

From Open Data Portals to Open Data-Ecosystems

Experiences & thoughts on how to proceed from
monitoring and analyzing Open Data Portals

WU

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Axel Polleres

Institute for Data, Process & Knowledge Management

<http://www.polleres.net/>

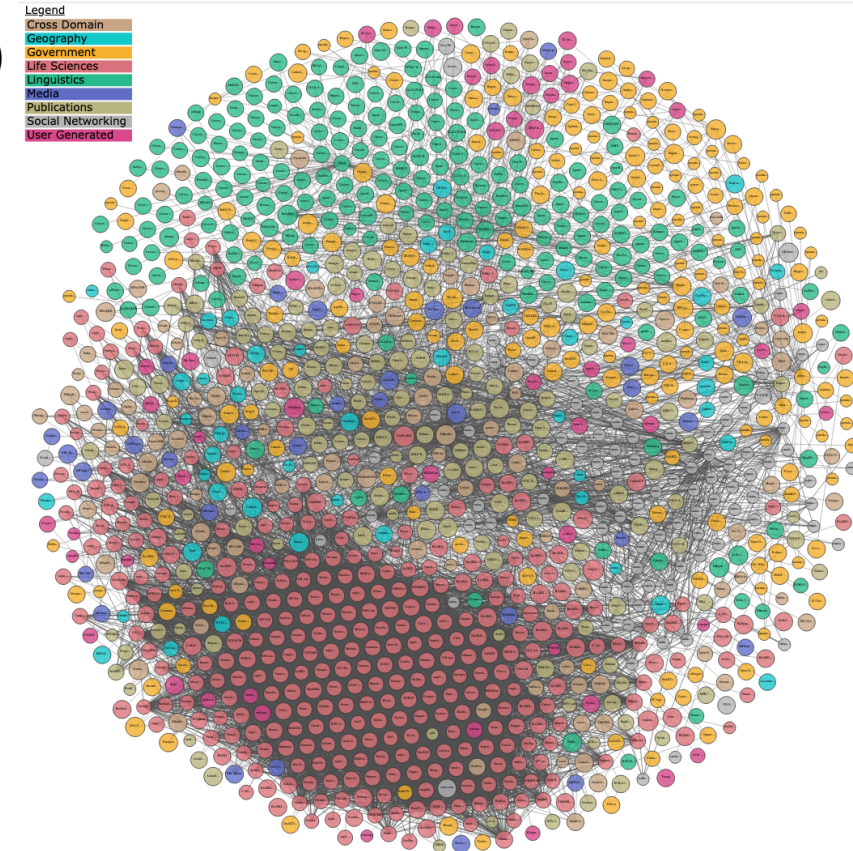


My background

- from Open Data-Ecosystems
- to Open Data

My background - The "Web of Data"

- past ~15 years research in Semantic Web and Linked (Open Data)
 - Standards (RDF, SPARQL, etc.) for
 - machine-readable
 - decentralized
 - Data Publishing & Processing
- ... which has led to a network of distributed, *partially* connected *Knowledge Graphs*

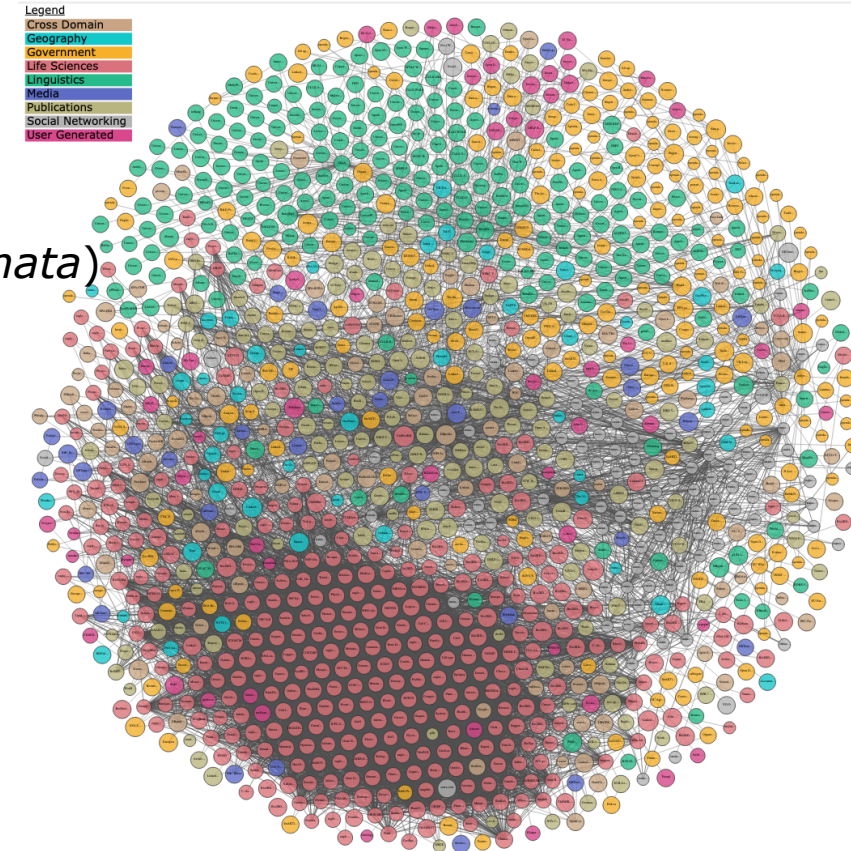


Armin Haller, Javier D. Fernández, Maulik R. Kamdar, and Axel Polleres. What are links in linked open data? a characterization and evaluation of links between knowledge graphs on the web. *ACM Journal of Data and Information Quality (JDIQ)*, to appear, accepted for publication, 2020. Pre-print available at <https://epub.wu.ac.at/7193/>.

My background - The "Web of Data"

- state of affairs:

- + common data and meta-data format (RDF)
- + Ontologies are reused widely (~ i.e., widely reused meta—data *schemata*)
(while maybe less expressive than originally expected) **Schema.org**
- + fairly scalable methods for distributed processing and querying
(still a very active, stimulating academic research area!)
- + widely used "centernodes" around which knowledge can "chrystalize"



The Linked Open Data Cloud from lod-cloud.net

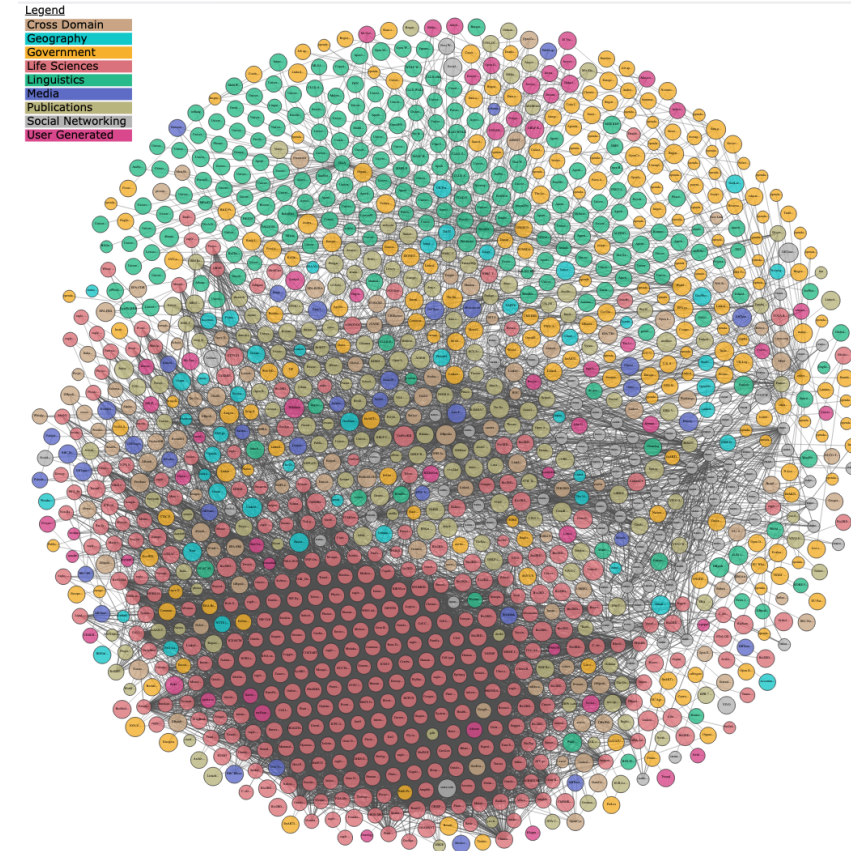
Armin Haller, Javier D. Fernández, Maulik R. Kamdar, and Axel Polleres. What are links in linked open data? a characterization and evaluation of links between knowledge graphs on the web. *ACM Journal of Data and Information Quality (JDIQ)*, to appear, accepted for publication, 2020. Pre-print available at <https://epub.wu.ac.at/7193/>.

My background - The "Web of Data"

- state of affairs:
 - lack of instance data links
(28% of datasets do not link to external instances)
 - broken class and property links
(schema definitions disappear, or exist multiple times,
hard to maintain by academics)
 - plethora of data and metadata formats, (still) missing
publishing best practices

GO FAIR ? <https://www.go-fair.org/fair-principles/>

Findability, (machine-) **A**ccessability, **I**nteroperability and **R**euse/Reusability only partially given...



The Linked Open Data Cloud from lod-cloud.net



My background – Meanwhile in parallel...

- We got quite excited about Open (Government) Data developments!

Open Data portals...

The screenshot shows the DATA.GOV portal with the following details for the Housing Affordability Data System (HADS):

- Metadata Updated:** March 8, 2017
- Access & Use Information:** Public. This dataset is intended for public access and use. License: No license information was provided.
- Downloads & Resources:** Comma Separated Values File (17730 views)
- Dates:** Metadata Created Date: March 7, 2014; Metadata Updated Date: March 8, 2017
- Additional Metadata:** Resource Type: Dataset; Metadata Created Date: March 7, 2014; Metadata Updated Date: March 8, 2017; Publisher: US Department of Housing and Urban Development; Unique Identifier: HUD031; Maintainer: Shula Markland; Maintainer Email: Shula.Markland@HUD.gov

The screenshot shows the data.gv.at portal with the following details for the 'Bildungsausgaben' dataset:

- Suchbegriff (z.B. Finanzen, Wahlen):** Suchbegriff
- Suche starten** button
- Katalog durchstöbern** button
- Startseite** | **Daten** | **Dokumente** | **Anwendungen** | **Infos**
- Katalog Bildungsausgaben**
- Veröffentlichende Organisation bzw. Person:** Statistik Austria
- Kategorie:** Bildungsausgaben; Regionale Gliederung; Bildungseinrichtung
- Daten und Ressourcen:** OGD_bildungsausgaben_BILDAUS_1, OGD_bildungsausgaben_BILDAUS_1_HEADER
- Entdecke** button

Uptake and use mostly in the form of Apps, Data journalism, etc. using single datasets.

The screenshot shows the metadata page for 'Bildungsausgaben' with the following details:

- 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://stabcube.at/stabcube/opedatabase?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]
- 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

Uniform metadata, accessible via common APIs (mostly JSON, rather than RDF).

The screenshot shows the Eurostat portal with the following details for the 'CITIES (URBAN AUDIT)' database:

- Overview 'Regions & cities' Background**
- City statistics (urb)**
 - Cities and greater cities (urb_cgc)**
 - Population on 1 January by age groups and sex - cities and greater cities (urb_cp)
 - Population structure - cities and greater cities (urb_cpopstr)
 - Population by citizenship and country of birth - cities and greater cities (urb_cpoc)
 - Fert
 - Livi
 - Edu
 - Cult
 - Lab
 - Eco
 - Trar
 - Envi
 - Functional**
 - Pop
 - Pop
 - Pop
 - Fert
 - Livi
 - Edu
 - Trar
 - Envi
 - Percep
- Population on 1 January by age groups and sex - cities and greater cities**
 - Title:** Population on 1 January by age groups and sex - cities and greater cities
 - Code:** urb_cpop1
 - Last update of data:** 23/05/2019
 - Last table structure change:** 15/04/2019
 - Number of values:** 366311
- Data availability - by country:**

Code	Label	Oldest data
-	overall	1990

Can (Open) Data portals be a more viable basis for making the Web of Data a reality?

- Big hopes (also for us as researchers!) & political push!



May 2009 – launch data.gov

March 2010 – Data.gov.uk launched

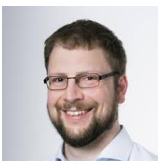


December 2011 – Neelie Kroes EU VP for Digital Agenda Data announced
EU Open Data Strategy

May 2013 – Obama's [executive order](#) makes open data default for gov.information

- Stefan Bischof, Axel Polleres, and Simon Sperl. City data pipeline - a system for making open data useful for cities. In *I-SEMANTICS 2013 Posters & Demonstrations Track*, volume 1026 of *CEUR Workshop Proceedings*, pages 45--49, Graz, Austria, September 2013. CEUR-WS.org. [[.pdf](#)]
- Jürgen Umbrich, Sebastian Neumaier, and Axel Polleres. Quality assessment & evolution of open data portals. In *IEEE International Conference on Open and Big Data*, Rome, Italy, August 2015.

Can we use Open Data as a Web of Data?



Stefan Bischof, Axel Polleres, and Simon Sperl. City data pipeline - a system for making open data useful for cities. In *I-SEMANTICS 2013 Posters & Demonstrations Track*, volume 1026 of *CEUR Workshop Proceedings*, pages 45--49, Graz, Austria, September 2013. CEUR-WS.org. [[.pdf](#)]

[PhD thesis 2017](#)

- Can we collect and use Open Data to answer relevant questions?
 - e.g. developments of Cities (in Europe, worldwide)*

City

Answer:

- partially,
- manual integration of a handful of sources
- lots of missing data
- We can develop new methods for automatic, scalable integration and model building (using Semantic Web technologies and Machine Learning!)

Hope:

- The more data becomes available, the better (data quality & amount) it will work!***

3. Anal

2. Sen

1. Per

City Open Data Portals:
Berlin, Vienna, London, ...

Other Open Data Sources:
DBpedia THE WORLD BANK Open Data UNdata
A world of information

Hope: The more Open data becomes available, the better it will work!

Does Open data develop into a viable Data-Ecosystem?



Jürgen Umbrich, Sebastian Neumaier, and Axel Polleres. Quality assessment & evolution of open data portals. In *IEEE International Conference on Open and Big Data*, Rome, Italy, August 2015.

Can we monitor Open Data development and quality?

Can we automatically collect and integrate Open Data?

Can we make Open Data more **FAIR** *automatically*?

Hope: The more Open data becomes available, the better it will work!

Does Open data develop into a viable Data-Ecosystem?



Jürgen Umbrich, Sebastian Neumaier, and Axel Polleres. Quality assessment & evolution of open data portals. In *IEEE International Conference on Open and Big Data*, Rome, Italy, August 2015.

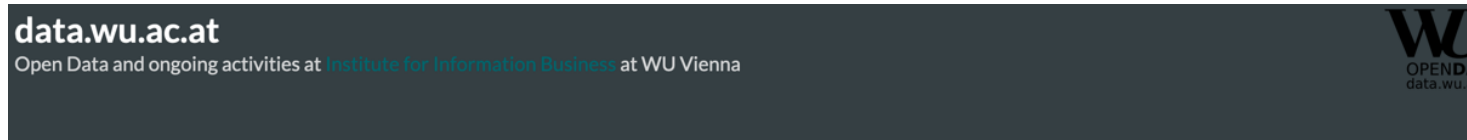
Can we monitor Open Data development and quality?

Can we automatically collect and integrate Open Data?

Can we make Open Data more **FAIR** automatically?

Monitoring and Integrating Open Data

- Open Data Quality and Usability
- Semantic enrichment of Data Catalogs
- Dataset Search
- <https://data.wu.ac.at/>
- <http://adequate.at/>
- <https://www.communidata.at/>



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

Applications

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

Enabling Spatio-temporal search in Open Data
Spatio-temporal labelling of datasets

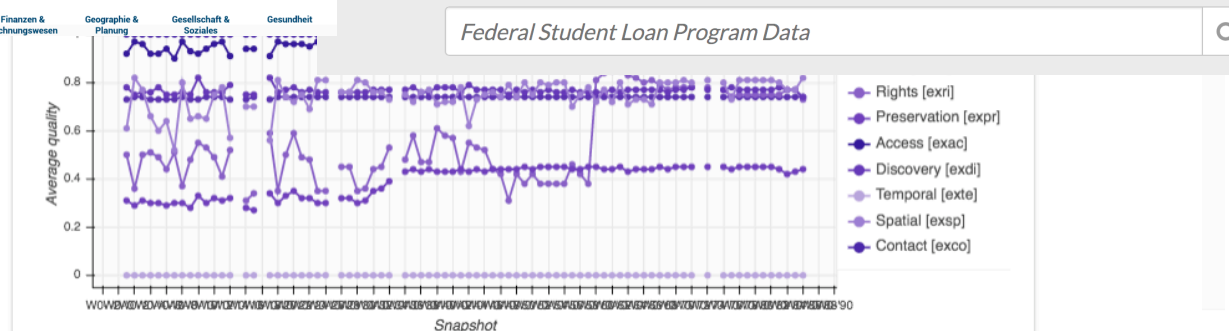
Annotation of datasets found on Open Data catalogs and their respective metadata descriptions to enable structured, spatio-temporal search.

Open Data Portalwatch (collecting and assessing metadata quality over time):

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

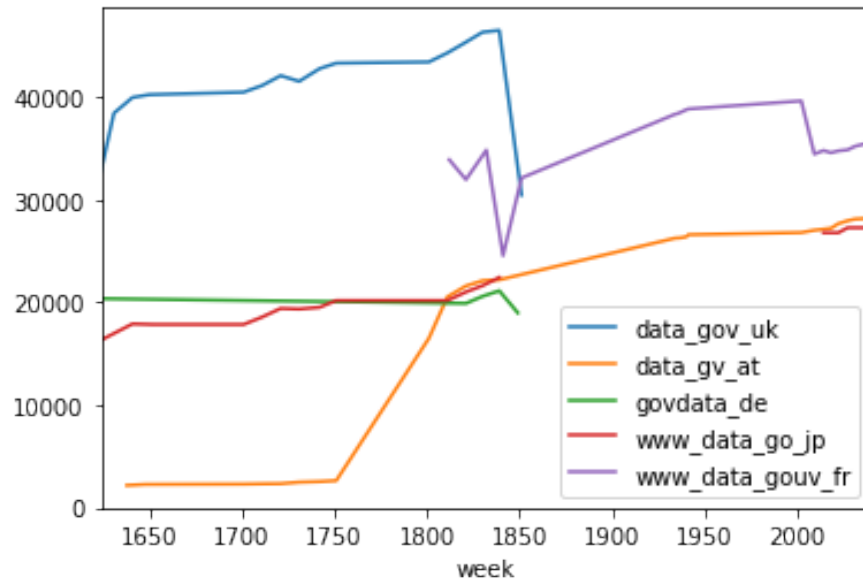
The collage shows several open data portals. On the left, 'Open Government Data Wien' features a 'Datenkatalog: Umwelt' section with categories like 'Bevölkerung', 'Bildung', and 'Umwelt'. In the center, 'UNdata' displays 'Open Data Berlin'. To the right, 'london.gov.uk' has a 'Welcome to the London Datastore' page. Below these, 'data.gov' is prominently featured with a blue header and the text: '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.' Below this text, it says 'GET STARTED SEARCH OVER 170,714 DATASETS'.

The screenshot shows the 'Open Data Österreich' portal. It has a blue header with the title and a search bar. Below the search bar, it displays statistics: '24.818 Datensätze', '465 Anwendungen', and '60 Organisationen'. At the bottom, there are icons for various categories: Arbeit, Bevölkerung, Bildung & Forschung, Finanzen & Rechnungswesen, Geographie & Planung, Gesellschaft & Soziales, and Gesundheit.



Does actually more (meta-)data become available?

- no unique answer...
 - affected by:
 - failed crawls
 - wrong metadata
 - portal maintenance
 - changing interfaces/APIs
 - merged/consolidated datasets
 - last but not least: (our own) crawler maintenance/reengineering
 - ... but
 - Therein lies the problem!
 - Why is the metadata not made available historically in first place?
 - Last, but not least: what should we measure?
 - #datasets? #resources?
 - #size_of_datasets/resources?
 - how? in fact:
 - advertised sizes differ from downloadable file sizes,
 - compression, etc.
 - "current" vs "incremental" datasets
 - Help? Google?



Does actually more (meta-)data become available?

■ more sample issues(?) ...

Dear Mr Harm,

Thank you for contacting European Data Portal Helpdesk.

The catalogue was deleted by accident, we apologise. Harvester is running again and restoring all datasets.

Please let me know if you need further assistance from our services.

Best regards,

Laia Guinovart
EDP Helpdesk



EDP Helpdesk created an issue

Welcome to European Data Portal Helpdesk / **DESK-7965**
[data.gov.at datasets are missing](#)

Issue Type: Report a bug or an error

Assignee: [EDP Helpdesk](#)

Created: 26/Aug/20 11:48 AM

Priority: Medium

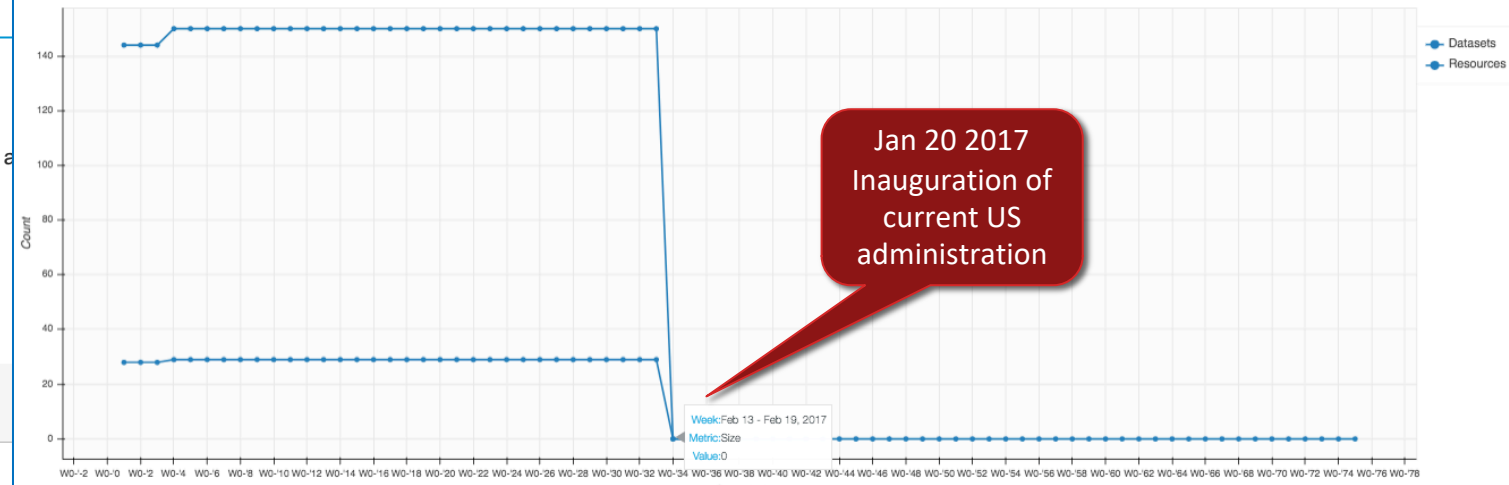
Reporter: [EDP Helpdesk](#)

Dear EDP-team,
can you please provide details why the datasets from [data.gov.at](#) are not available anymore at the moment? (<https://www.europeandataportal.eu/data/datasets?catalog=data-gv-at&showcataloguedetails=true&locale=en>)

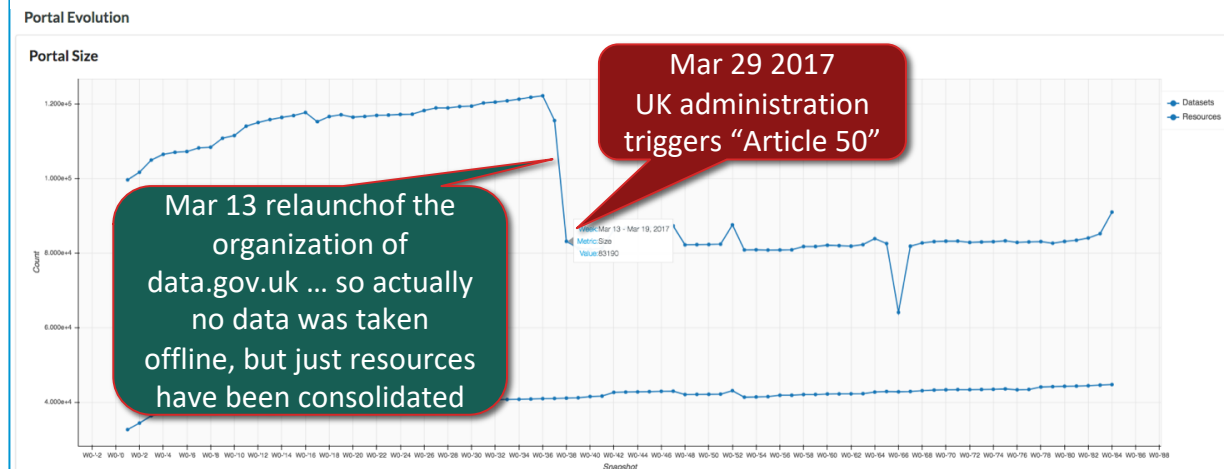
This is rather critical for us and a prompt fix would be much appreciated!

best,
Robert

http://data.wu.ac.at/portalwatch/portal/open_whitehouse_gov/1804/



http://data.wu.ac.at/portalwatch/portal/data_gov_uk/1815/evolution



- ... Lessons learnt:
 - **meta-data quality** and **portal QoS** needs to be monitored **at the portals**
 - **metadata integration** (CKAN, SOCRATA, DCAT, schema-org) needed, pre-dating/feeding into Google's dataset search
 - accumulated a large corpus of Open data portal metadata and data to learn from and do further research... but a real **sustainability strategy and a WebArchive for OpenData** are needed!

Meta-data quality and portal QoS best need to be monitored at the portals

- Quality-check button (regular automated quality checks and feedback option for users) on the portal



data.gv.at - Open Data Österreich Neue Leitfäden zu Open Data Analyse & Open Data Governance

Startseite Daten Dokumente Anwendungen Infos News

News

Qualität der Datensätze ab sofort Adequate-checked

18.5.2018 ADEQUATE, Datenprüfung, Datenqualität, Verbesserung der Metadaten

Im Rahmen des Projekts ADEQUATE - Open Data wird an Mechanismen zur Evaluierung und semi-automatischen Verbesserung von Metadaten und Daten im Bereich Open (Government) Data (siehe: <https://www.adequate.at/>) gearbeitet.

Alle auf data.gv.at bzw. opendataportal.at publizierten Datensätze werden hierbei regelmäßig automatisiert überprüft, und jeder Datensatz (der analysiert wurde) bietet ab sofort den Button: **ADEQUATE checked im** Bereich der Metadaten, welcher interessierte Userinnen und User zur entsprechenden Analyse des Datensatzes auf www.adequate.at führt.

ADEQUATE Checked ✓

Veröffentlichende Organisation
bzw. Person

Das ADEQUATE Service bietet damit

- Benutzerinnen und Benutzern zusätzliche Information pro Datensatz,
- interessierten Community-Mitgliedern die Option, direkt an einer Kopie des Datensatzes zu arbeiten und
- Publishern von Datensätzen die Möglichkeit, mittels dem ADEQUATE Service mögliche Verbesserungen in den Metadaten und Daten zu erkennen und am jeweiligen Portal (data.gv.at und/oder ODP) durchzuführen.

Thurnay, L., Lampoltshammer, T. J., Neumaier, S., & Knap, T. (2019). ADEQUATE: A Community-Driven Approach to Improve Open Data Quality. In W. Abramowicz & A. Paschke (Eds.), *Business Information Systems Workshops* (pp. 555–565). Springer International Publishing. https://doi.org/10.1007/978-3-030-04849-5_48

Metadata integration (CKAN, SOCRATA, DCAT, schema-org) pre-dating/feeding into Google's dataset search

Google search results for "Instrucciones particulares de Mercaderies Perilloses". The search bar shows the query. Below the search bar are filters for "Last updated", "Download format", "Usage rights", "Topic", and "Free". Three datasets are found:

- Instruccions particulars de Mercaderies Perilloses** (datos.gob.es), updated 6 de jul. 2017. Includes a button "Explore at data.wu.ac.at" and a "csv" download option.
- Instruccions particulars de Mercaderies Perilloses** (data.wu.ac.at), updated Jul 6, 2017. Includes a "csv" download option.
- Instrucciones particulares de MP** (datos.gob.es), updated 6 Jul 2017.

The main result for "Instruccions particulars de Mercaderies Perilloses" shows:

- Dataset updated** Jul 6, 2017
- License** Attribution 4.0 (CC BY 4.0). License information was derived automatically.
- Area covered** <http://datos.gob.es/recurso/sector-publico/territorio/Autonomia/Cataluna>
- Description** Table indicating the APB's internal code and description of the different specific transport and storage requirements for HG on port premises.

Open Data Portal Watch: Mapping and export of Schema.org metadata descriptions for over 250 Open Data portals.

[Data Portal List](#)
[Spatio-Temporal Dataset Search](#)
[Data Portal Monitor](#)
[About](#)

Dataset Information

- Title** Instruccions particulars de Mercaderies Perilloses
- Description** Table indicating the APB's internal code and description of the different specific transport and storage requirements for HG on port premises.
- Publisher**
- Portal Information**
- Homepage** <http://datos.gob.es/>
- Country** ES
- Software** Underlying software: XMLDCAT

Schema.org

```
{
  "@context": "http://schema.org",
  "@id": "http://datos.gob.es/catalogo/ea0001302-instrucciones-particulares-de-mpi",
  "@type": "Dataset",
  "about": "http://datos.gob.es/kos/sector-publico/sector/transporte",
  "datePublished": "2017-07-06T00:00:00+02:00",
  "description": "Table indicating the APB's internal code and description of the different specific transport and storage requirements for HG on port premises.",
  "distribution": [
    {
      "@id": "http://datos.gob.es/catalogo/ea0001302-instrucciones-particulares-de-mpi/resource/e7569b49-4cfb-4f1d-8c9b-7ecb8798a590",
      "@type": "DataDownload",
      "contentSize": "8464.0",
      "contentUrl": "http://opendata.portdebarcelona.cat/FitxersEDI/Instruccions_particulars_de_Mercaderies_Perilloses/mi/PortBcn-Instru_Part.csv",
      "encodingFormat": "csv",
      "license": "https://creativecommons.org/licenses/by/4.0/deed.es",
      "name": "Instrucciones particulares de MP"
    }
  ],
  "inLanguage": "es",
  "includedInDataCatalog": {
    "@id": "http://datos.gob.es/",
    "@type": "DataCatalog",
    "description": "Underlying software: XMLDCAT",
    "spatialCoverage": "ES",
    "url": "http://datos.gob.es/"
  },
  "keywords": [
    "EDI"
  ],
  "name": "Instruccions particulars de Mercaderies Perilloses",
  "publisher": {
    "@id": "http://datos.gob.es/recurso/sector-publico/org/Organismo/EA0001302",
    "@type": "Organization"
  },
  "spatialCoverage": "http://datos.gob.es/recurso/sector-publico/territorio/Autonomia/Cataluna"
}
```

Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Lifting data portals to the web of data. In *10th Workshop on Linked Data on the Web (LDOW2017)*, Perth, Australia, April 2017. [[.pdf](#)]

Large corpus of OpenData portal metadata & data

Open Data Portal Watch

The Open Data Portal Watch framework is a scalable quality assessment and evolution monitoring framework for Open Data (Web) portals. We harvest the metadata of 280 Web catalogues and perform a quality assessment of the metadata along 6 dimensions and 19 metrics.

Available Content

Quality Metrics

6 Dimensions 19 Metrics

More information about our quality dimensions and metrics.

Portals List

280 registered Portals

Basic information such as the country, software and the number of datasets.

SPARQL Endpoint

Structured Queries

Query all datasets and quality assessments at a SPARQL endpoint

API

RESTful API

Documentation of the RESTful ODPW API

Data dumps

All datasets as Turtle files

We provide the collected metadata for each portal and snapshot.

Statistics

Top 10 portals (by datasets)

Portal	Datasets	Resources
transparenz.hamburg.de	138911	232585
open.canada.ca	86390	217524
www.data.gouv.fr	35633	193159
offenedaten.de	28661	85350
data.gv.at	28199	41113
www.data.go.jp	27635	351328
data.opendatasoft.com	20042	86246
data.humdata.org	17690	276700
data.gov.gr	10552	25947
data.gov.ie	9976	28273

Top 10 file formats

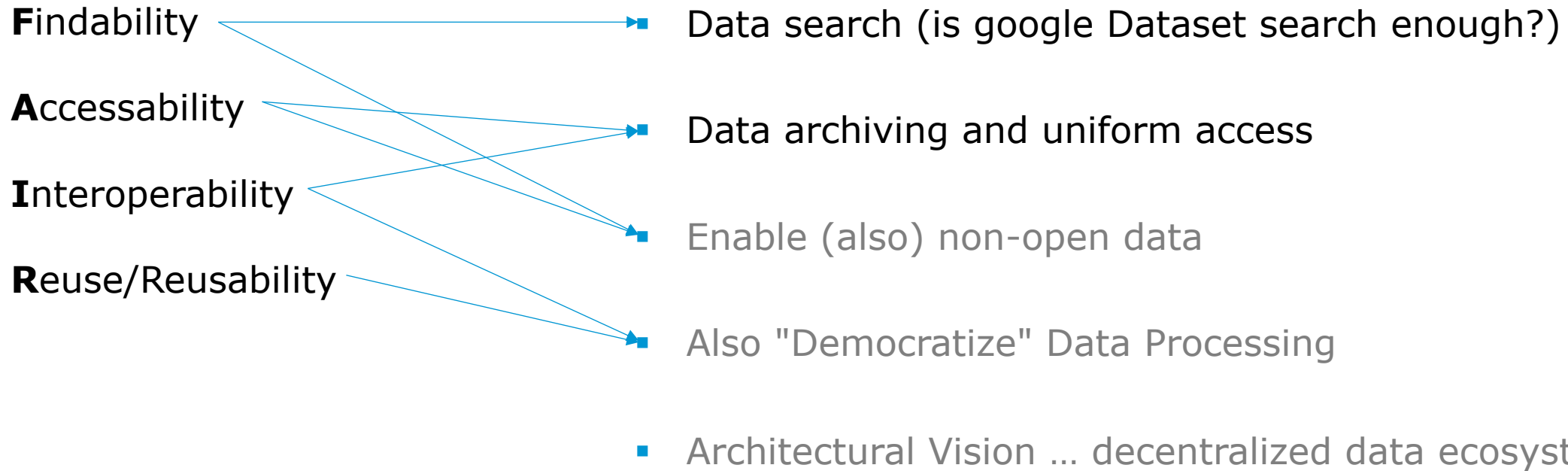
Format	Count	Percentage
CSV	253472	12.35%
PDF	249675	12.17%
HTML	240061	11.7%
TIFF	190876	9.3%
XLS	164917	8.04%
json	120369	5.87%
other	112875	5.5%
csv	107370	5.23%
xml	59073	2.88%
ZIP	45706	2.23%

Top 10 licenses

License	Count	Percentage
Open Government Licence - Canada	201722	15.51%
Creative Commons Namensnennung 4.0 International	71900	5.53%
Creative Commons Attribution	64030	4.92%
Creative Commons Attribution 4.0	55828	4.29%
Creative Commons Attribuzione	53193	4.09%
Creative Commons Atribuição	53193	4.09%
Creative Commons Αναφορά	53193	4.09%
dl-de-by-2.0	47278	3.64%
Public Domain	42996	3.31%
CC BY 4.0	34847	2.68%

- Johann Mitlöhner, Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Characteristics of open data CSV files. In *2nd International Conference on Open and Big Data*, August 2016. Invited paper. [[.pdf](#)]

What's next? What's missing for a FAIR (European) Data Ecosystem?



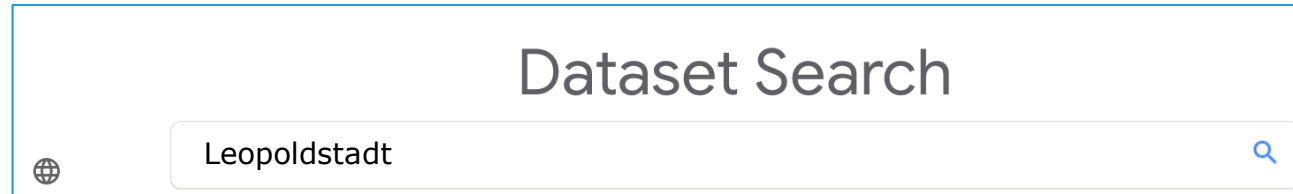
starting
points
are
there

beyond
current
OpenData
platforms

Data search – Is google Dataset search enough?

Findability,
Accessability
Interoperability
Reuse/Reusability

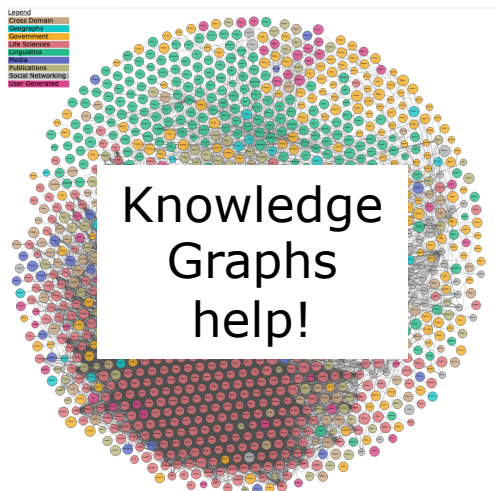
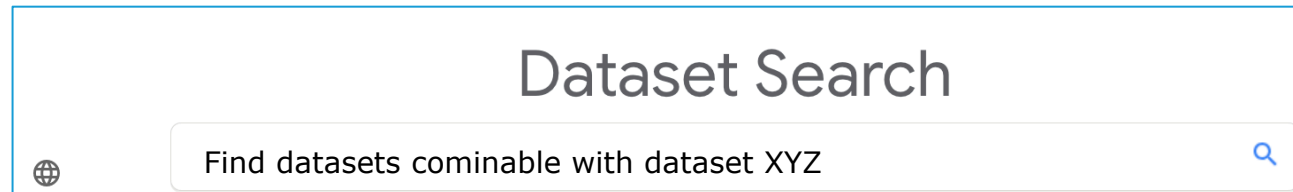
- Spatio-temporal search



- Search for organisations



- Find semantically comparable datasets
 - *(needs semantic labeling of datasets)*



Data search – Is google Dataset search enough?

Findability,
Accessibility
Interoperability
Reuse/Reusability

data.gv.at - Open Data Österreich

Suche

Leopoldstadt

Suchergebnisse - Daten & Dokumente

Stadtplan von Anton Behsel 1825 29.03.2019
Wien
Stadt Wien/Magistratsabteilung 8 - Wiener Stadt- und Landesarchiv
Georeferenzierter Stadtplan der Inneren Stadt und der Vorstädte (Bezirke innerhalb des Gürtels sowie...
WMS GIF JPEG PNG
Alle Datensätze Alle Dokumente

Suchergebnisse - Anwendungen

Keine Anwendungen gefunden!
Alle Anwendungen

Suchergebnisse - News

Keine News gefunden!
Alle News

Google Dataset Search Beta

Leopoldstadt

16 results found

statista Mitarbeiter der Erste Group nach Ländern 2018
de.statista.com

statista Cost-Income-Ratio der Erste Group bis 2018
de.statista.com

statista Betriebserträge und -aufwendungen der Erste Group bis 2018
de.statista.com

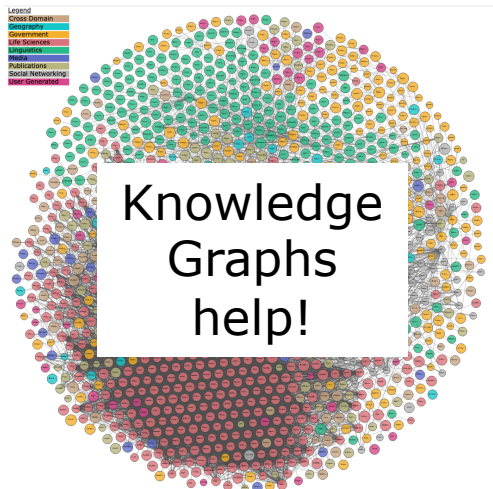
Mitarbeiter der Erste Group nach Ländern 2018
Explore at de.statista.com

Dataset provided by Statista

Time period covered 2018

Area covered Österreich

Description Die Statistik zeigt die Anzahl der Mitarbeiter der Erste Group nach Ländern im Jahr 2018. Die Erste Group Bank AG wurde im Jahr 1819 in der Leopoldstadt, einer Vorstadt Wiens, als 'Erste österreichische Spar-Casse' gegründet. Heute ist sie eine der größten Bankengruppen in Zentral- und



Solution:
(i) look into data, not only metadata!
(ii) link to Existing Knowledge Graphs!

Sebastian Neumaier and Axel Polleres. Enabling spatio-temporal search in open data. *Journal of Web Semantics (JWS)*, 55:21--36, March 2019. [[DOI](#) | [http](#)]

Data search – Is google Dataset

- Findability,
- Accessibility
- Interoperability
- Reuse/Reusability

Metadata

Tourismus - Ankünfte und Nächtigungen

Ankünfte und Nächtigungen in den oberösterreichischen Meldegemeinden

Daten und Ressourcen

Ankünfte und Nächtigungen in OÖ seit dem Jahr 2000

Veröffentlichende Stelle: Land Oberösterreich

Datenverantwortliche Stelle: Land Oberösterreich, Abteilung Statistik

Lizenz: Creative Commons Namensnennung

Link zur Lizenz: <https://creativecommons.org/licenses/by/4.0/>

Attributbeschreibung: NUTS2 -> Bundesland Oberösterreich; Gemeindegemeinde -> Erhebungsgemeinde lt. Tourismus; Abs.7: Städte und Gemeinden mit Gästenächtigungen im Kalenderjahr

CSV

NUTS2	Gemeindegemeinde	Jahr	Ankünfte
AT31	Linz	2000	3
AT31	Steyr	2000	
AT31	Wels	2000	
AT31	Altheim	2000	
AT31	Aspach	2000	
AT31	Auerbach	2000	
AT31	Braunau a. Inn	2000	

Diagram showing relationships: Upper Austria (Oberösterreich) -> Linz -> Disambiguate -> Linz

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 [Stad 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- [Land Oberösterreich](#)

Finanzgebarung 2015 [http://data.gv.at](#)

Finanzdaten der 444 oberösterreichischen Gemeinden

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

Legend: Cross Domain, Government, Life Sciences, Media, Publications, Social Networking, User Generated

OpenStreetMap

GeoNames

WIKIDATA

Knowledge Graphs help!

GeoKn

Data search – Is google Dataset search enough?

- F**indability,
- A**ccessability
- I**nteroperability
- R**euse/Reusability

- Spatio-temporal search

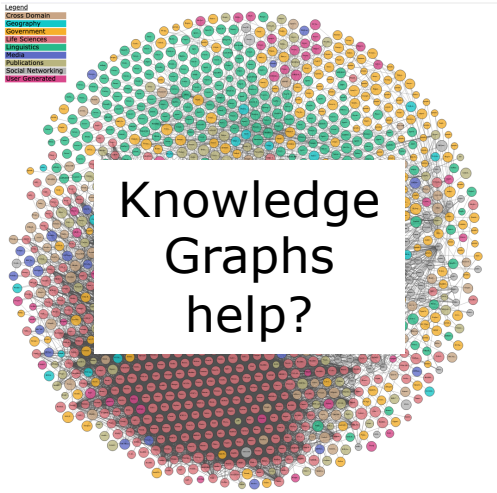


- Search for organisations



- Find semantically comparable datasets
 - *(needs semantic labeling of datasets)*

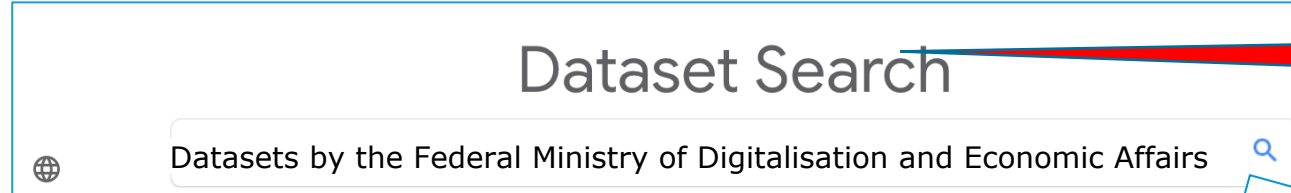
Currently still too hard



Data search – Is google Dataset search enough?

Findability,
 Accessibility
 Interoperability
 Reuse/Reusability

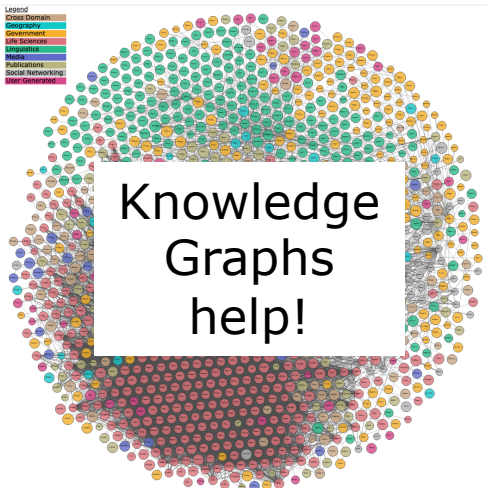
- Search for organisations



Currently still too hard

5 main challenges:

- (1) temporal changes in organizations and in the portal metadata,
- (2) lack of a base ontology for describing organizational structures+changes
- (3) metadata and KG data quality
- (4) multilinguality
- (5) disambiguating public sector organizations.



Knowledge Graphs help!

In 2000, the **Schüssel cabinet** merged the Ministry of Economy with the **Ministry of Labor**. Road and maintenance and a number of minor matters were assigned to the **Ministry of Traffic**; water hydraulic engineering went to the **Ministry of Agriculture**. The **Ministry of Environment** took over becoming the **Ministry of Social Affairs** in the process. The merger was undone in 2008. The Ministry's current name and portfolio dates to 2018.

Challenges of Linking Organizational Information in Open Government Data to Knowledge Graphs

Jan Portisch¹{0000-0001-5420-0863}, Omaiima Fallatah²{0000-0002-5406-9119}, Sebastian Neumaier^{3,4}{0000-0002-9804-8822}, Mohamad Yaser Jaradeh⁵{0000-0001-8777-2783}, and Axel Polleres^{1,4}{0000-0001-5670-1140}

¹ Data and Web Science Group, University of Mannheim, Germany
 jan@informatik.uni-mannheim.de

² Information School, The University of Sheffield, Sheffield, UK
 oofa11atah1@sheffield.ac.uk

³ Vienna University of Economics and Business, Austria
 {sebastian.neumaier, axel.polleres}@vu.ac.at

⁴ Complexity Science Hub Vienna, Austria

⁵ LIS Research Center, Leibniz University Hannover, Germany
 jaradeh@l3s.de

Abstract. Open Government Data (OGD) is being published by various public administration organizations around the globe. Within the metadata of OGD data catalogs, the publishing organizations (1) are not uniquely and unambiguously identifiable and, even worse, (2) change over time, by public administration units being merged or restructured. In order to enable fine-grained analyses or searches on Open Government Data on the level of publishing organizations, linking those from OGD portals to publicly available knowledge graphs (KGs) such as Wikidata and DBpedia seems like an obvious solution. Still, as we show in this position paper, organization linking faces significant challenges, both in terms of available (portal) metadata and KGs in terms of data quality and completeness. We herein specifically highlight five main challenges, namely regarding (1) temporal changes in organizations and in the portal metadata, (2) lack of a base ontology for describing organizational structures and changes in public knowledge graphs, (3) metadata and KG data quality, (4) multilinguality, and (5) disambiguating public sector organizations. Based on available OGD portal metadata from the Open Data Portal Watch, we provide an in-depth analysis of these issues, make suggestions for concrete starting points on how to tackle them along with a call to the community to jointly work on these open challenges.

Keywords: Open Data · Dataset Evolution · Entity Linking · Knowledge Graphs · Knowledge Graph Evolution.

Use of our corpus to analyse the problem:

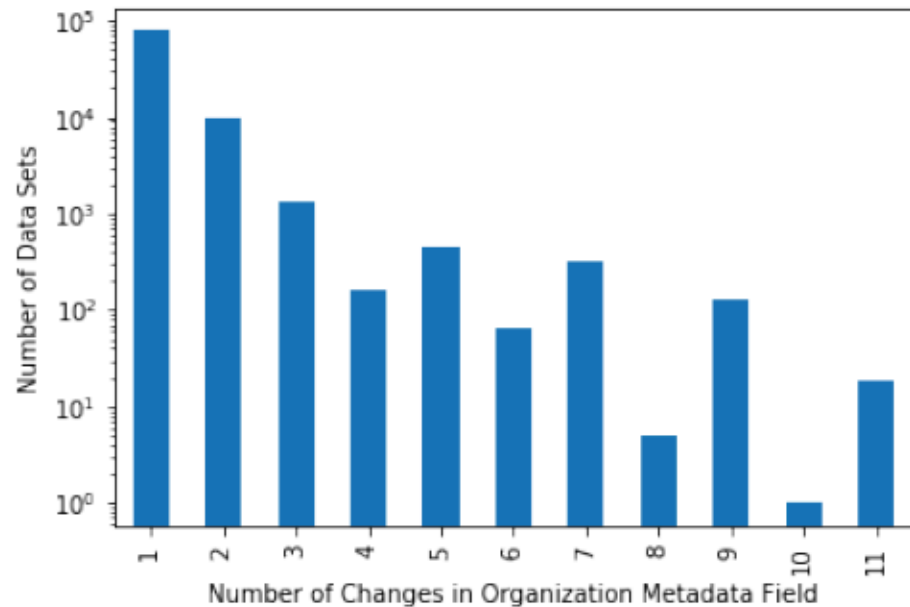


Fig. 1. The distribution of organizational changes for individual datasets. On the X-axis, the number of times an organization has been changed for a dataset on a particular data portal is shown while the overall frequency of such a change is reflected on the Y-axis. Note that the Y-axis is log-scale.

Use of our corpus to analyse the problem:

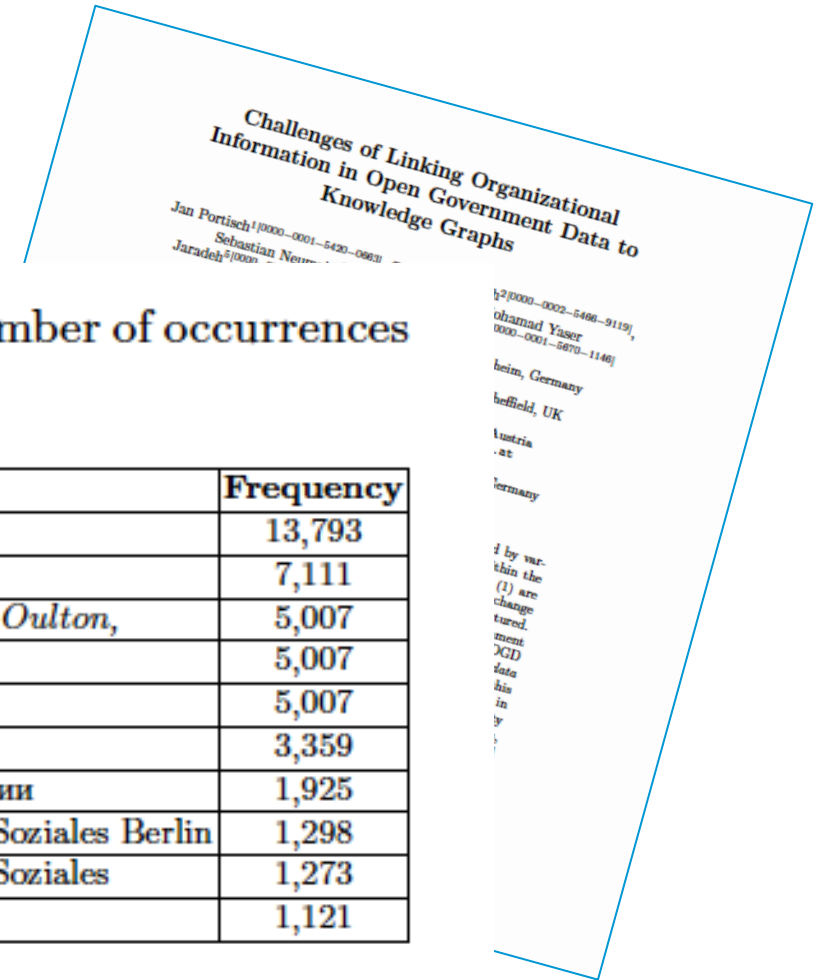


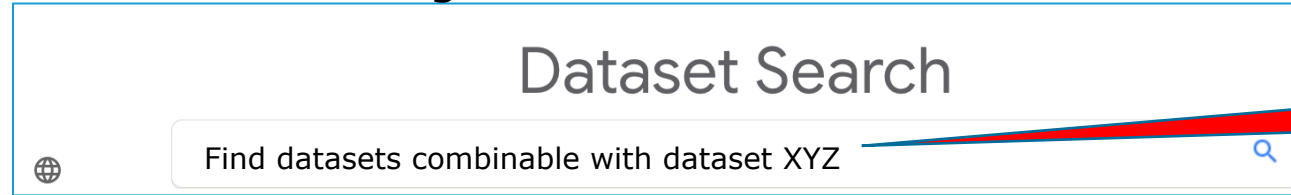
Table 2. Top 10 organization changes by label together with the number of occurrences of the change within datasets listed in ODPW.

Old Organization Label	New Organization Label	Frequency
NSGIC GIS Inventory (aka Ramona)	NSGIC GIS Inventory	13,793
Geoscience Australia	Corp	7,111
Daryl Beggs, Ruth Oulton, Benjamin Lang,	Benjamin Lang, Daryl Beggs, Ruth Oulton,	5,007
Daryl Beggs, Ruth Oulton, Benjamin Lang,	Engineering	5,007
Benjamin Lang, Daryl Beggs, Ruth Oulton,	Engineering	5,007
Ivan Begtin	Федеральная служба статистики	3,359
Archive bot	Национальный цифровой архив России	1,925
Senatsverwaltung für Gesundheit und Soziales	Senatsverwaltung für Gesundheit und Soziales Berlin	1,298
Senatsverwaltung für Gesundheit und Soziales Berlin	Senatsverwaltung für Gesundheit und Soziales	1,273
PAT S. Statistica	ISPAT	1,121

Data search – Is google Dataset search enough?

- Findability,
- Accessibility
- Interoperability
- Reuse/Reusability

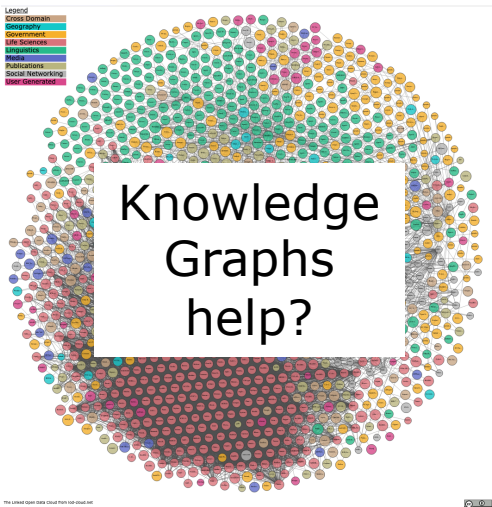
- Search for organisations



Currently still too hard

main challenges:

- (1) lack of coverage in current KGs
- (2) heterogeneous representations within KGs



SemTab: Semantic Web Challenge

This challenge aims at benchmarking systems dealing with the tabular data to knowledge graph matching problem.

Editions

- SemTab 2019: [\[description\]](#) [\[results\]](#) [\[datasets\]](#) [\[proceedings\]](#)
- SemTab 2020: [\[description\]](#) [\[results\]](#) [\[Synthetic dataset\]](#)

Publications

- Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, Kavitha Srinivas and Vincenzo Cutrona. *Results of SemTab 2020*. Vol-2775 of CEUR Workshop Proceedings. 2020. [\[Paper\]](#)
- Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen and Kavitha Srinivas. *SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems*. Extended Semantic Web Conference (ESWC). 2020. [\[Paper\]](#) [\[Presentation\]](#) [\[DOI\]](#)
- Vincenzo Cutrona, Federico Bianchi, Ernesto Jiménez-Ruiz and Matteo Palmonari. *Tough Tables: Carefully Evaluating Entity Linking for Tabular Data*. International Semantic Web Conference (ISWC). 2020. [\[Paper\]](#) [\[Presentation\]](#) [\[DOI\]](#)

In Wikidata the election class (Q40231) has 787 subclasses in total, they often represent parties and results differently

Data archiving and uniform access?

Findability,
Accessibility
Interoperability
Reuse/Reusability

Data archiving and uniform access

Crawl and index datasets, clean and index evolving, tabular datasets.
Search and query datasets over time!

- Total Corpus: 5.5 TB (CSV,JSON,XML,...)
- Latest Versions: 1.2 TB

The screenshot shows the 'Sparql Interface' of the OD Archiver. It includes a search bar with the URL 'https://archiver.ai.wu.ac.at/sparql', a 'Download files as ZIP' button, and a query editor. The query is as follows:

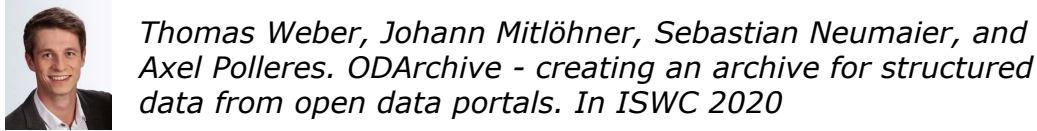
```

1 .:
2 PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
3 PREFIX arc: <https://archiver.ai.wu.ac.at/ns/csvw#>
4 PREFIX csvw: <http://www.w3.org/ns/csvw#>
5 PREFIX dcat: <http://www.w3.org/ns/dcat#>
6 PREFIX dc: <http://purl.org/dc/terms/1.1>
7
8 SELECT ?versionURL
9 FROM <https://archiver.ai.wu.ac.at/graph>
10 WHERE {
11   ?datasetURL arc:hasPortal ?Portal ; # ?datasetURL: the dataset URL
12   # ?Portal: a dcat:catalog
13   dc:hasVersion ?versionURL ; # ?versionURL: a crawled version
14   dcat:mediaType ?mediaType . # ?mediaType: media type as per
15
16   ?versionURL dc:issued ?dateVersion . # ?dateVersion: crawl time.
17
18   FILTER (?Portal = <http://data.gv.at> &&
19     ?mediaType = "text/csv" &&
20     strdt(?dateVersion, xsd:dateTimeStamp) >= "2020-05-10T00:00Z"^^xsd:dateTime)
21 }
22 LIMIT 10

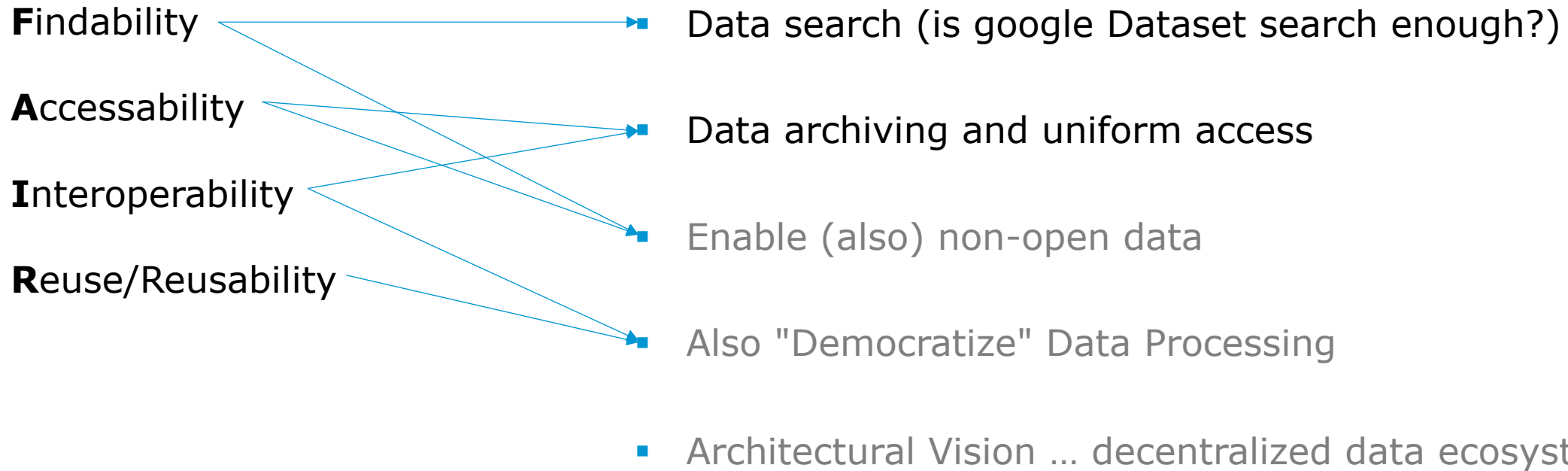
```

At the bottom of the interface, there are tabs for 'Table' and 'Response', and the URL 'https://archiver.ai.wu.ac.at/' is displayed.

Should it be really us
doing that or
... the data
providers/portals?
... te WebArchive?
... google?



What's next? What's missing for a FAIR (European) Data Ecosystem?



starting
points
are
there

beyond
current
OpenData
platforms

Data search – Is google Dataset search enough?

Findability,

Accessability

Interoperability

Reuse/Reusability

- How to Enable (also) non-open data?
 - Publish Open Metadata also for non-open (public) data!
(*e.g. micro-census data, public registries, e.g. COVID-vaccine-data?*)
 - Semantic descriptions of Provenance, Usage Policies
 - Accountability records (Blockchain?)
 - Enable automated-contracting of Data?
 - eIDs, verifiable credentials, formal policies ODRL

Data search – Is google Dataset search enough?

Findability,

Accessability

Interoperability

Reuse/Reusability

- How to also also "Democratize" Data Processing?
 - To really democratise Open Data, you need do democratize also computing power, or at least make it accessible to researchers
 - Marriage of "data portals" with "the cloud"
 - (e.g. "Open Data backend" for EOSC?)
 - Could mean a boost of European AI research?

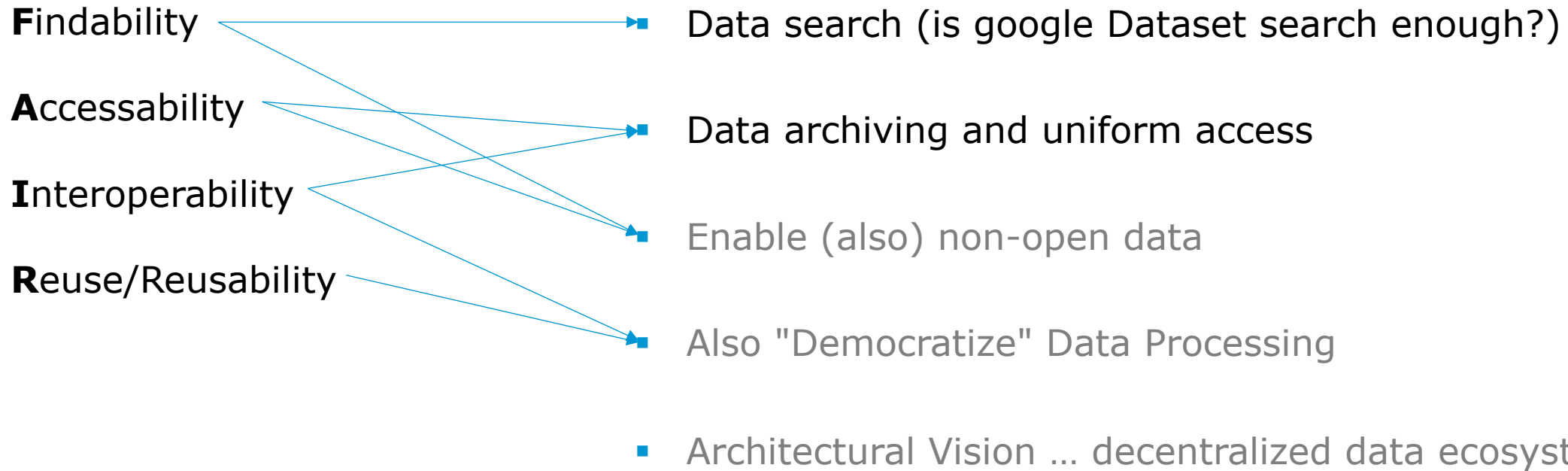
Forderungen der ASAI zu einer konkreten AI Strategie in Österreich

vom Vorstand der Austrian Society for Artificial Intelligence (ASAI)

Österreichische AI Strategie aus Sicht der Wissenschaft

Bernhard Moser, Georg Dorffner, Thomas Eiter, Wolfgang Faber, Günther Klambauer, Robert Legenstein, Bernhard Nessler, Axel Polleres, and Stefan Woltran.
Österreichische AI Strategie aus Sicht der Wissenschaft: Forderungen der ASAI zu einer konkreten AI Strategie in Österreich. *OCG Journal*, 01/2020:14--17, 2020.
Invited article (in German). [[http](http://ocg-journal.com/issue/view/issue-14-17)]

What's next? What's missing for a FAIR (European) Data Ecosystem?

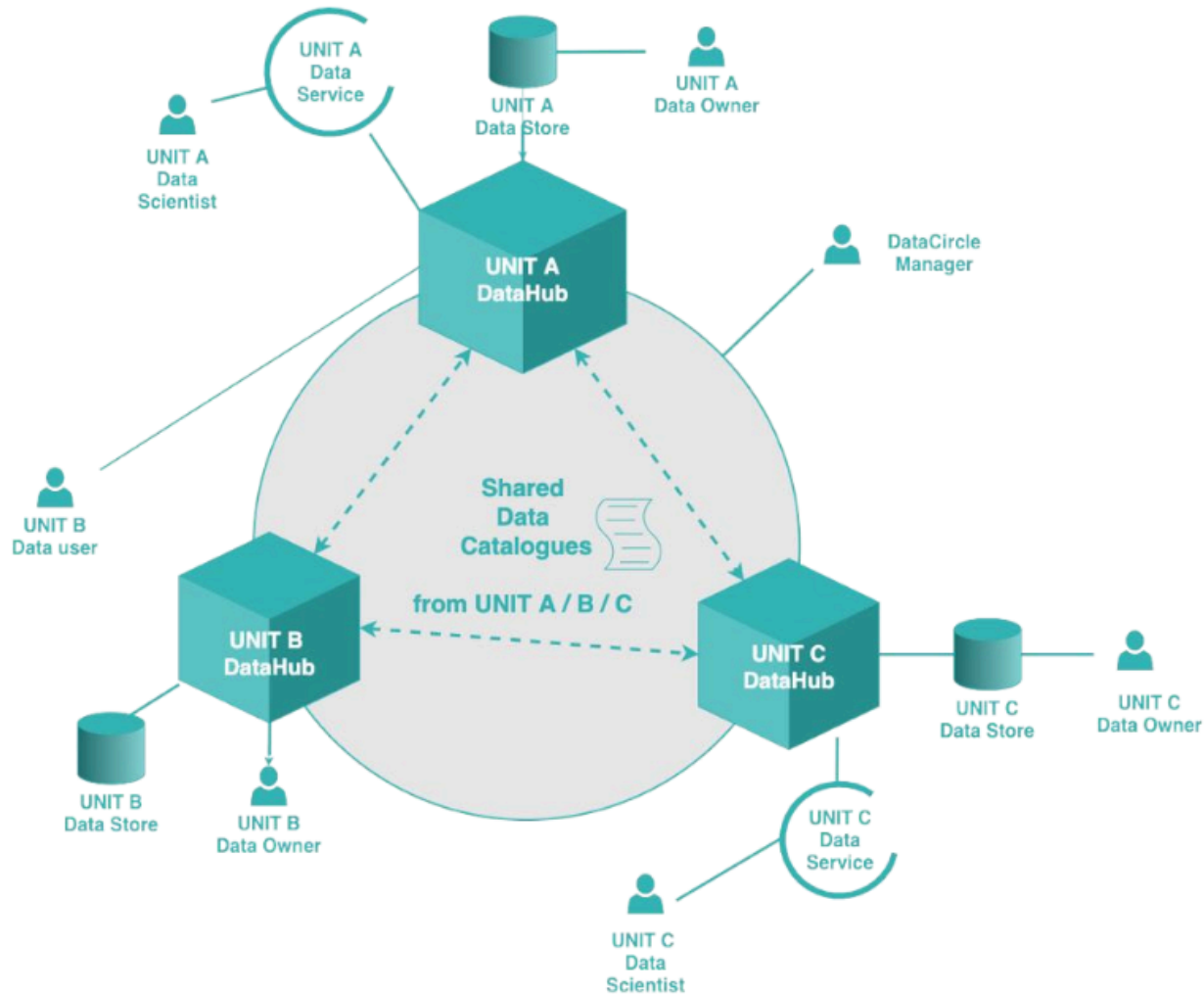


starting
points
are
there

beyond
current
OpenData
platforms

Vision: What's next?

Data-Ecosystem for Controlled Decentralized Data Sharing:



- FAIR (findable, accessible, interoperable and re-usable) **Open & Closed Data!**
- Data Sharing & Daten-Pooling
- Keep data sovereignty & trust
- Decentralisation
- Out of the box Data Science Pipelines & Cloud Services

- Shared Data Catalogs & Knowledge Graph are only a part of it!

Thanks to collaborators & project partners:



COOPERATION OGD  ÖSTERREICH

 DIO
Data Intelligence Offensive

Take-home Summary:

Knowledge Graphs/Linked Data can help to strengthen Open Data

They're both puzzle pieces for Open Data Ecosystems..
... or what I would rather call "**The Web of Data**"

We need to think further, particularly:

- include closed data into our considerations
- mix top-down (best practices and standards) with bottom up population of this "Web of Data"
- we will need a working decentralized architecture for this (supporting QoS/QoD monitoring, access control, etc.)

- From an academic point of view:

Still enough work for a couple of PhDs...

→ if you want to help? We're hiring! 😊

Thank you! Questions?



[@AxelPolleres](https://twitter.com/AxelPolleres)

data.wu.ac.at