



WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS



Metadata Quality: Learning from Open Data Portalwatch

Axel Polleres

twitter: @AxelPolleres

web: polleres.net



SPARQL



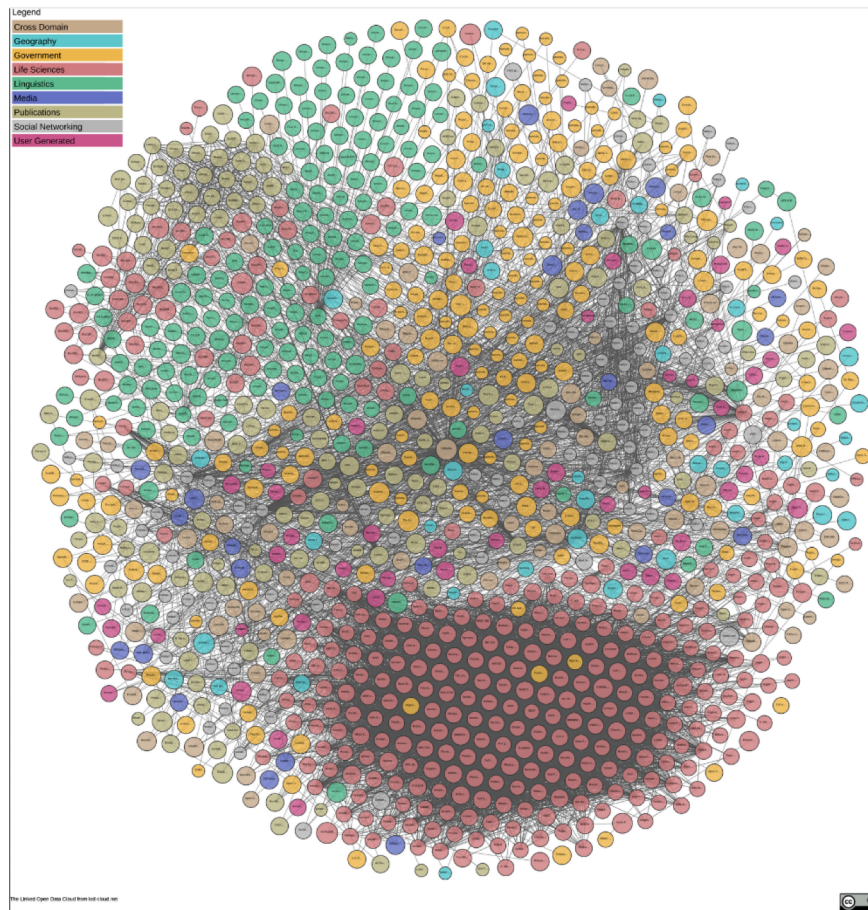
**Open
Data**



**Linked
Data**

Linked Data

But: **Open Data** is more than Linked Open Data...



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.at **Open Government Data**
Offene Daten für Wien

UNdata

Open Data Berlin

london.gov.uk

Opening up Europe's public data

DATA.GOV

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 234,627 DATASETS

Manufacturing & Trade Inventories & Sales

A lot of Open Data is not Linked Data

- Cities, International Organizations, National and European Portals, Int'l. Conferences:



We are aware of currently 260 active such portals worldwide.

Different portals...

DATA.GOV DATA TOPICS - IMPACT APPLICATIONS DEVELOPERS CONTACT

DATA CATALOG / Datasets Organizations ?

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 dataset 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 (17730 views) [Download](#)

Dates

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

Metadata Source

Data.json Metadata [Download Metadata](#)

Harvested from HUD.JSON

affordability cost fmr households housing income rent renter

data.gv.at data.gv.at - offene Daten Österreichs

Suchbegriff (z.B. Finanzen, Wahlen) [Suche starten](#)

Datenkatalog Apps & News [Katalog durchstöbern](#)

API

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-BABE1-0](#) [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

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=71137735-2c65-328f-b57d-be941ada765e

RSS-Feeds für Statistik Austria

geänderte Datensätze

Letzte Änderung

30.04.2018 00:59:46

C-A10-0;Zeit:C-BARG-0;Regionale Gliederung:C-BABE1-0;Bildungseinrichtung:F-INSG;Ausgaben (gesamt);F-TR_PA;Personalaufwand;F-TR_SA;Sachaufwand;F-

What's the problem(s)?

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

Open Data Portal Software

CKAN ... <http://ckan.org/>

- almost „de facto“ standard for Open Data Portals
- facilitates search, metadata (publisher, format, publication date, license, etc.) for datasets

- <http://data.gov/>
- <http://data.gv.at/>

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

The screenshot shows the homepage of data.gv.at. The browser address bar displays 'http://www.data.gv.at/'. The page features a search bar with the placeholder text 'Suchbegriff (z.B. Finanzen, Wahlen)' and a 'Suche starten' button. Below the search bar, there are navigation links for 'Datenkatalog', 'Anwendungen & News', and 'Katalog durchstöbern'. A main navigation menu includes 'Startseite', 'Katalog', 'Anwendungen', 'News', 'Hintergrund-Infos', 'Netiquette', and 'Kontakt'. The main content area has the heading 'offene Daten Österreichs – lesbar für Mensch und Maschine' and a sub-heading 'Vielfalt, Transparenz, Offenheit, Demokratie'. It describes the portal as a 'Katalog offener Datensätze und Dienste' and provides information on how to use the data. A diagram on the right shows a computer monitor displaying binary code, with arrows pointing to a group of people and a smartphone, illustrating the accessibility of the data for both humans and machines.

Our solution:

- Open Data Portalwatch
 - Monitoring Metadata quality
 - Mapping to standard vocabularies
 - Enriching Metadata to improve search

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

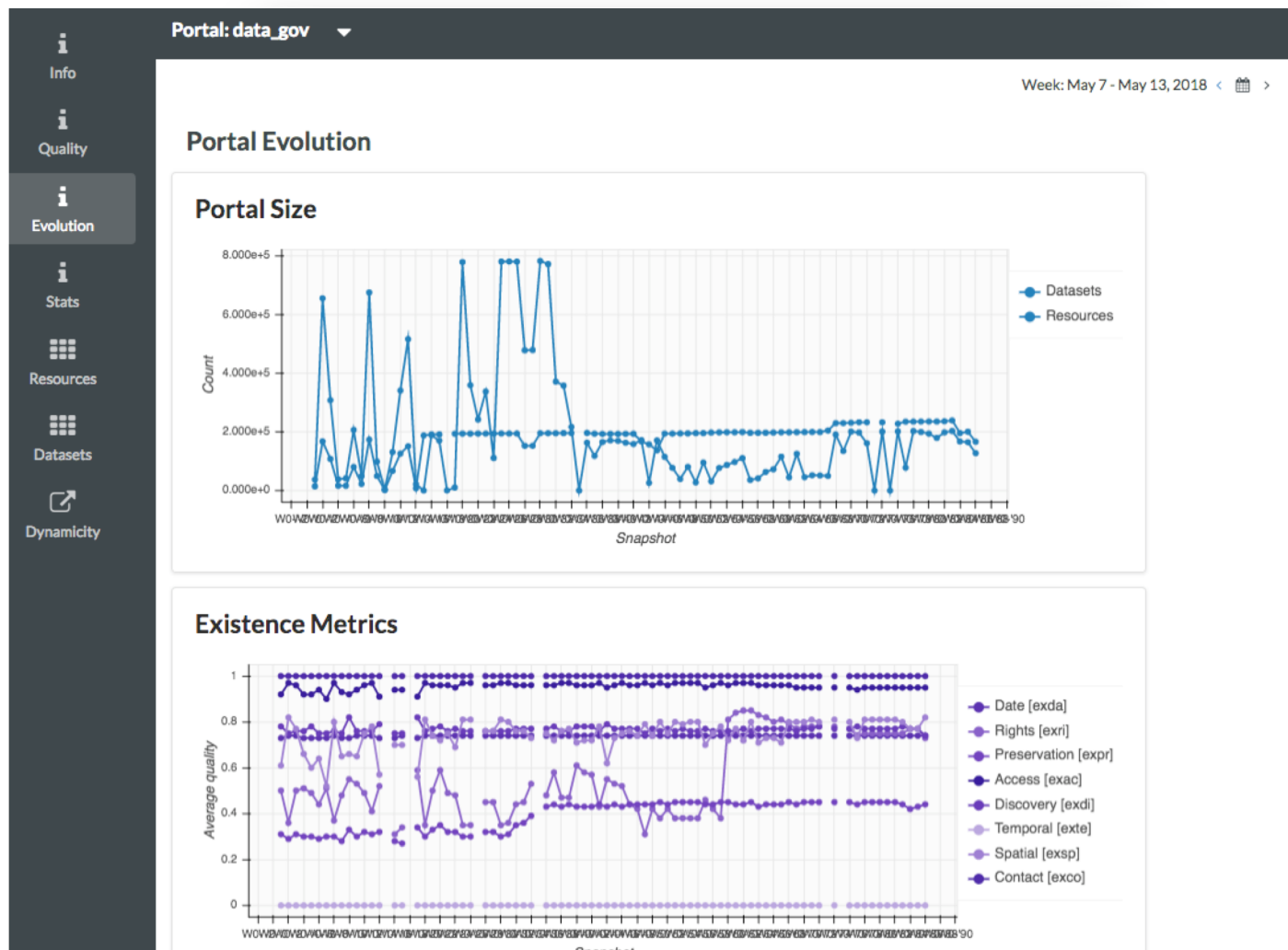
[1] Towards assessing the quality evolution of open data portals. In ODQ2015: Open Data Quality Workshop, Munich, Germany

[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)

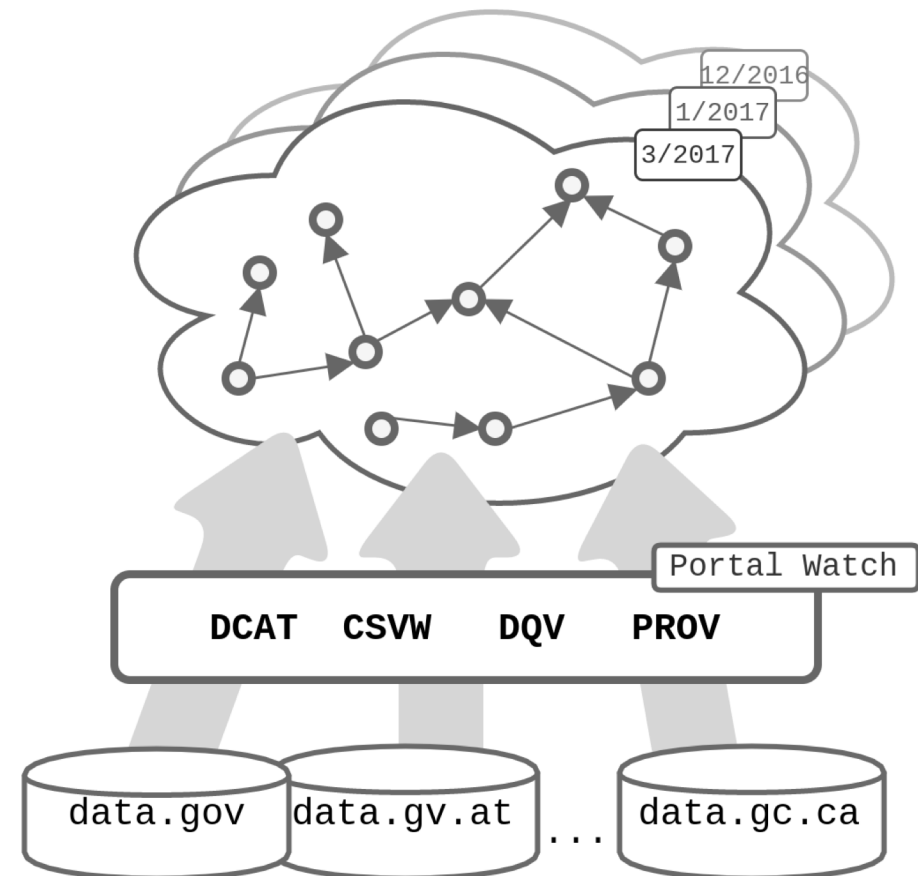
Demo:

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/>

Finally: 3) Why is Search in Open Data a problem?

data.gv.at – offene Daten Österreichs

Leopoldstadt

Daten & Dokumente ● Apps & News → Katalog durchstöbern

Startseite **Daten** ▼ Dokumente ▼ Anwendungen ▼ Infos ▼

Katalogsuche

Leopoldstadt Wildcards (*) für Suche nach Wortteilen werden unterstützt.

Filter

Suchergebnis zu "Leopoldstadt" (0 gefunden) Seite 1 von 0

alle Datensätze anzeigen Ergebnisseiten: ← Erste Letzte (0) →

Suchergebnisse von opendataportal.at (0 gefunden)

Titel	Veröffentlichende Stelle / Datenverantwortliche	Veröffentlicht auf	Letzte Änderung auf	Format	Lizenz
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

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

VS.

Katalog
Bevölkerung in Wien: Bezirk - Geschlecht

HTML Tables

Beer	Company	ABV	IBU	SRM	Color
Novik Wolf Light	A.B. Pipsor Bryggerier (Sweden)	4.7	110		
Turbodog	Abba Brewing Company	5.6	166	15	28 80
Abbey Ale	Abba Brewing Company	6.0	230	18	32 25
Piccan	Abba Brewing Company	5.0	150	11	20 19
Jockamo	Abba Brewing Company	6.5	190	13	52 16
Red Ale	Abba Brewing Company	5.2	151	11	30 16
Amber	Abba Brewing Company	4.5	128	10	17 15
Rock	Abba Brewing Company	6.5	187	16	25 13
Fall Fest	Abba Brewing Company	5.4	167	15	20 12
Reclamation	Abba Brewing Company	5.0	167	15	20 9
Andygar	Abba Brewing Company	6.0	235	19	25 8
Purple Haze	Abba Brewing Company	4.2	128	11	13 8
Batsuma	Abba Brewing Company	5.1	155	11	17 5
Strawberry	Abba Brewing Company	4.2	120	11	13 5
Save Our Shore	Abba Brewing Company	7.0	200	15	30 4
Wheat	Abba Brewing Company	4.2	125	10	15 3
Golden	Abba Brewing Company	4.2	125	10	11 3
Light	Abba Brewing Company	4.0	118	8	10 3
Christmas Ale	Abba Brewing Company	7.5			30

research.google.com/tables

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

Data Integration as Search

Coffee Consumption around the world

population

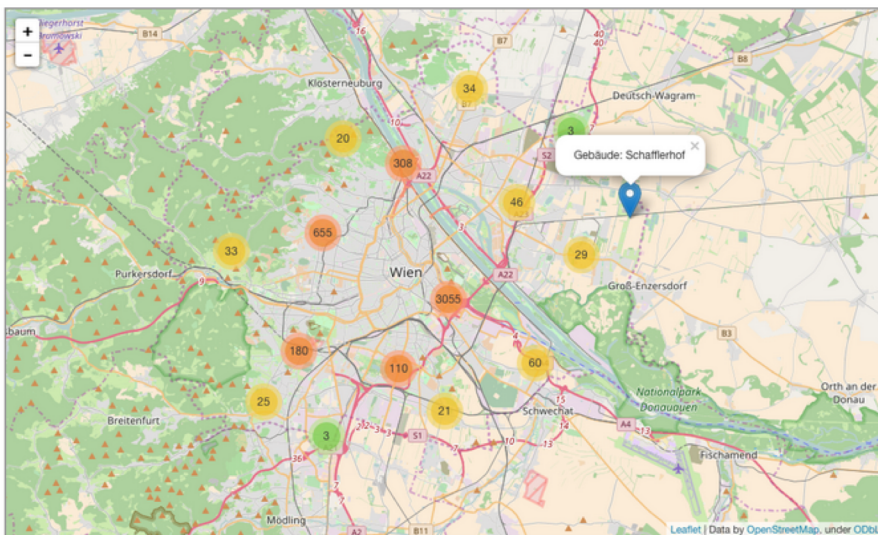
- World Population 2 97% of world population
- World Merged 97% of world population
- FARA, GARCIA 97% of world population
- World Countries 97% of world population

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"
- c) Open Data is not properly geo-referenced

Some starting points:

- First baby steps on building an Open Data Knowledge Graph:
- Ongoing work to enable **spatio-temporal search** in Open Data e.g. in our project communidata.at (just submitted to JWS)



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 (Linked) Data. Most existing approaches to “semantically label” such tabular data rely on mappings of textual information to classes, properties, or instances in RDF knowledge bases in order to link – and eventually transform – tabular data into RDF. However, as we will illustrate, Open Data tables typically con-

Enabling Spatio-Temporal Search in Open Data

Sebastian Neumaier^{a,1}, Axel Polleres^{a,b,c,2}

^aVienna University of Economics and Business, Vienna, Austria

^bComplexity Science Hub Vienna, Austria

^cStanford University, CA, USA

Abstract

Intuitively, most datasets found in Open Data are organised by spatio-temporal scope, that is, single datasets provide data for a certain region, valid for a certain time period. 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. Therefore, we argue that spatio-temporal search is a crucial need for Open Data portals and across Open Data portals, yet - to the best of our knowledge - no working solution exists. We argue that - just like for for regular Web search - *knowledge graphs* can be helpful to significantly improve search: in fact, the ingredients for a public knowledge graph of geographic entities as well as time periods and events exist already on the Web of Data, although they have not yet been integrated and applied - in a principled manner - to the use case of Open Data search. In the present paper we aim at doing just that: 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, (iii) enable structured, spatio-temporal search over Open Data catalogs through our spatio-temporal knowledge graph, both via a search interface as well as via a SPARQL endpoint, available at data.vu.ac.at/odgraphsearch/

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

Demo:

<http://data.wu.ac.at/odgraphsearch/>

Spatio-temporal search in Open Data



Q Search SPARQL API About

Search Open Data

Temporal filters

Gemeindebezirk Leopoldstadt

Republic of Austria > Wien > Wien Stadt > Gemeindebezirk Leopoldstadt

Spatial entity or Full-text results

Hunde pro Bezirk Wien - Anzahl der Hunde pro Bezirk [Stadt Wien](#)
<http://data.gv.at/>

NUTS1	NUTS2	NUTS3	DISTRICT_CODE	SUB_DISTRICT_CODE	Postal_CODE	Dog Breed	Anzahl	Ref_Date
AT1	AT13	AT113	90200	.	1020	Zwergspitz / Mischl...	10	20171130

Publizistikförderung - Publizistikförderung [RTR-GmbH](#)
<http://data.gv.at/>

Im Rahmen der Publizistikförderung werden periodische Druckschriften gefördert, die sich mit politischen, kulturellen oder weltanschaulichen Themen befassen. Beschreibungen der Daten sowie der Möglichkeit der Datenabfrage über eine REST-Schnittstelle siehe <https://data.rtr.at/Publizistik>.

gesetzlichegrundlage	zeitschrift	foerderungswerber	strasse	plz	ort	foerderbetrag	jahr	status
Abschnitt II PubFG 1...	TARANTEL	Werkkreis Literatur ...	Vivariumstraße 8/4/1...	1020	Wien	5742.22	2017	abgesch

Top Locations Wien - top-locations-wien.csv [Stadt Wien](#)
<http://data.gv.at/>

Touristische Auswahl der wichtigsten POIs in Wien. Ca 140 POIs in den Kategorien Sightseeing, Museen, Gastronomie, Nightlife, Musik, Shopping, Cafés und Restaurants. Jede Location enthält allgemeine Infos wie z.B. Adresse, Telefonnummer sowie eine Kurzbeschreibung und die Geodaten.

title	category	Beschreibung	address	zip	city	geo_latitude	geo_longitude	tel_1	tel_:
Wiener Sängerknaben	musicstage	Die "jüngsten musika	Obere Augartenstraße	1020	Wien	48,224458	16,3734017		

Idea:

Link OD to geospatial and temporal Knowledge bases:

- Geonames
- OpenStreetMap
- Perio.do
- Wikidata
- ...

Take-home messages:

- Scalable monitoring of metadata quality and evolution is possible and useful
- Meta-Data Quality can be improved by IE and looking into the data
- Linked Data and standard protocols (SPARQL, Memento)
 - helps to provide an integrated view
 - Enable combination with other sources to improve search

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
Tools and services for processing and enriching CSV files

CSV Search

Available Services

- CSV Clean**: A service to parse and clean various types of CSV descriptions (or character-separated values files). The cleaned file is UTF-8 encoded and uses the "T" as value separator.
- CSV Profiler**: This service analyzes the input CSV and provides basic information and metrics such as the resulting column data types, and the completeness of columns.
- CSV Metadata Editor**: This is a prototype of forms for generating metadata about a submitted CSV file. The metadata is compatible with the CSV on the Web metadata specification.
- API**: Documentation of the RESTful CSV Engine API.

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.