WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

Metadata Quality: Learning from Open Data Portalwatch

Axel Polleres twitter: @AxelPolleres web: polleres.net

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

SPARQL

Open Data

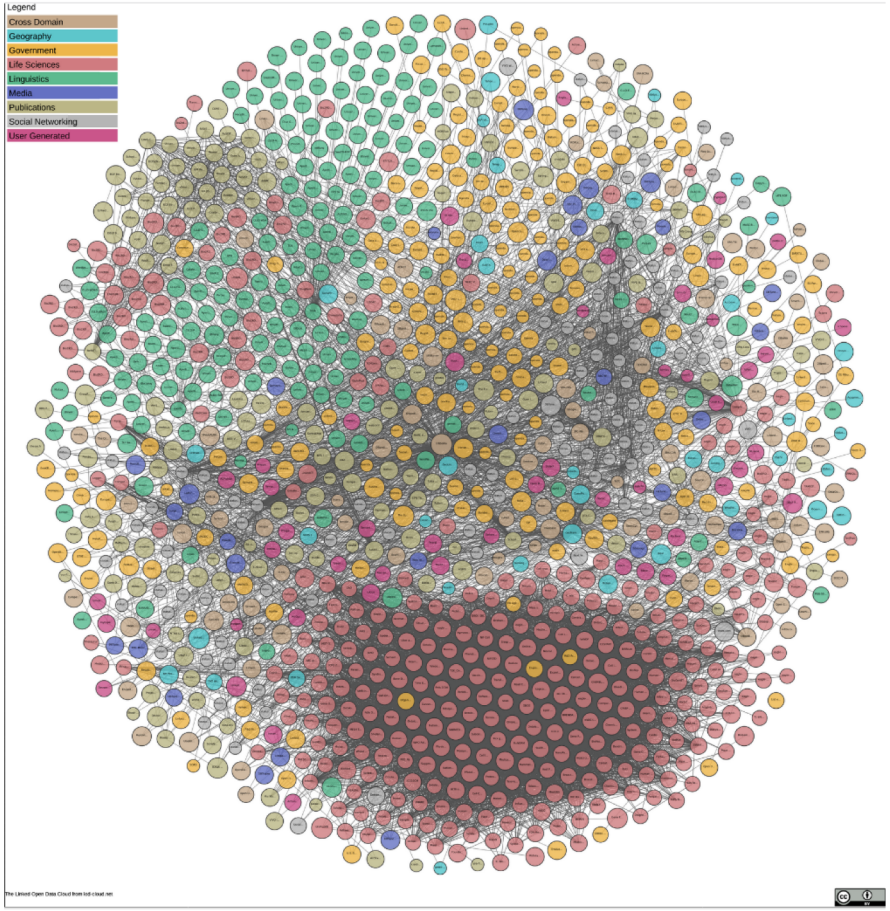
Linked Data

Latest release 04-30-2018- 1184 Datasets

Linked Data

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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



Latest release 04-30-2018- 1184 Datasets

Open Data is a Global Trend!

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

EFMD EQUIS ACCREDITED

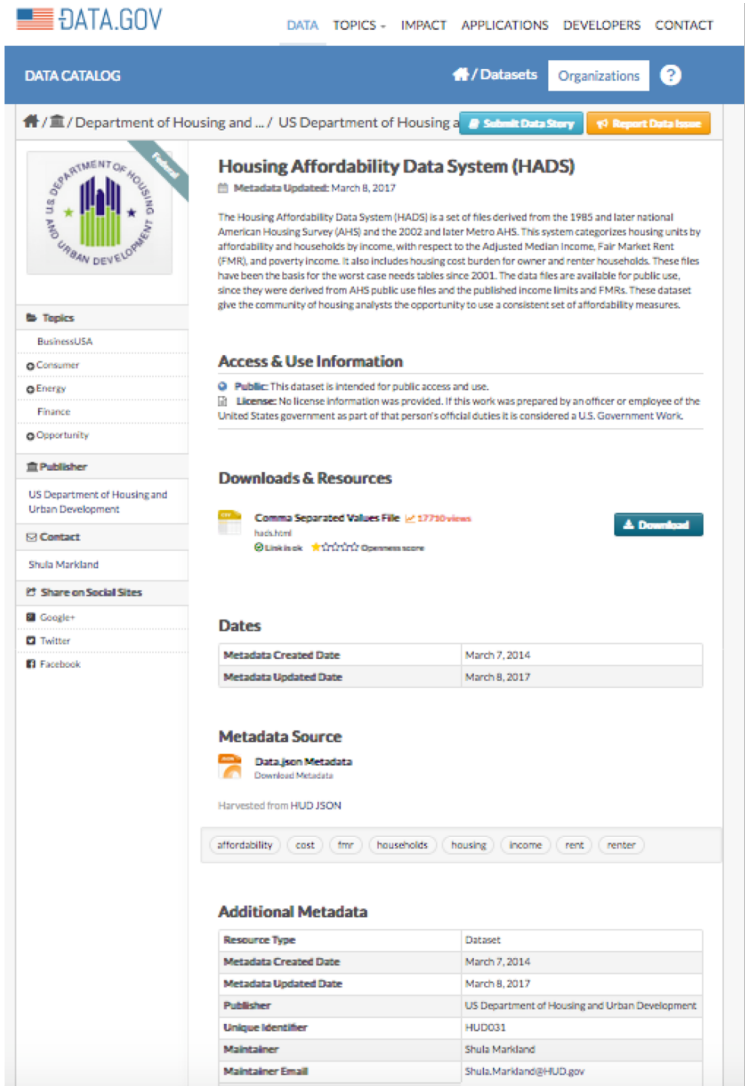
A lot of Open Data is not Linked Data

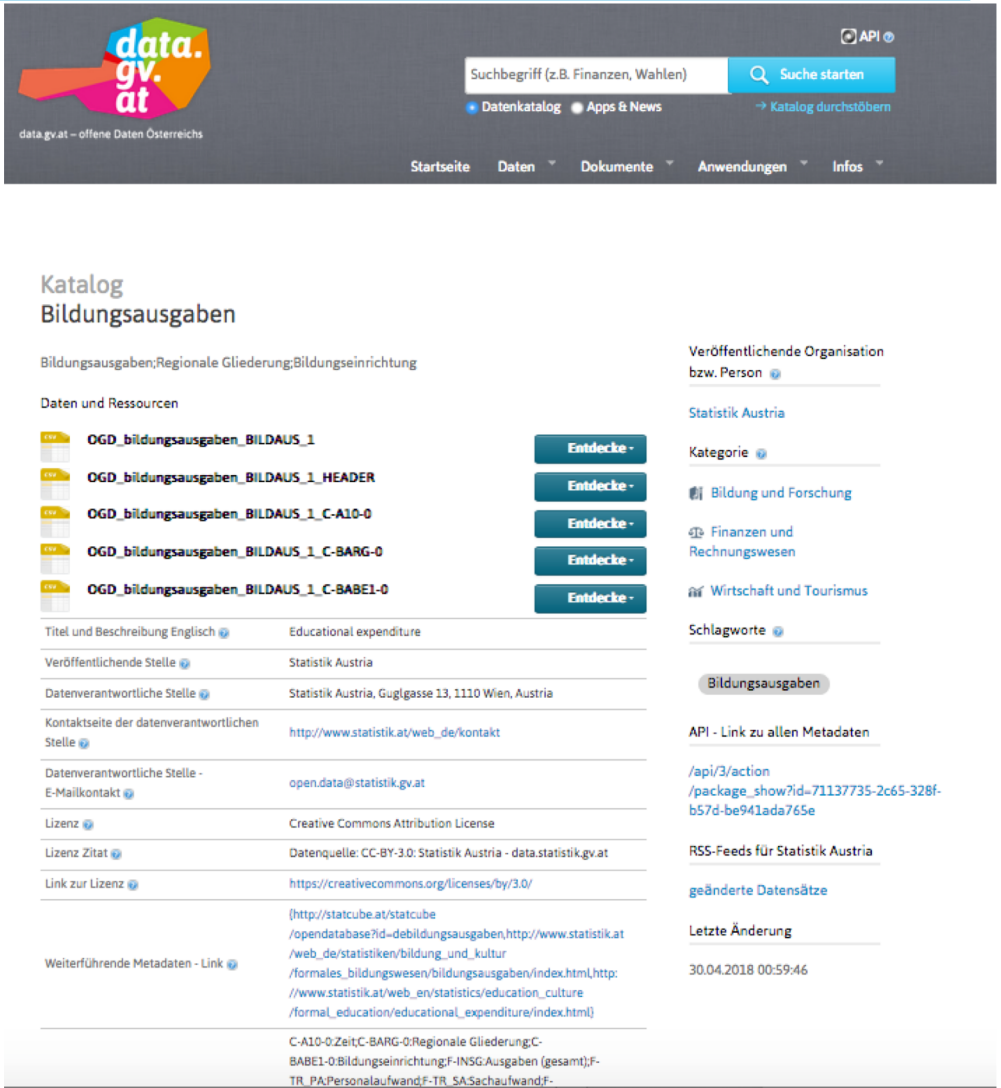
- Cities, International Organizations, National and European Portals, Int'l. Conferences:

We are aware of currently
260 active such portals worldwide.

Different portals...

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS





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- (Meta-)data becomes **stale/offline/outdated**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

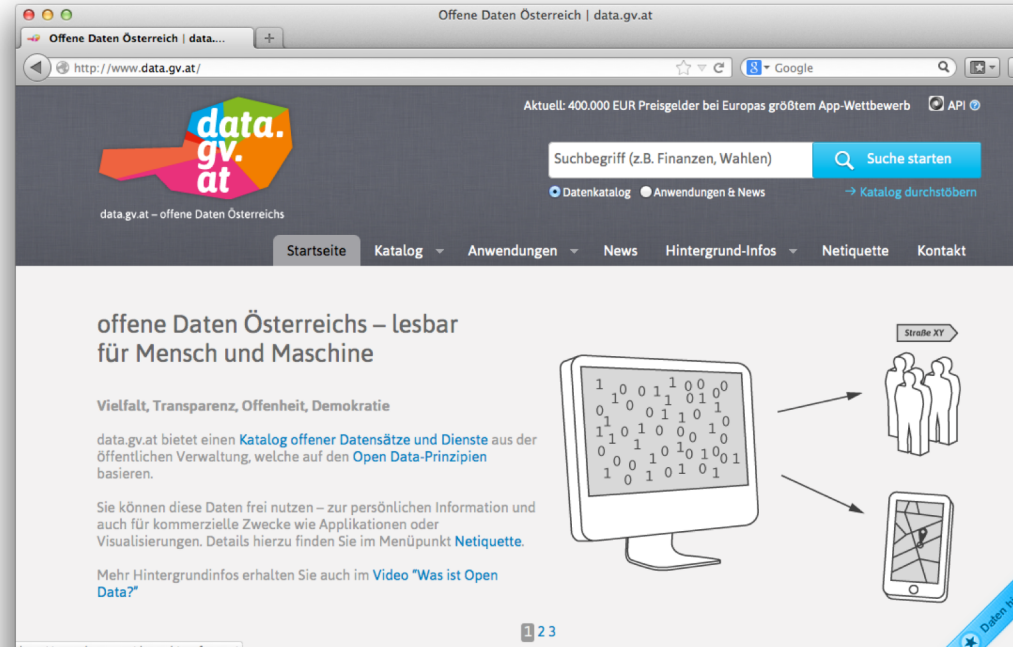
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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**



Our solution:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

EFMD EQUIS ACCREDITED

1) Monitoring and QA over evolving data portals

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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

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

6/2016 [3]:
- 260 portals
- 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

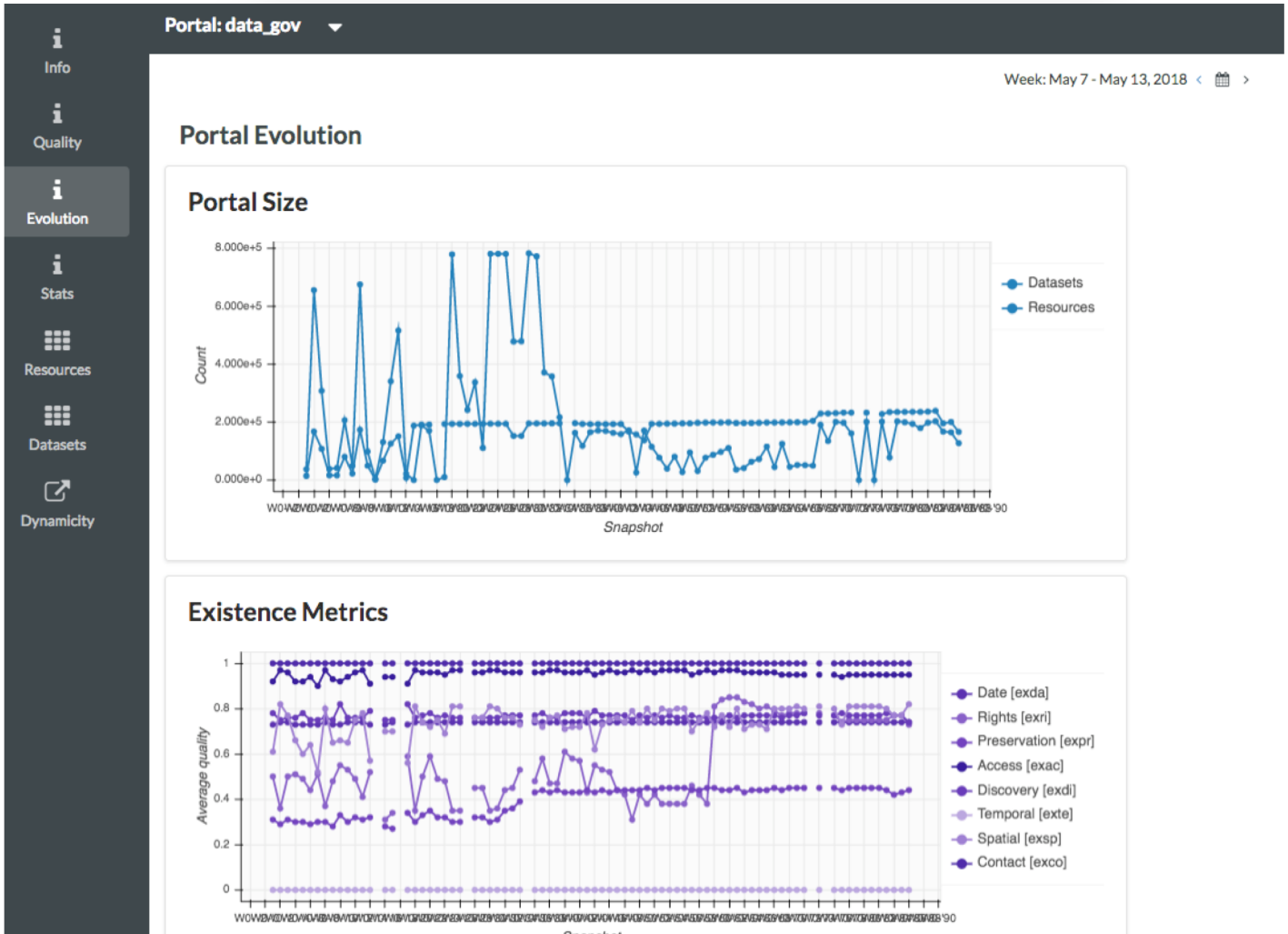
[2] Quality assessment & evolution of open data portals. In: International Conference on Open and Big Data, Rome, Italy (2015)

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

Demo:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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



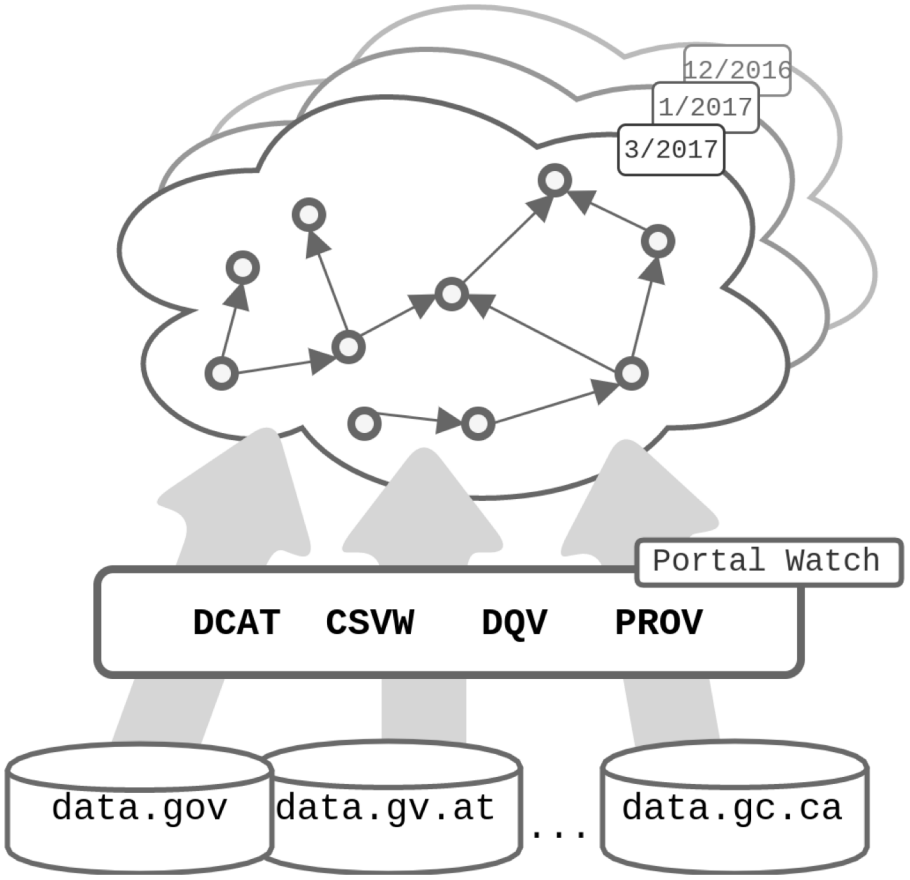
EFMD EQUIS ACCREDITED

2) Mapping to Standard vocabularies & Linked Data

- Mapping & Heuristic Enrichment
  - DCAT
  - PROV
  - CSVW
  - Schema.org
- Enable uniform access:

  →SPARQL endpoint

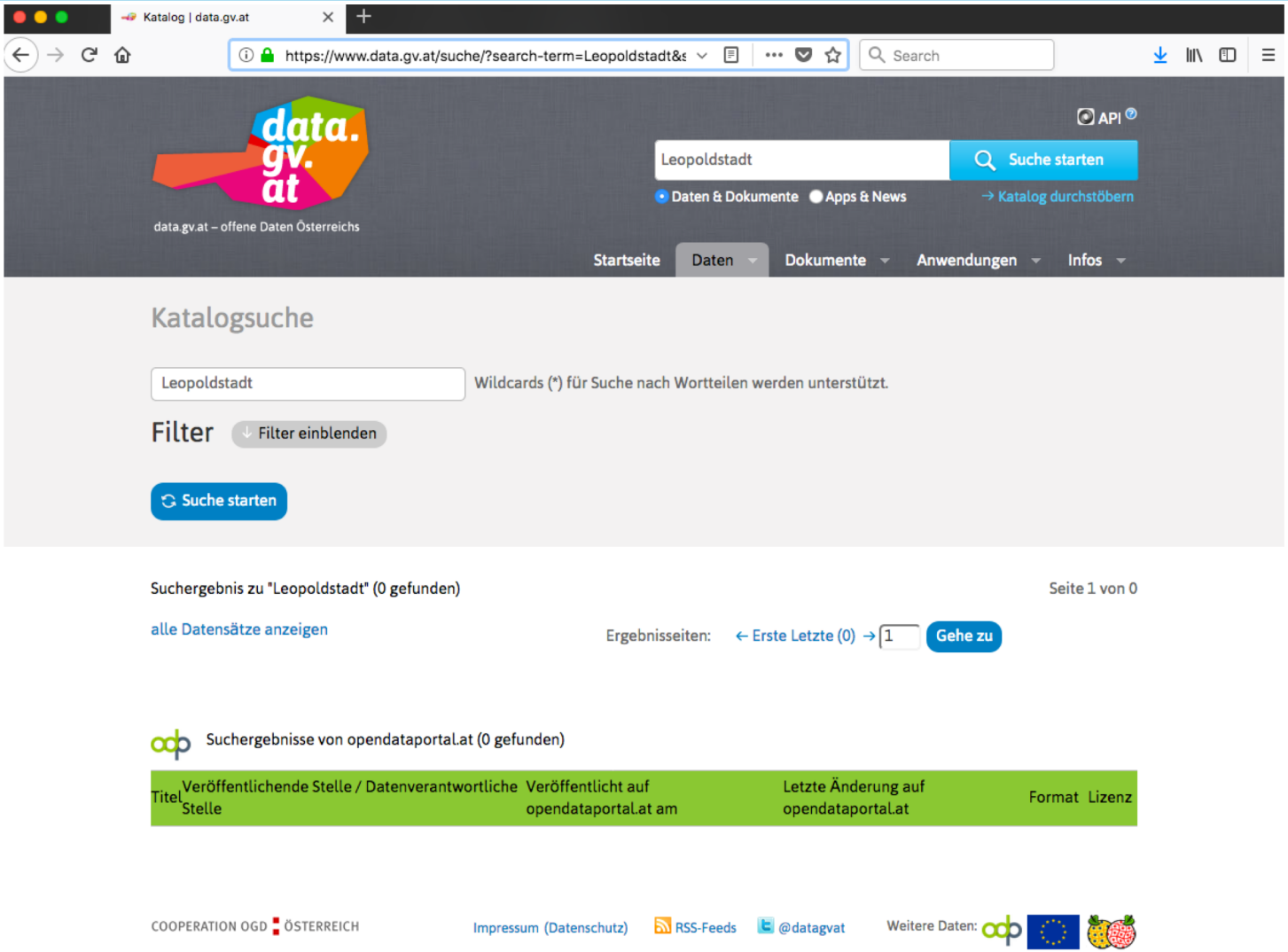
  → Linked Data & Memento Protocol



[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?

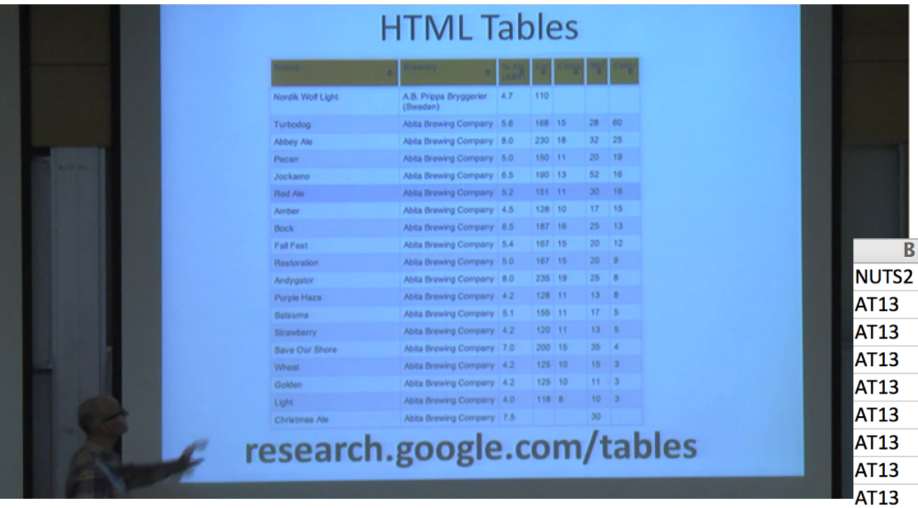


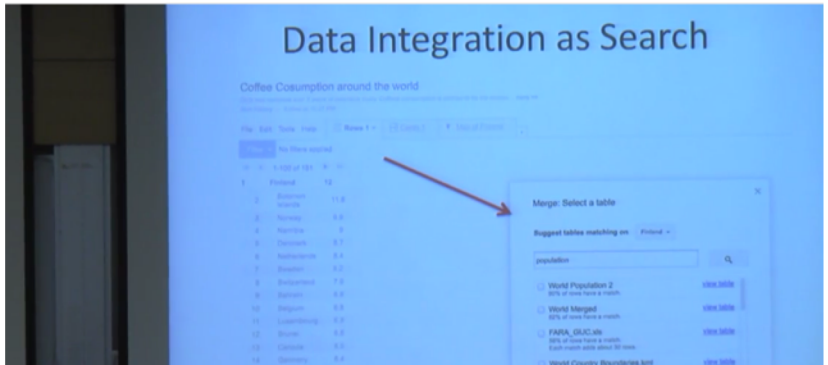
Why is Search in Open Data a problem?

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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"

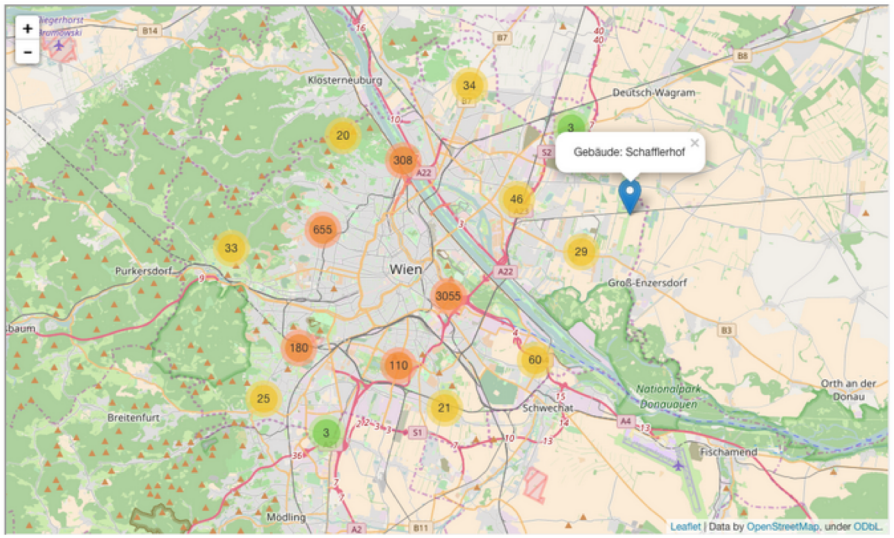
c) Open Data is not properly geo-referenced

Some starting points:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- 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 Neumaier[1], Jürgen Umbrich[1], Josiane Xavier Parreira[2], and Axel Polleres[1]

[1] Vienna University of Economics and Business, Vienna, Austria
[2] Siemens AG Österreich, Vienna, Austria

Abstract. With the success of Open Data a huge amount of tabular data sources became available that could potentially be mapped and linked into the Web of (Linked) Data. Most existing approaches to "semantically label" such tabular data rely on mappings of textual information to classes, properties, or instances in RDF knowledge bases in order to link – and eventually transform – tabular data into RDF. However, as we will illustrate, Open Data tables typically con-

Enabling Spatio-Temporal Search in Open Data

Sebastian Neumaier[a,1], Axel Polleres[a,b,c,2]

[a] *Vienna University of Economics and Business, Vienna, Austria*
[b] *Complexity Science Hub Vienna, Austria*
[c] *Stanford University, CA, USA*

Abstract

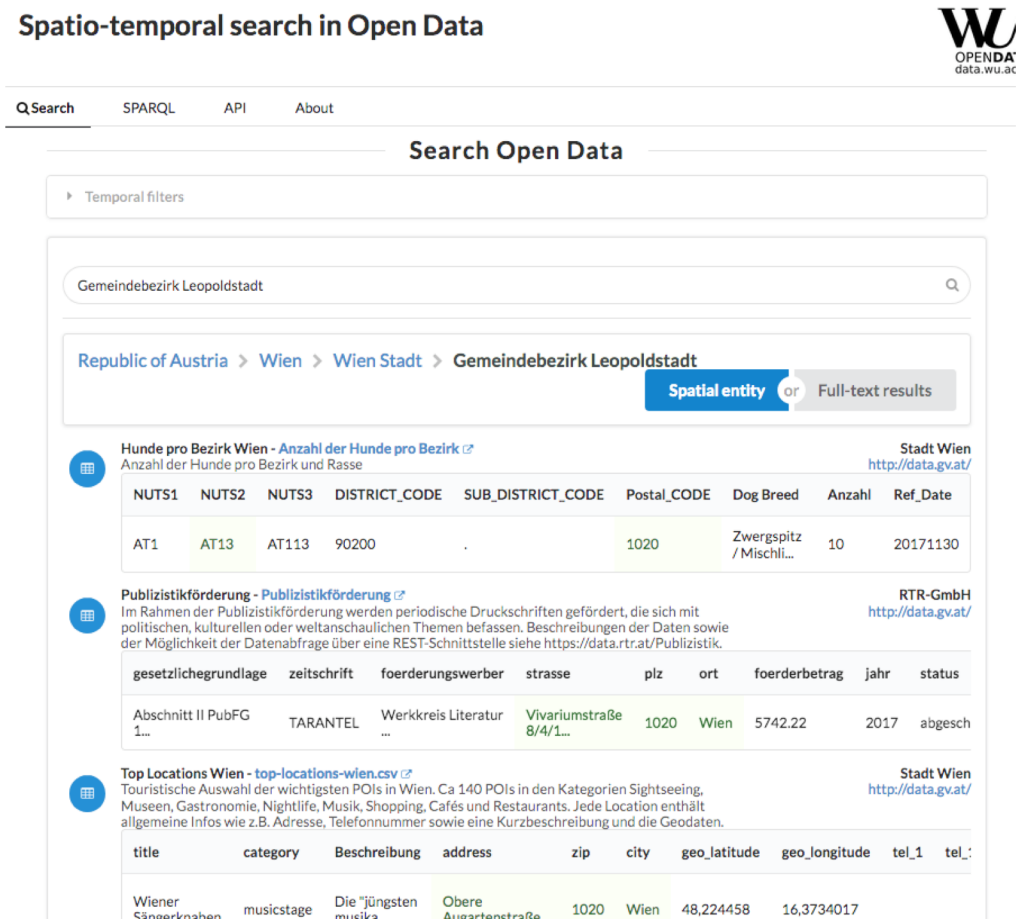
Intuitively, most datasets found in Open Data are organised by spatio-temporal scope, that is, single datasets provide data for a certain region, valid for a certain time period. For many use cases (such as for instance data journalism and fact checking) a pre-dominant need is to scope down the relevant 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.wu.ac.at/odgraphsearch/`

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

Demo:

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

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



Idea:

Link OD to geospatial and temporal Knowledge bases:

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

EFMD EQUIS ACCREDITED

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)

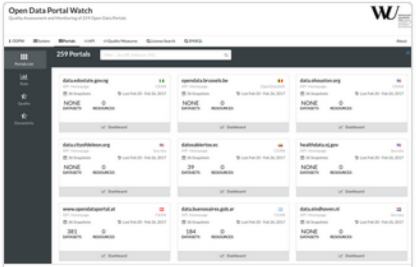
Projects

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

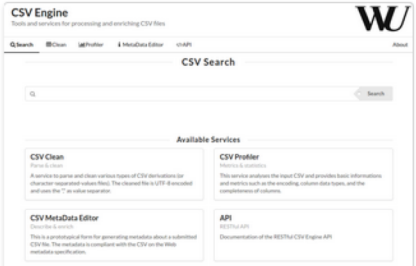


Open Data Portal Watch

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

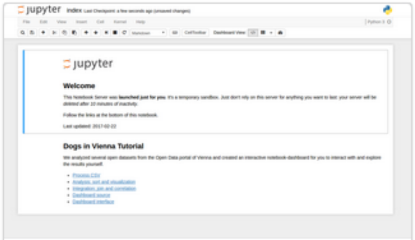
Search & enrich CSVs

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.

EFMD EQUIS ACCREDITED