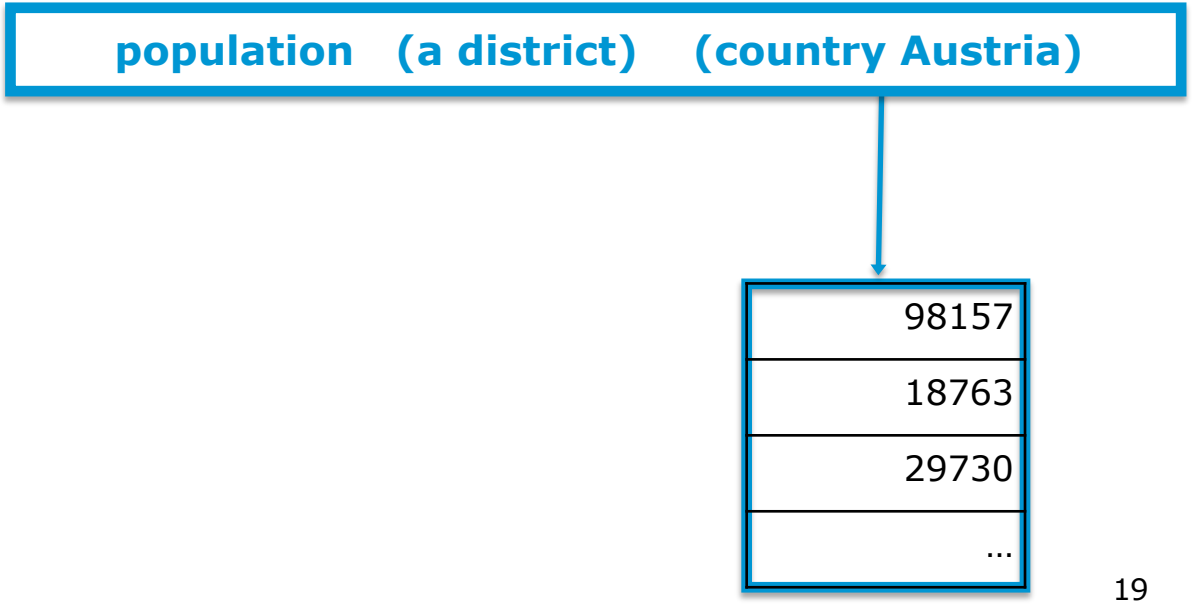
Why not use the numeric values?

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

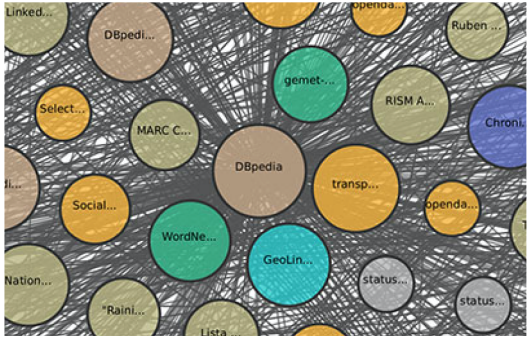
NUTS2	*LAU2_NAME*	*YEAR*	*SEX*	*P_TOTAL*
AT31	Linz	2013	1	98157
AT31	Steyr	2013	1	18763
AT31	Wels	2013	1	29730
…	…		…	…

Why not use numeric values?

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



Background Knowledge Graph

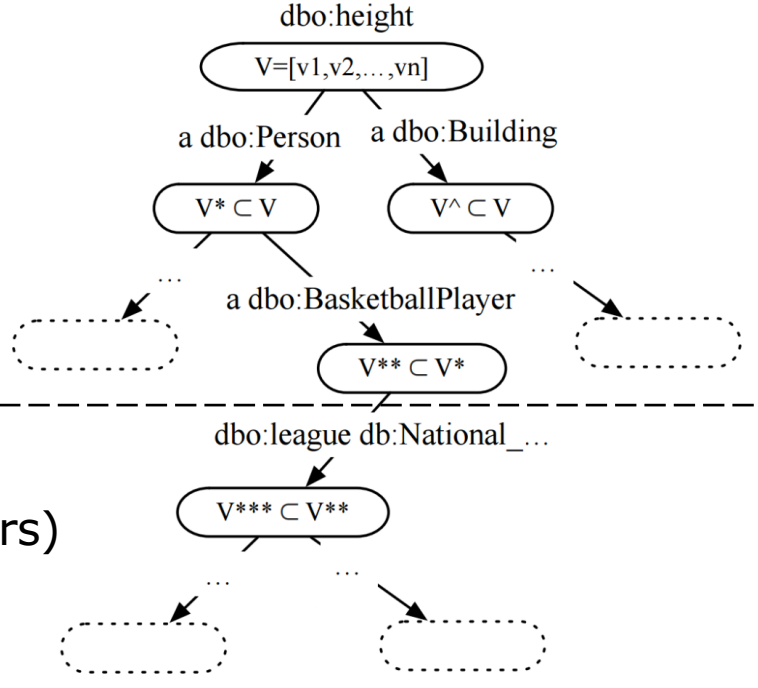


What's in there?

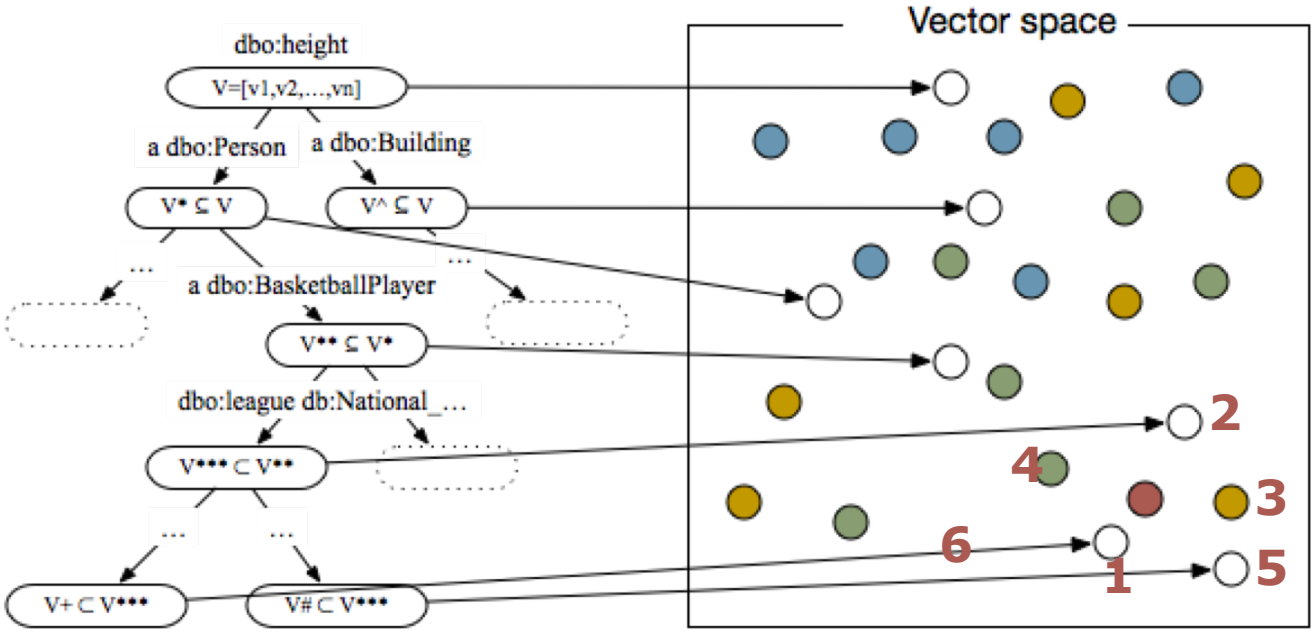
- Cities
  - **Population**
  - **Area**
  - Country
  - Location (**Coordinates)**
  - Economic indicators
  - …
- Organisations:
  - **Revenues**
  - Board members
  - **…**
- Persons (e.g. celebrities, sports)
  - Name
  - Profession
  - **Height**
- Landmarks (e.g. famous buildings)
  - **Country**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

Source: http://data.wu.ac.at/iswc2016_numlabels/submission/col14.html

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 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

What else can we do/use?

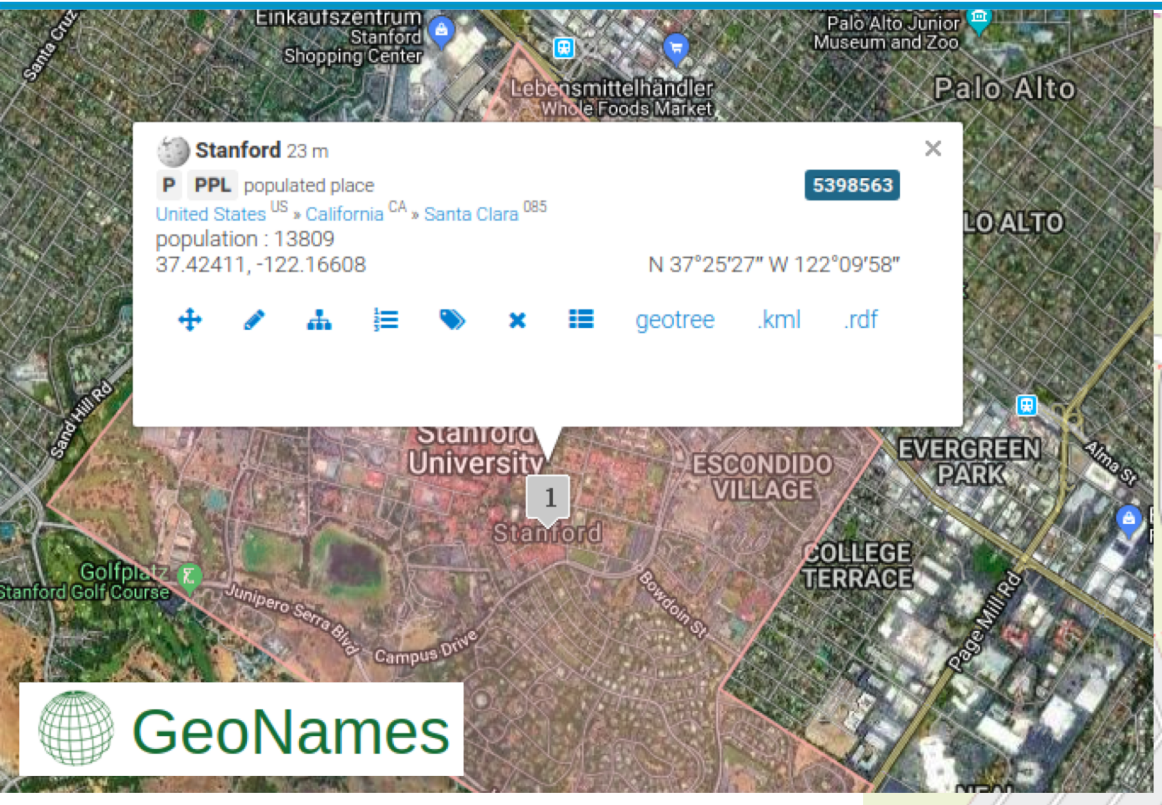
Focus on specific dimensions:

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

NUTS2	*LAU2_NAME*	*YEAR*	*SEX*	*AGE_TOTAL*
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



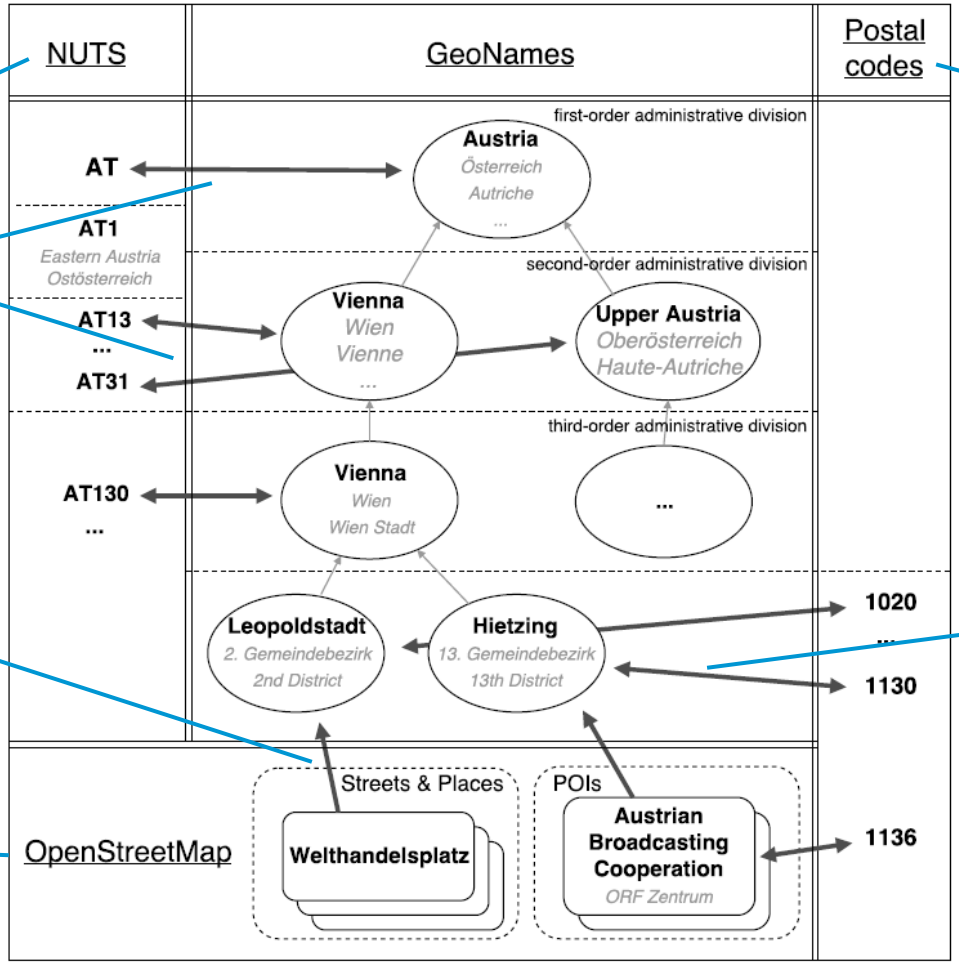
Geo-Knowledge Graph Construction

European Classification of Territorial Units

Wikidata links

Mapping OSM entities to GeoNames regions

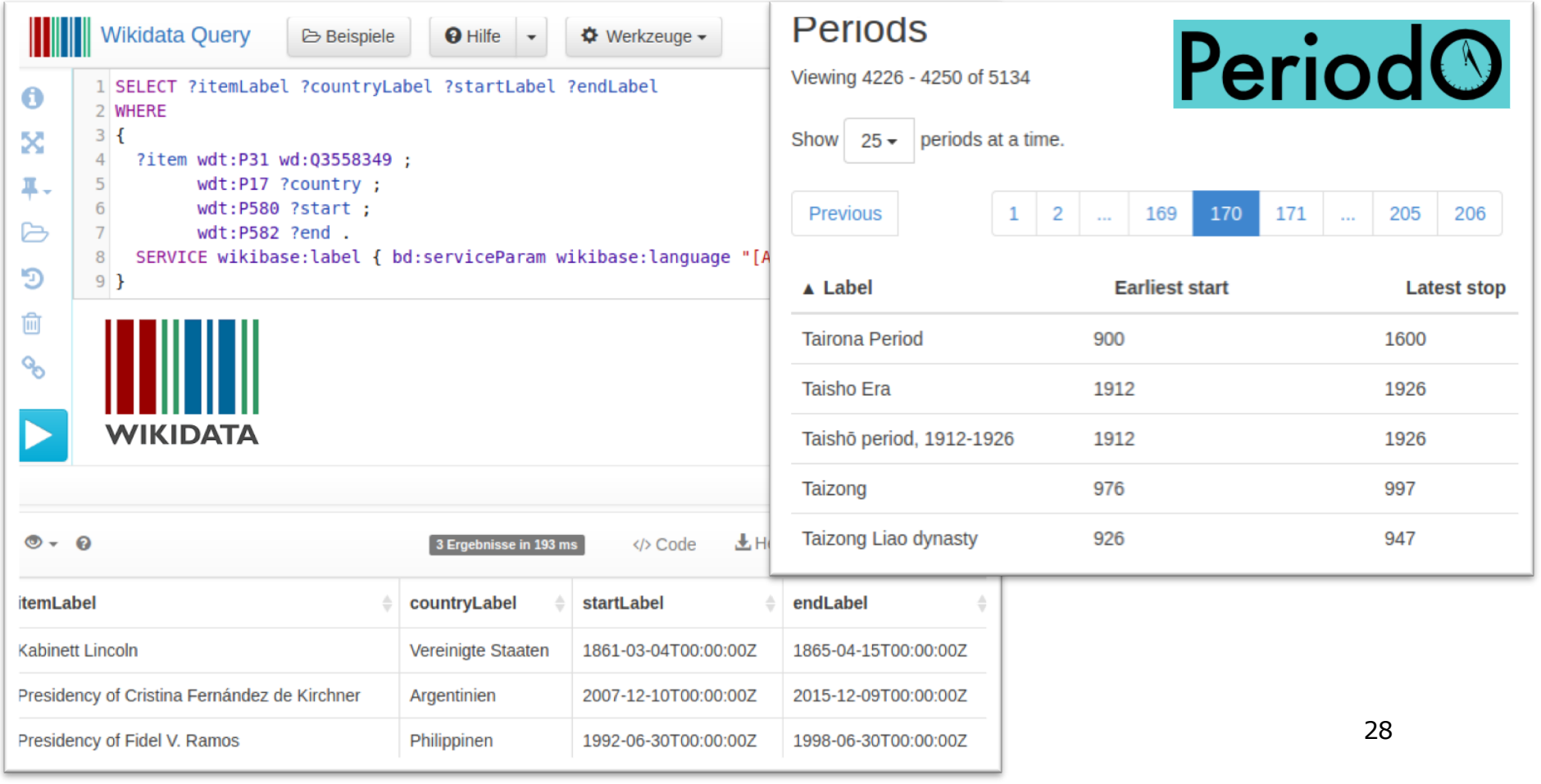
Extracting OSM streets and places



Wikidata, GeoNames

Wikidata links

Available Temporal Knowledge



Temporal Knowledge Graph Construction

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

```
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 )
}
```

```
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 ) }
```

- Named events and their labels
- Links to parent periods
- Temporal extent: a single beginning and end date
- Links to the spatial coverage

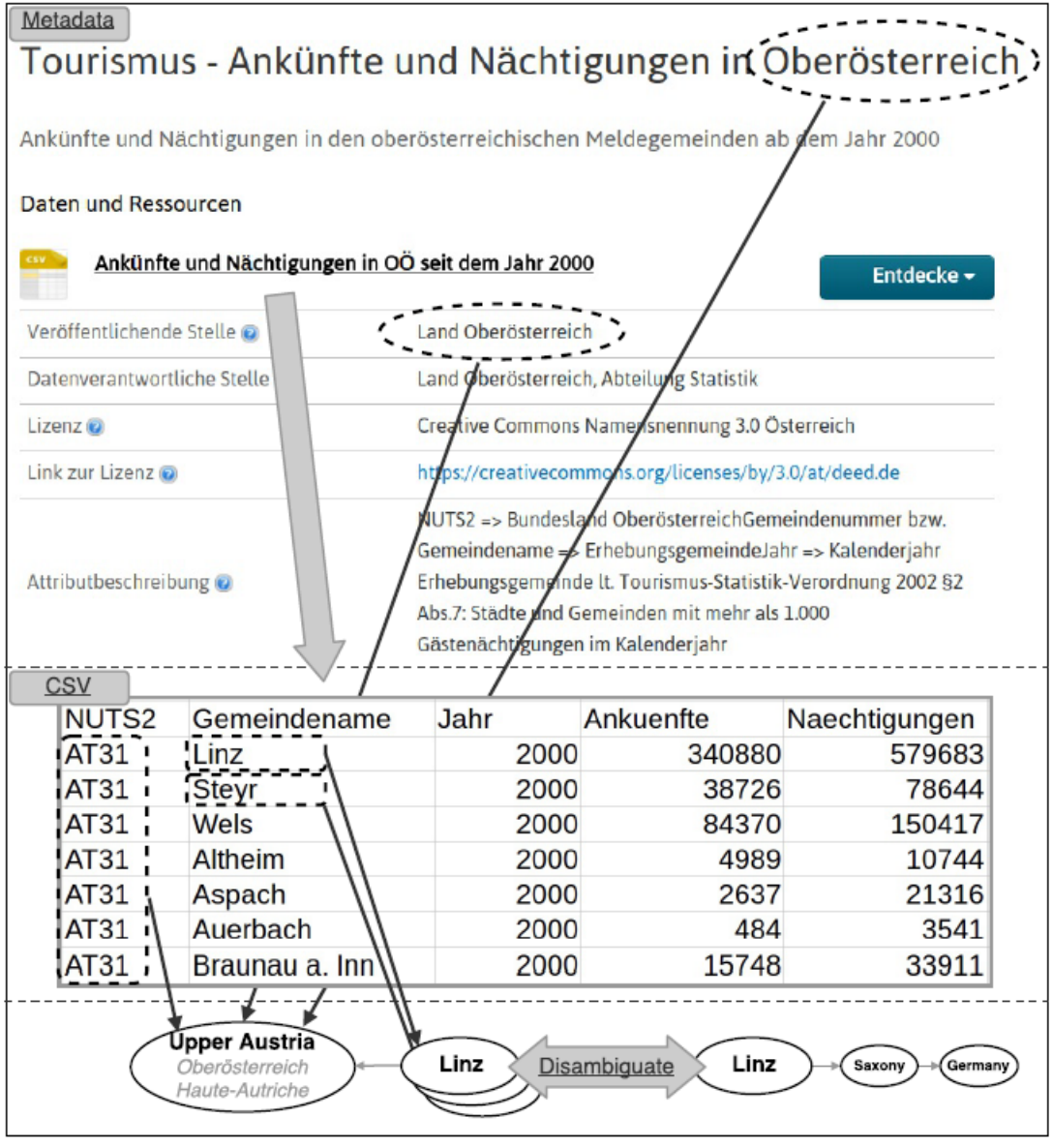
Dataset Labelling

Metadata descriptions

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

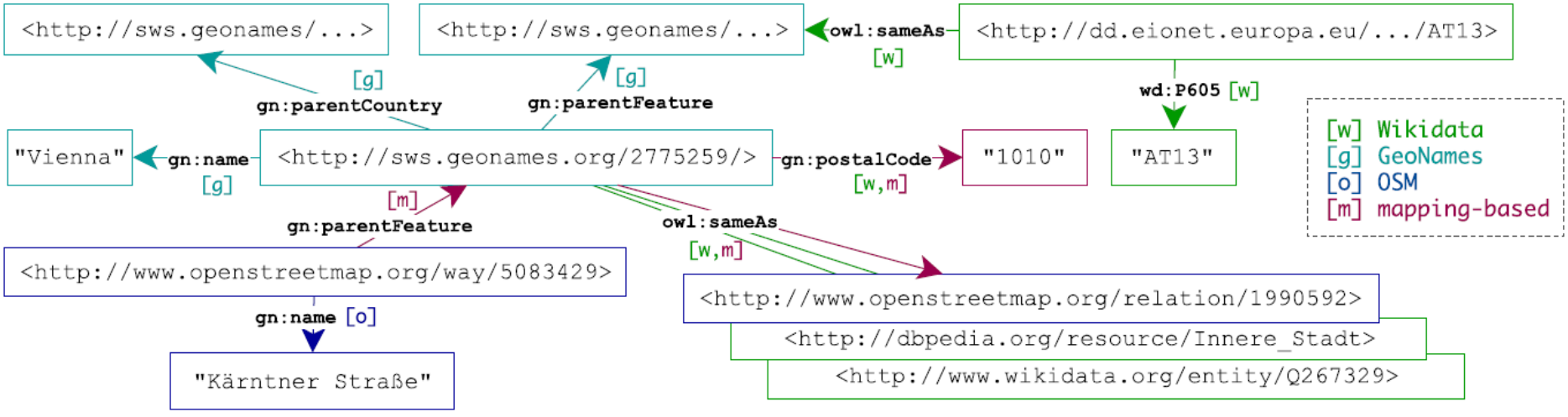
CSV cell value disambiguation

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



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

RDF Export 1/2: Knowledge Graph

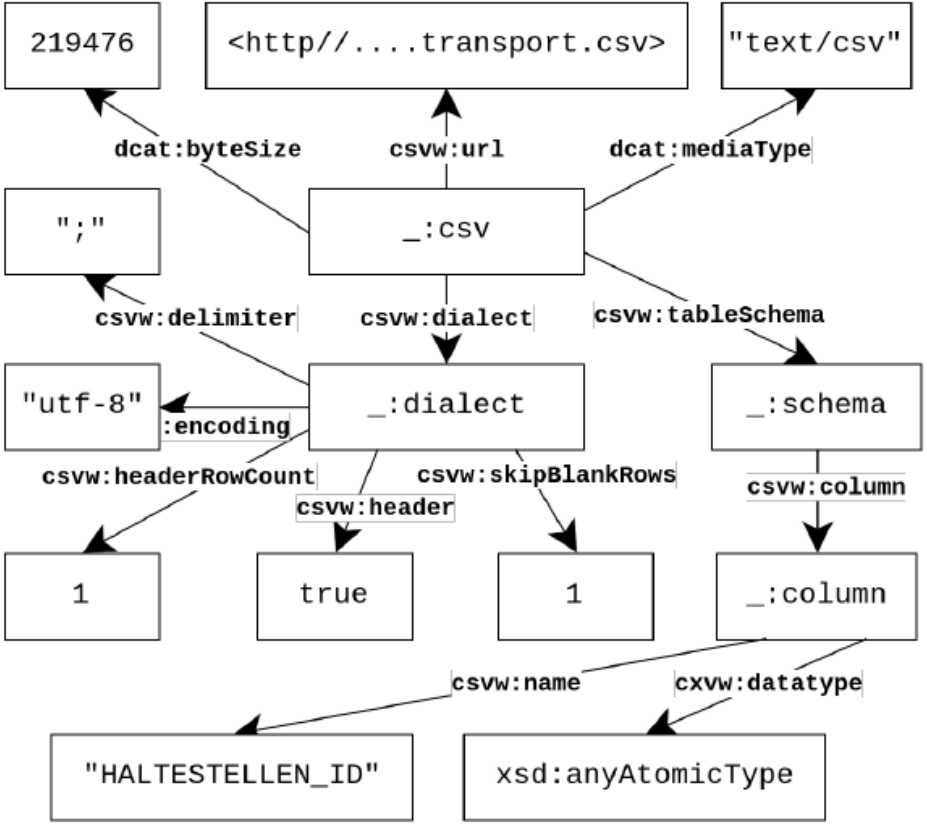


- Spatial and temporal base knowledge graph
- Annotated data points in metadata and CSV cells
- CSV metadata using CSVW vocabulary
  - e.g., delimiter, encoding, header, …

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

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

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



[4] R. Pollock et al., Metadata Vocabulary for Tabular Data, W3C CSV on the Web (2015)

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))
}
```

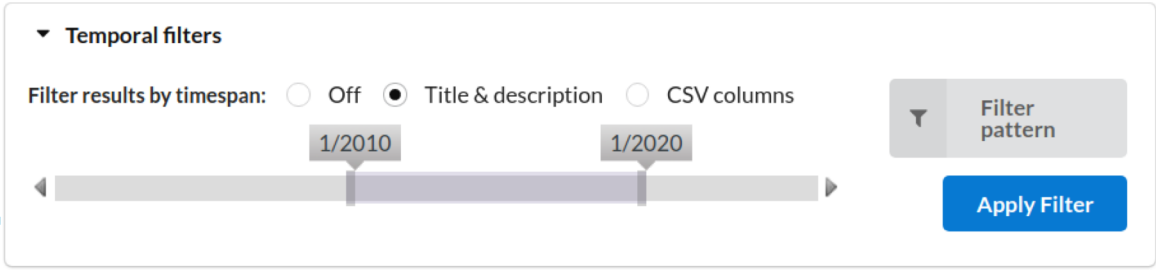
Search Interface

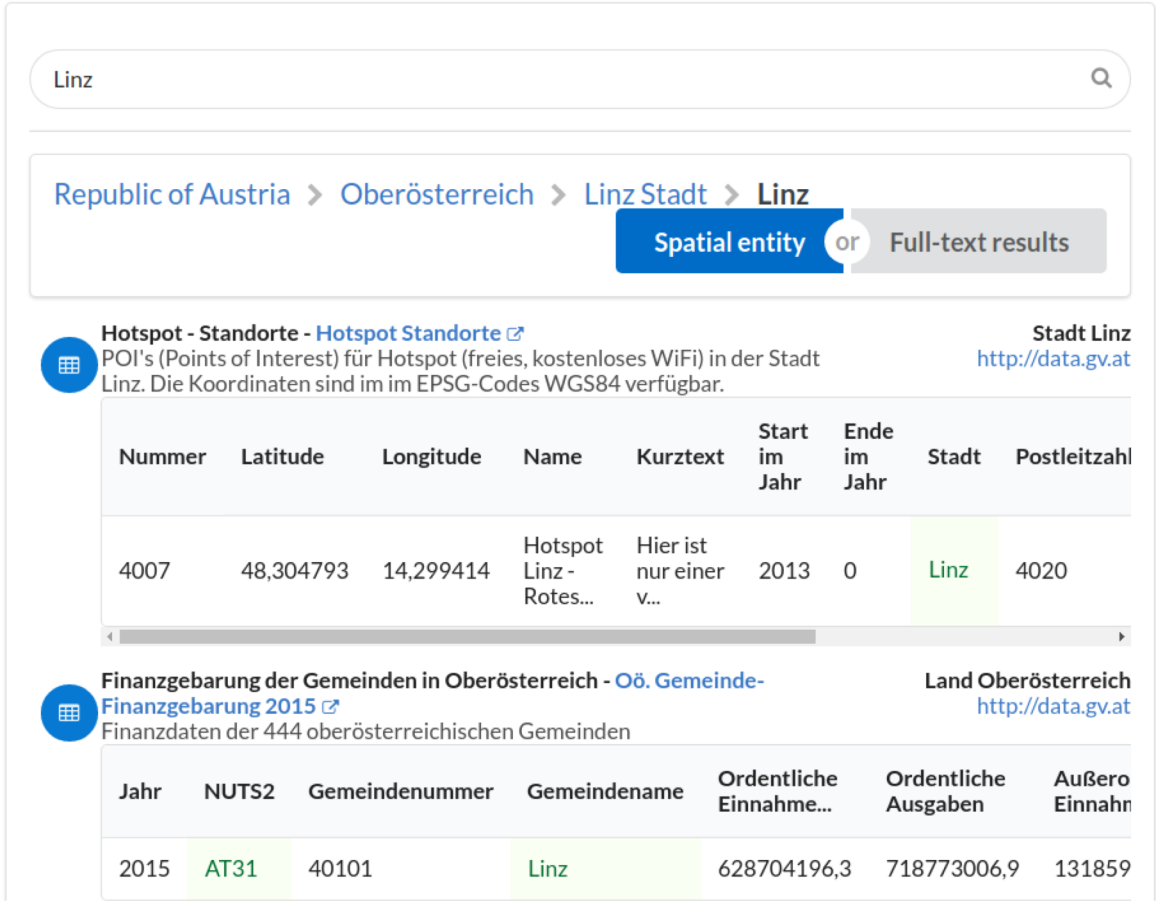
Faceted query interface:

- 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





How well does it work? Indexed Datasets

portal	datasets	CSVs	indexed
total			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

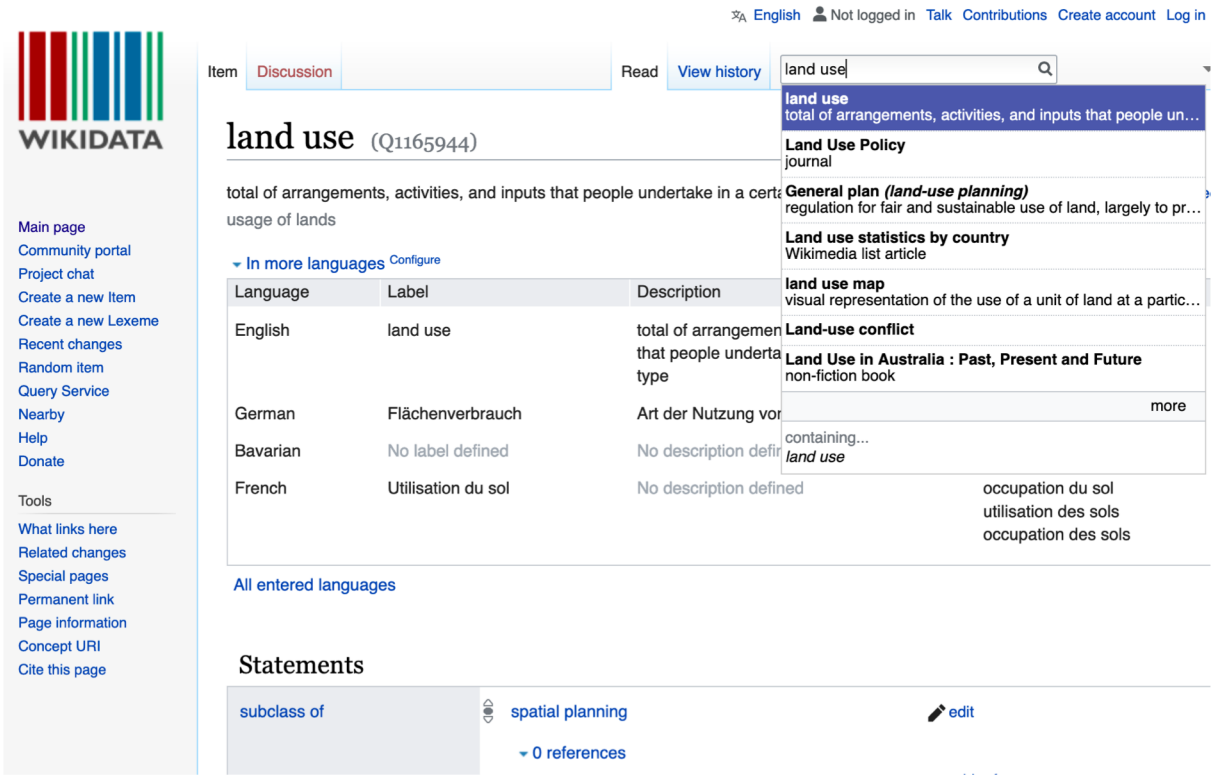
Intuitively, most datasets found on governmental Open Data portals are organized by spatio-temporal criteria, that is, single datasets provide data for a certain region, valid for a certain time period. Likewise, for many use cases (such as, for instance, data journalism and fact checking) a pre-dominant 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



- 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?

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- 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?*


PhD topics
inside

Backup slides:

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

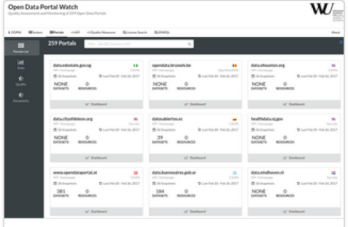
Projects

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

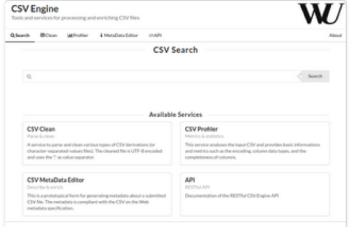


Open Data Portal Watch

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

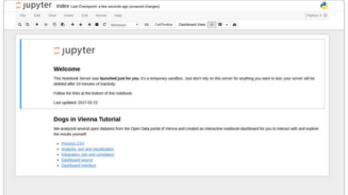
Search & enrich CSVs

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.gv.at and opendataportal.at.

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

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

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

6/2016 [3]:
- 260 portals
- CKAN, **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

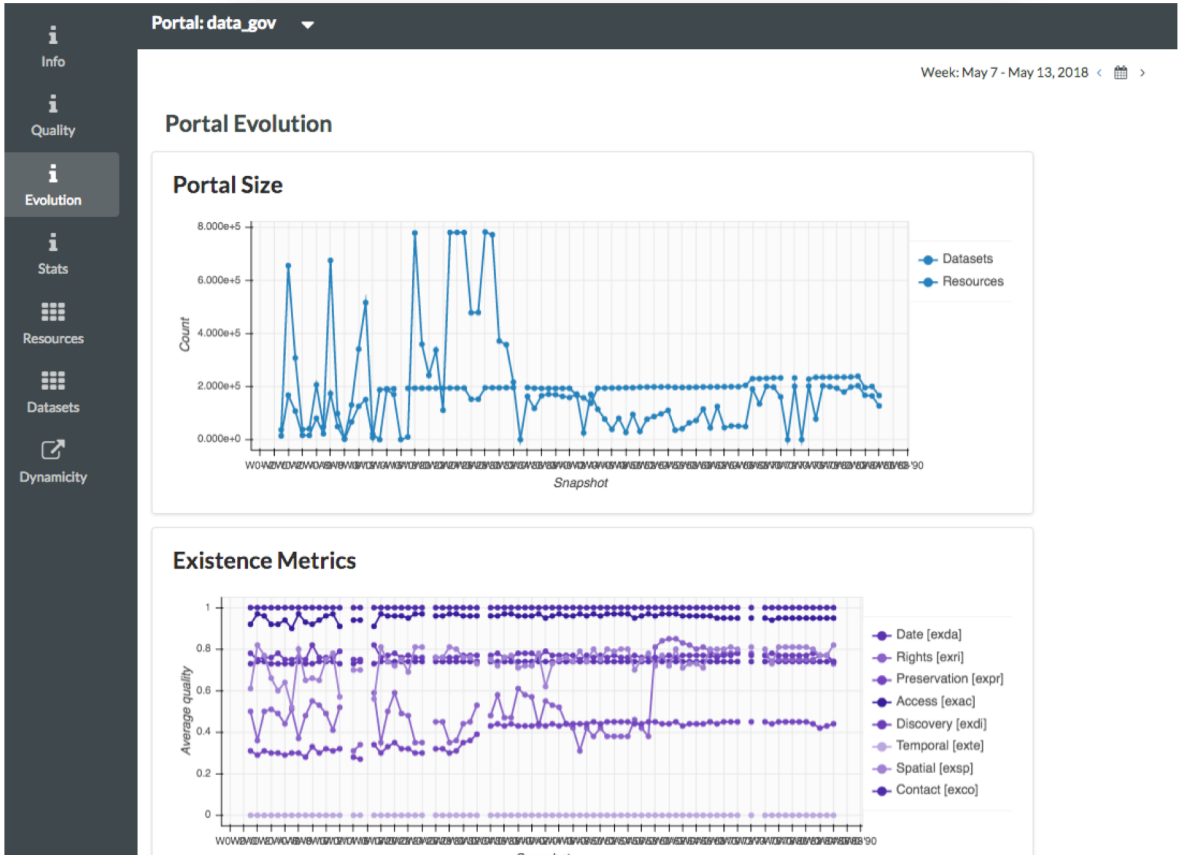
[1] Towards assessing the quality evolution of open data portals. In ODQ2015: Open Data Quality Workshop, Munich, Germany

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

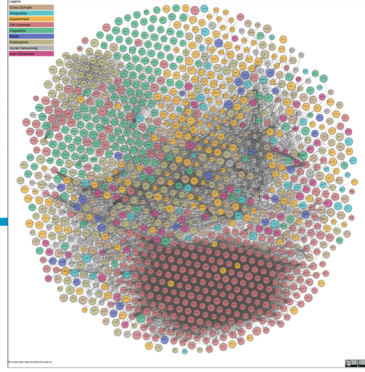
[3] Automated quality assessment of metadata across open data portals. ACM Journal of Data and Information Quality (2016)

Demo:

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



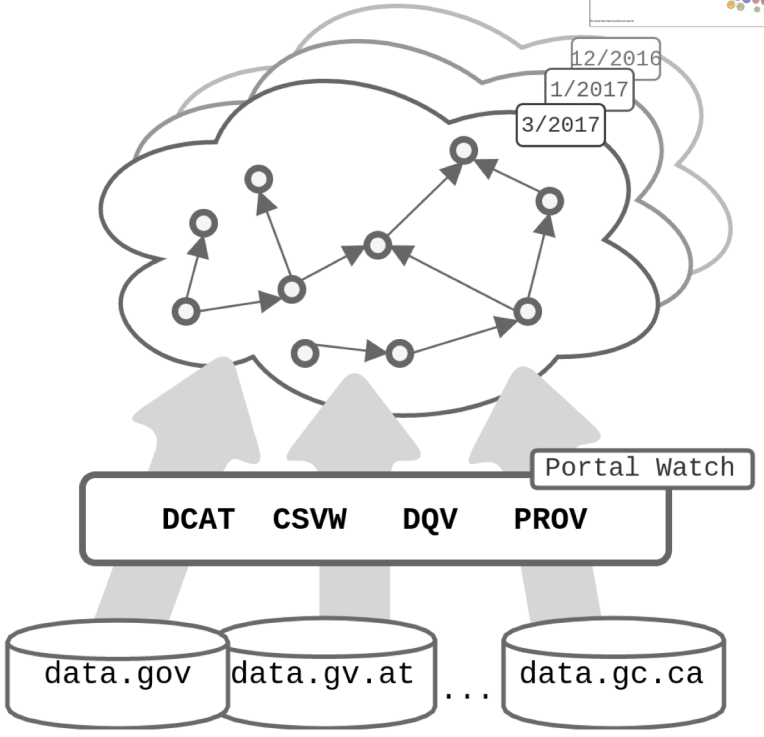
2) Mapping to Standard vocabularies & Linked Data



- 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/