



# The web of data: how are we doing so far?

---

ELENA SIMPERL

KING'S COLLEGE LONDON

@ESIMPERL

A network diagram consisting of numerous black circular nodes of varying sizes connected by thin, light grey lines. The nodes are scattered across the white background, with some forming small clusters and others standing alone. The lines represent connections between the nodes, creating a complex web-like structure.

The web has shaped our understanding and interactions with data

Microsoft Bing Search results for "bayern gegen psg".

911,000 Results | Date | Open links in new tab

### B. Munich vs PSG

Champions League Schedule - Standings

**LATEST MATCH** | MATCHES

Champions League - Round of 16 - 2nd leg

**B. Munich** (3) **2 - 0** (0) **PSG**

Today - FT

Gnabry, Serge 89' | Choupo-Moting, Eric Maxim 61'

Allianz Arena - Ref: Daniele Orsato

**Game stats** | Timeline | B. Munich lineup | PSG lineup

Shots on goal	6	4
Shots	10	10
Possession %	47.0%	53.0%

See more

All times are GMT Standard Time

---

### FC Bayern Munich

German football club

Founded: 27 February 1900 | Stadium: Allianz Arena | President: Herbert Hainer | Head coach: Julian Nagelsmann

---

### Paris Saint-Germain F.C.

French football club

Paris Saint-Germain Football Club, commonly referred to as Paris Saint-Germain, Paris, Paris SG or simply PSG is a professional football club

Answering factual questions

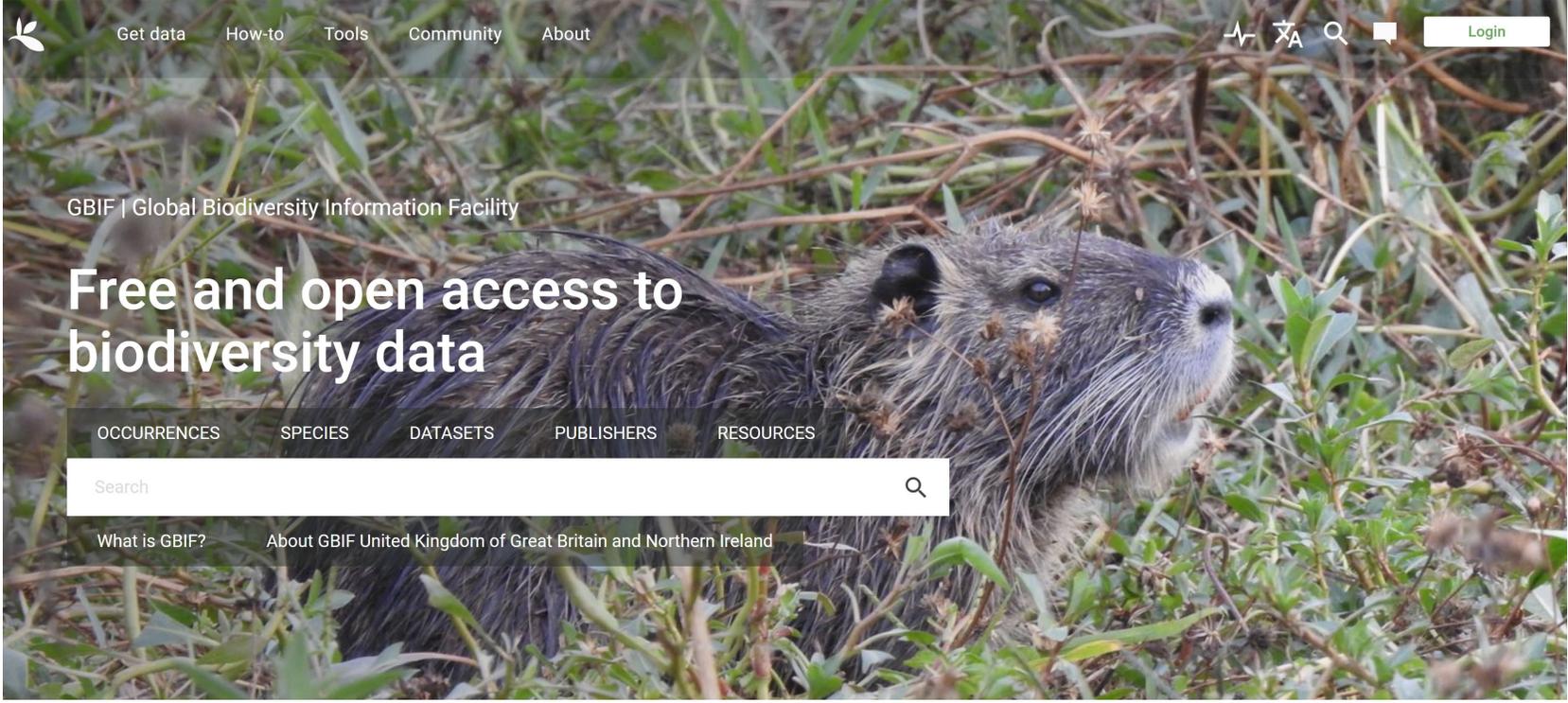
Book Orders.xlsx ☆ 📄 ☁

File Edit View Insert Format Data Tools Add-ons Help Last edit was made on 23 May 2017 by Amanda East

100% \$ % .0 .00 123 Calibri 11 B I S A

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Date ordered	Seller	Title	Copies Ordered	Cost	Delivery	Total Cost	Order number	Charged to	Charge Code	Dispatched	Delivered to Uni	Request by?
10	28/04/15	Amazon	Data Mining with R: Learning with Case Studies (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)	1	£ 57.99	£2.75	£ 60.74	204-5906330-5131568	Elena CC	EDSA	yes	4/27/2015	Elena
11	21/05/15	Amazon	Hadoop: The Definitive Guide	1	£ 33.50				Elena CC	WD Aqua	yes	26/5/15	Elena
12	21/05/15	Amazon	Writing That Works, 3e: How to Communicate Effectively in Business	1	£ 9.62	£0.00	£ 88.92	204-3028730-2112362	Elena CC	WD Aqua	yes	26/5/15	Elena
13	21/05/15	Amazon	The Elements of Style	5	£ 9.16				Elena CC	WD Aqua	yes	26/5/15	Elena
14	21/05/15	Amazon	On Writing Well: The Classic Guide to Writing Nonfiction	5	£ 3.58	£14.00	£ 31.90	203-5118543-0033118	Elena CC	WD Aqua	yes	4/6/2015	Elena
15	21/05/15	Amazon	R Cookbook	1	£ 26.50			203-0322709-5321116	Elena CC	EDSA	yes	26/5/15	Chris
16	21/05/15	Amazon	Agile Data Science: Building Data Analytics Applications with Hadoop	1	£ 26.50				Elena CC	EDSA	yes	22/5/15	Chris
17	21/05/15	Amazon	Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython	1	£ 16.46	£4.52	£100.48		Elena CC	EDSA	yes	26/5/15	Chris
18	21/05/15	Amazon	Programming Collective Intelligence: Building Smart Web 2.0 Applications	1	£ 26.50				Elena CC	EDSA	yes	26/5/15	Chris
19	23/6/15	Amazon	Complexity Theory and the Social Sciences: The start of the art	1	26.09		£ 26.09		Ordered by Nichola via Shaun Ford	WD Aqua		23/6/2015	Elena

# Sharing data online



Publishing data for others to use



OpenAddresses

# The free and open global address collection

[Learn more](#) ▾

Creating datasets in  
collaboration



# Creating digital traces

---



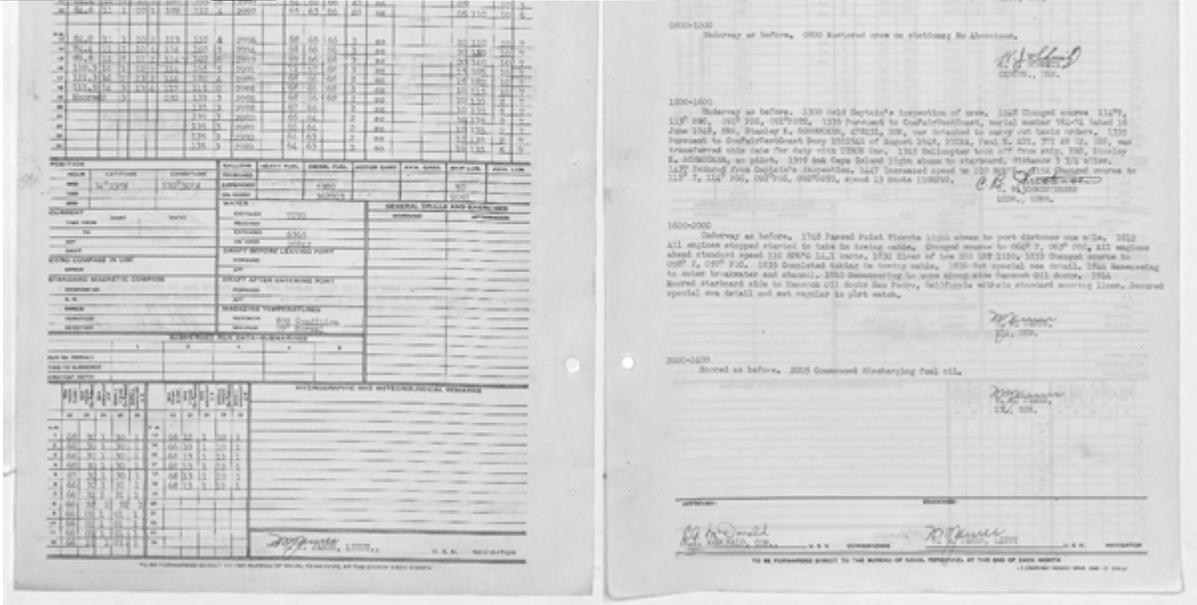
DECK LOG—COLUMN SHEET CONFIDENTIAL

USS CHARLESTON (PG-51)  
land, wn TO Sunday 15 April 1945  
7 W. R. MAKOSKY Comdr. U. S. N. COMMANDING

TIME	COURSE MILES	WIND DIRECTION SPEED	BAROMETER CORRECTED	TEMPERATURE		WIND DIRECTION	WIND SPEED	CLOUDS					VISIBILITY	
				WET BULB	WIND DIRECTION			LOW	MIDDLE	HIGH	DELTA	WIND SPEED		WIND DIRECTION
085	315	2	3020	54	53	50	50	Na	-	-	15	315	10	7
085	315	2	3026	54	53	50	50	Na	-	-	15	315	10	7
085	300	2	3024	53	54	49	50	Ns	-	-	15	300	10	7
085	300	4	3023	55	54	48	2	Sc	ΔC	-	30	300	8	7
085	295	6	3022	50	51	48	50	Na	-	-	40	295	10	7
085	290	4	3021	50	51	48	3	Sc	-	-	25	290	10	7
085	290	4	3021	50	51	48	2	Sc	ΔC	-	20	290	10	7
085	290	4	3020	50	51	48	2	Sc	ΔC	-	25	295	10	6
085	290	4	3020	50	51	48	2	Sc	ΔC	-	20	280	10	7
085	290	4	3020	50	51	48	50	Sc	ΔC	-	25	295	10	8

TASK TUTORIAL

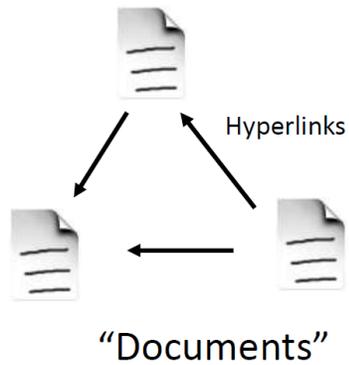
Navigation  
See FIELD GUIDE > Navigation for step-by-step instructions.  
Enter the Date (day-month-year / dd-mm-yyyy)



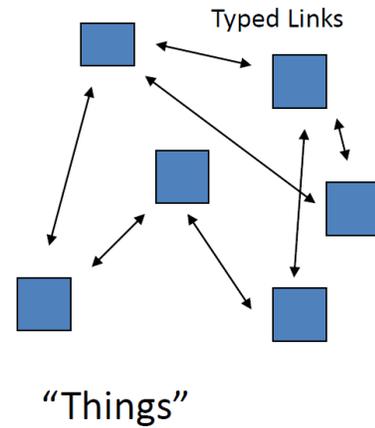
# Making algorithms smarter with data

Logbook of the USS *Burton Island* from a voyage to the Arctic in 1948

Web of Documents



Web of Data



(Source: Fensel, 2013)

Tim Berners-Lee  
 Date: 2006-07-27, last change: \$Date: 2009/06/18 18:24:33 \$  
 Status: personal view only. Editing status: imperfect but published.  
 Up to Design Issues

Linked Data

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF\*, SPARQL)
4. Include links to other URIs, so that they can discover more things.

Simple. In fact, though, a surprising amount of data isn't linked in 2006, because of problems with one or more of the steps. This article discusses solutions to these problems, details of implementation, and factors affecting choices about how you publish your data.

The four rules

I'll refer to the steps above as rules, but they are expectations of behavior. Breaking them does not destroy anything, but misses an opportunity to make data interconnected. This in turn limits the ways it can later be reused in unexpected ways. It is the unexpected re-use of information which is the value added by the web.

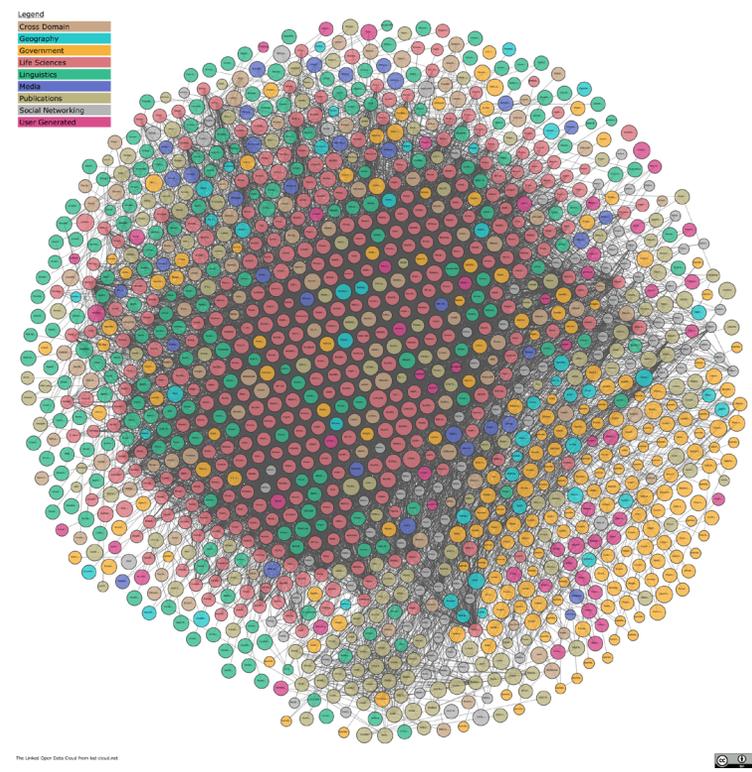
The first rule, to identify things with URIs, is pretty much understood by most people doing semantic web technology. If it doesn't use the universal URI set of symbols, we don't call it Semantic Web.

The second rule, to use HTTP URIs, is also widely understood. The only deviation has been, since the web started, a constant tendency for people to invent new URI schemes (and sub-schemes within the `urn: scheme`) such as LSIDs and handles and XRI's and DOIs and so on, for various reasons. Typically, these involve not wanting to commit to the established Domain Name System (DNS) for delegation of authority but to construct something under separate control. Sometimes it has to do with not understanding that HTTP URIs are names (not addresses) and that HTTP name lookup is a complex, powerful and evolving set of standards. This issue discussed at length elsewhere, and time does not allow us to delve into it here. I @ref TAG



(Source: W3C, 2006)

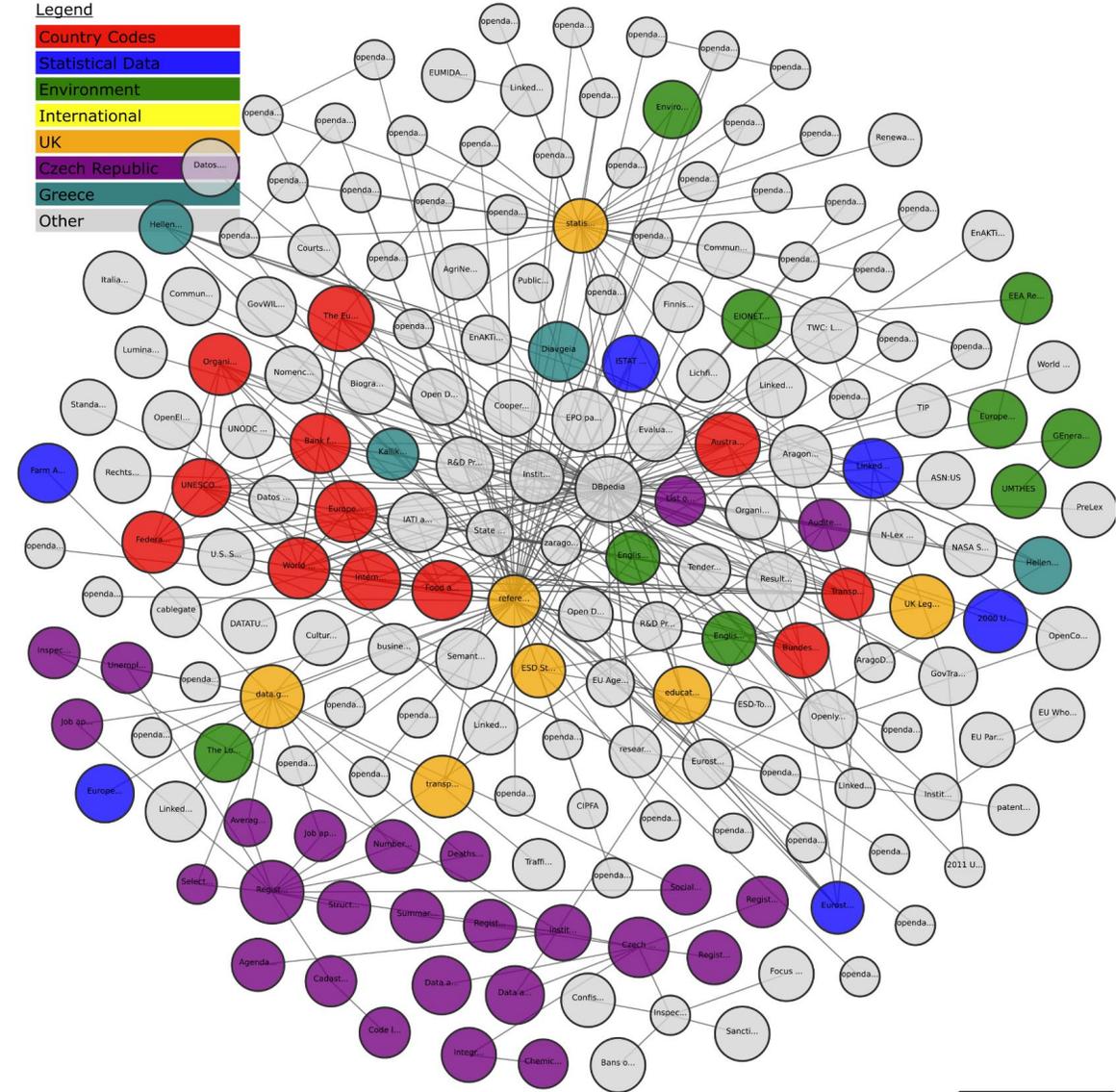
The theory and practice of the web are not the same



Release	Datasets	Links
2022	1573	2.6 M
2020	1440	2.5 M
2018	1308	2.4 M

Legend

- Country Codes
- Statistical Data
- Environment
- International
- UK
- Czech Republic
- Greece
- Other



The Government Linked Open Data Cloud from lod-cloud.net



### Datasets

data.europa.eu (1.6M datasets from 170 publishers in 36 countries)

Top 12 Catalogues

Weight

Top 12
Top 20
all Catalogues
100 Points
100 Points
110 Points
75 Points
20 Points
405 Points

Country	Name	Findability	Accessibility	Interoperability	Reusability	Contextuality	Rating
	European Union Intellectual Property Office (EU)	100 / 100	100 / 100	80 / 110	75 / 75	5 / 20	Excellent
	European Central Bank (EU)	98 / 100	97 / 100	77 / 110	75 / 75	10 / 20	Excellent
	Directorate-General for Migration and Home Affairs (EUROPE)	100 / 100	100 / 100	60 / 110	75 / 75	15 / 20	Good
	Directorate-General for Education, Youth, Sport and Culture (EUROPE)	90 / 100	100 / 100	72 / 110	75 / 75	10 / 20	Good
	European Research Council Executive Agency (EU)	100 / 100	100 / 100	60 / 110	75 / 75	10 / 20	Good
	Executive Agency for Small and Medium-sized Enterprises (EUROPE)	100 / 100	100 / 100	60 / 110	75 / 75	10 / 20	Good
	European Agency for Safety and Health at Work (EU)	90 / 100	100 / 100	70 / 110	75 / 75	10 / 20	Good
	European Political Strategy Centre (EU)	80 / 100	100 / 100	80 / 110	75 / 75	5 / 20	Good
	European Union Agency for Cybersecurity (EU)	90 / 100	100 / 100	60 / 110	75 / 75	13 / 20	Good

## Directorate-General for Migration and Home Affairs

Directorate-General for Migration and Home Affairs

**Catalogue dashboard**

Distribution accessibility issues (0) DCAT-AP Schema Violations (54)

### Rating evolution

Good

#### Findability

100 / 100

Time based search 100% Geo search 100% Keyword usage 100% Categories 100%

#### Accessibility

100 / 100

Most frequent acces... 200 Download URL 100% Most frequent downl... 200

#### Interoperability

60 / 110

Non-proprietary 0% Format / Media type ... 100% Machine readable 100% DCAT-AP compliance 0%

Media type 100% Format 100%

#### Reusability

75 / 75

Contact point 100% License information 100% License vocabulary 100% Access restrictions 100%

Publisher 100% Access restrictions v... 100%

Datasets by data theme

Datasets per origin

Datasets per origin and catalogue

Total Number of Datasets: 1 311 527

Data theme ^	Trend per year	Percentage of total datasets	Number of Datasets
Agriculture, fisheries, forestry and food	0.06%	29%	379 333
Economy and finance	-0.03%	5.25%	67 337
Education, culture and sport	-0.14%	1.5%	21 034
Energy	0.10%	0.75%	8 781
Environment	-0.00%	20.75%	273 604
Government and public sector	0.05%	7.5%	98 575
Health	-0.13%	1.25%	16 899
International issues	0.01%	0.09%	1 143
Justice, legal system and public safety	0.00%	14.5%	188 893
Population and society	-0.01%	4.25%	54 282
Provisional data	-	0.01%	125
Regions and cities	0.19%	4.25%	55 050
Science and technology	0.10%	7.5%	98 616
Transport	0.00%	3.75%	47 855

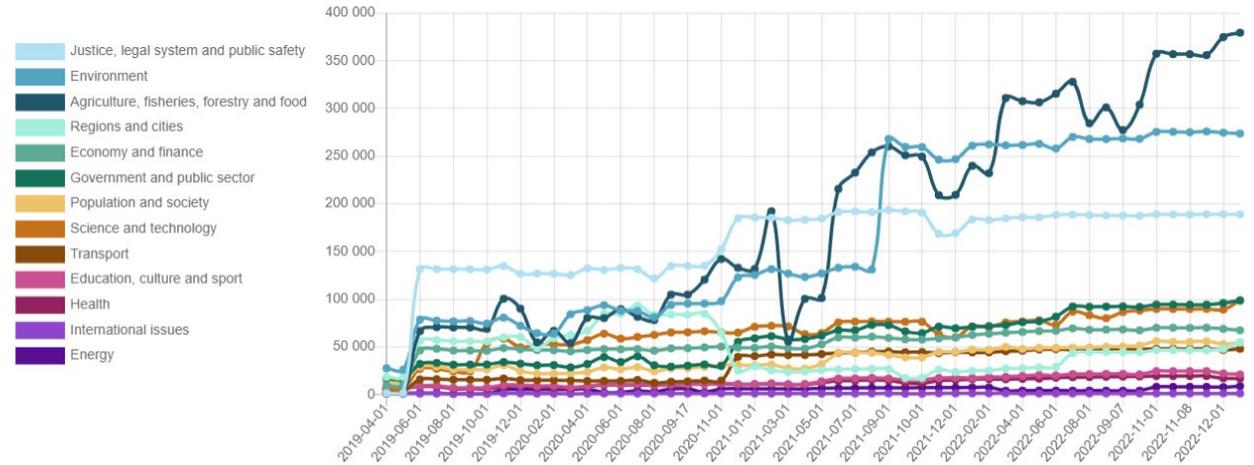
## Evolution over time

Total Number of Datasets

Datasets by data theme

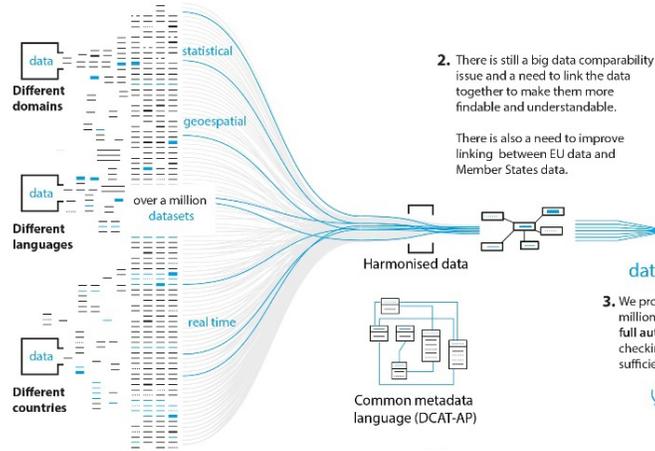
Datasets per origin

Datasets per origin and catalogue



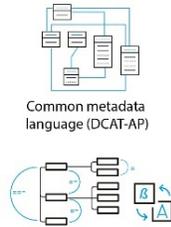
Unique features of data.europa.eu

1. The challenge is to present a heterogeneous and large amount of data coming from different countries and domains in a way that makes sense.



2. There is still a big data comparability issue and a need to link the data together to make them more findable and understandable.

There is also a need to improve linking between EU data and Member States data.

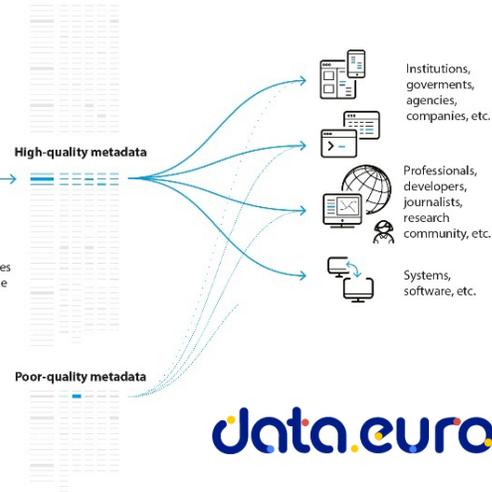


Multilingual EU reference data

Data become harmonised by being categorised into domains via data themes. They are described with the same metadata language (DCAT-AP), harmonised and linked using multilingual EU reference data.

From quantity to quality

Current priorities are to improve the quality of metadata, to allow for interoperability, and the relevance and quality of the data.



3. We provide access to over a million datasets, which requires full automation, hindering the checking of quality at a sufficient level of detail.



data.europa.eu



Navigation bar for data.europa.eu with links for Data, Studies, data.europa academy, News, and Contact.

Secondary navigation bar with links for Dataset, Categories, Quality, Similar Datasets, Feedback, Share, Dataset Feed, and Cite (highlighted with a red box).

Consolidated list of persons subject, under EU sanctions, to travel restrictions

Directorate-General for Financial Stability, Financial Services and Capital Markets Union Publisher: Directorate-General for Financial Services and Capital Markets Union

In its policy, the European Union intervenes when necessary to prevent conflict or in response to emerging or actual crises. In certain cases, EU intervention can take the form of restrictive measures or 'sanctions'. One type of measure are restrictions on admissions (travel bans or travel restrictions) targeting individuals. This dataset contains the Consolidated list of persons subject, under EU sanctions, to travel restrictions.

Disclaimer: While every effort is made to ensure that the database and the consolidated list correctly reproduce all relevant data of the officially adopted texts published in the Official Journal of the European Union, the Commission cannot accept any liability for possible omissions of relevant data or mistakes, and for any use made of the database or of the consolidated list. Only the information published in the Official Journal of the EU is deemed authentic.





### Country Insights

This section provides insights into the status of open data in European countries. The section includes information about countries' national open data portal, the level of open data maturity, use cases, catalogues, events, reports, news, webinars, and interviews about open data. [Read more...](#)



### Open Data Impact

The 2020 report on the Economic Impact of Open Data researches the value created by open data in Europe. It is the second study by the European Data Portal, following the 2015 report. The study forecasts the open data market size and the employment growth for 2025. Impact is exemplified by efficiency gains and cost savings due to open data. [Read more...](#)



### Open Data Maturity

A series of indicators have been selected to measure open data maturity across Europe. These indicators cover the level of development of national policies promoting open data, an assessment of the features made available on national data portals as well as the expected impact of open data. [Read more...](#)



### Studies

A collection of papers that discuss various relevant topics around open data, such as barriers in working with open data, high-value datasets, and business-to-government data sharing. [Read more...](#)



### Use Cases

This part of the portal contains use cases from both publishers and re-users of open data. The use cases provided below are gathered via the 'Tell us your story' form on the portal or shared with us by other contacts. This section of the portal will grow over time. [Read more...](#)



### Tell us your story

Share your story about how you make use of open data. [Read more...](#)



The screenshot shows the 'data.europa.eu' website. At the top, there is a navigation bar with 'Home', 'Datasets', 'Documentation', 'Publications', 'data.europa academy', 'News & events', and 'Contact us'. Below this, a search bar is visible with the text '900+ use cases!!!' overlaid in an orange cloud. The 'Use cases' section is active, showing filter options for Country, Region, Sector, and Type of Use Case. Search results are displayed for 'EUTwinnings' and 'Fake News Shield', both with PDF icons. The 'EUTwinnings' result includes a description: 'The goal of the EU Twinnings application is to allow users to find their area's twin (an area with the most similar statistics data) and explore similar regions of the EU. Simply fill in the city name and...'. The 'Fake News Shield' result includes a description: 'Fake News Shield is an application that detects fake news for the population by scanning several sources for their credibility using machine learning. As a stand alone app, it can be installed on the...'. Other results include 'Fix my Berlin' and 'Irrigation Calculator'.

## An Assessment of Adoption and Quality of Linked Data in European Open Government Data \*

Luis-Daniel Ibáñez<sup>1</sup>[0000-0001-6993-0001], Ian Millard<sup>2</sup>, Hugh Glaser<sup>2</sup>, and Elena Simperl<sup>1</sup>[0000-0003-1722-947X]

<sup>1</sup> University of Southampton  
 {l.d.ibanez,e.simperl}@soton.ac.uk  
<sup>2</sup> Seme4 Ltd.  
 {ian.millard,hugh.glaser}@seme4.com

**Abstract.** The European Commission has adopted Linked Data principles and practices with the purpose of increasing the accessibility, interoperability and value of the data that is made available openly by European public sector organisations. This includes investment in metadata development for describing open datasets, catalogs of resources with persistent URIs, and the European Data Portal (EDP), which provides a single point of access, search and exploration of European open data. As the Public Sector Initiative (PSI) Directive is being revised, a critical question for the commission is the extent to which open government data publishers have adopted Linked Data, and how they are applying the underlying technologies. In this paper, we undertake a quantitative analysis to support this. We explore if and how open data portals indexed by the EDP are using Linked Data and assess the quality of the datasets according to multiple dimensions.

**Keywords:** Linked Data, open government data, data quality

### 1 Introduction

Linked Data refers to a set of principles, technologies and practices that facilitate data integration. Publishers are encouraged to adopt them to make their data more useful [5]. Linked Data makes it easier for developers to access and combine datasets from different sources. To unlock the value of their data, publishers are advised to [5]:

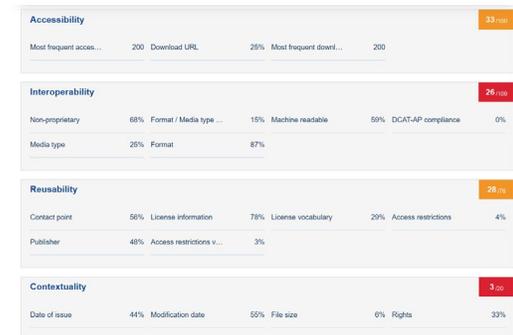
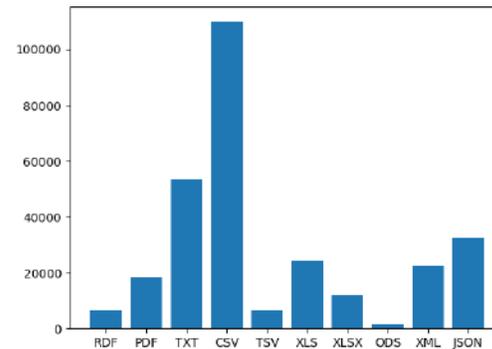
1. use URIs to name things and relationships among things;
2. use HTTP URIs so those names can be looked up (a technique called *dereferencing*);
3. return useful information upon lookup of URIs, using open standards such as RDF; and

\* Supported by the European Data Portal, an initiative funded by the European Union

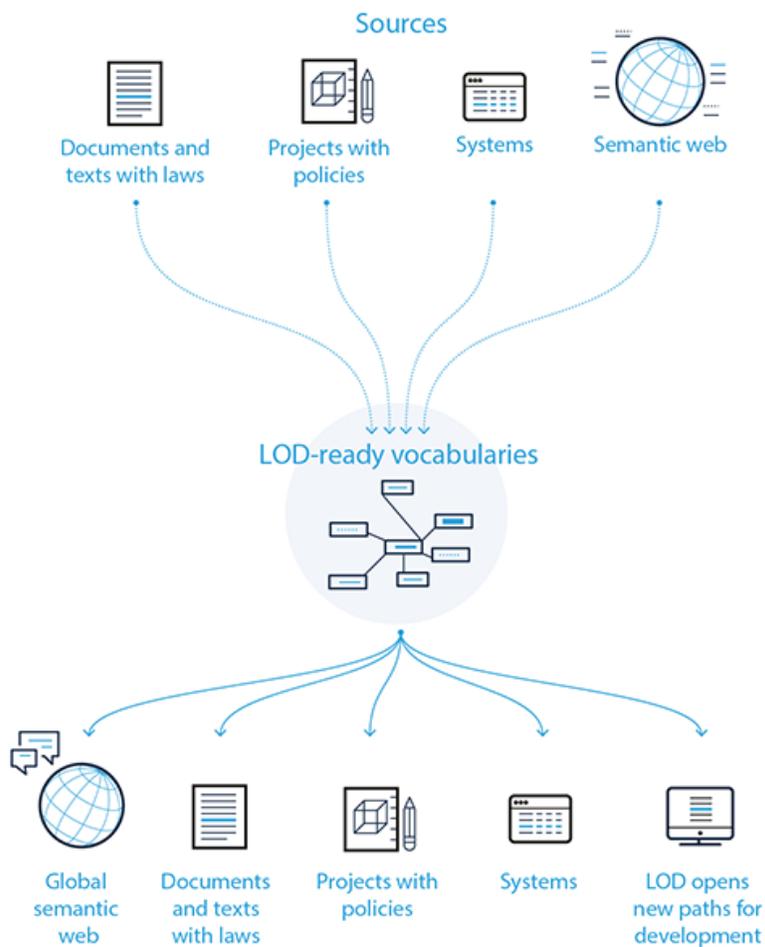
Total datasets	6636	Total distributions	8780
Successful distribution download	8016	Failed distribution downloads	764
Successful distribution validation	6990	Failed distribution validation	1026
Datasets with at least one valid distribution	5856	Triples inspected	137,208,657

Domain	# (%) of datasets
www.dati.lombardia.it	2836 (48.4%)
opendata.aragon.es	1252 (21.4%)
dati.opendataground.it (Comune AlbanoLaziale)	1011 (17.3%)
datos.gijon.es	357 (6.1%)
opendata.caceres.es	259 (4.4%)
www.dati.milivenezia.furla.it	232 (2.9%)
datos.santander.es	172 (2.9%)
nexo.carm.es (Region Murcia)	126 (2.2%)
opendata.camden.gov.uk	100 (1.7%)
datos.madrid.es	76 (1.3%)
Other 64 hosts	275 (4.7%)

Low uptake of linked data, limited vocabulary reuse, proprietary, non-dereferenceable vocabularies, reasonable metadata quality



The use of linked open data improves clarity and can be of great help in correlating pieces of information in different areas or systems (e.g. documents and texts with laws, projects with policies) and identifying relationships that might not otherwise be apparent



**Cellar**  
The semantic repository of the Publications Office

Publications Office

The image shows a 3D cube filled with white, molecular-like structures (trifurcated shapes) against a blue background. The top of the image features the European Union flag. The bottom left corner shows the Publications Office logo.

## EU Vocabularies

Share Help

Home Controlled vocabularies Models Business collections Online tools Releases Help

EU Vocabularies Business collections

### Business collections



#### AKN4EU

AKN4EU is the future machine-readable structured format for the exchange of legal documents in the EU decision-making process.

[Learn more >](#)

#### BITS (Book Interchange Tag Suite)

The Publications Office has selected BITS (Book Interchange Tag Suite) as an XML mark-up model suitable for the production of general publications.

[Learn more >](#)

#### CDM (Common Data Model)

CDM is the Common Data Model of the resources published by the Publications Office of the European Union.

[Learn more >](#)

Collections of semantic assets linked to concrete projects give an easy overview of the applied vocabularies and provide an introductory platform for authors to present their projects.

#### Common European Skills Data Space

Common European Skills Data Space is a European framework to support the transparency and understanding of skills and qualifications acquired in format...

[Learn more >](#)

#### Core Vocabularies

Core Vocabularies are simplified, re-usable and extensible data models that capture the fundamental characteristics of an entity in a context-neutral fashion.

[Learn more >](#)

#### Corporate Reference Data Management

The objective of the Corporate Reference Data Management coordination group is to build consensus on a corporate reference data management policy.

[Learn more >](#)

#### DCAT-AP for data portals in Europe

The DCAT Application profile for data portals in Europe (DCAT-AP) is a specification based on the Data Catalogue vocabulary (DCAT) for describing public sector datasets ...

[Learn more >](#)

#### Digital Europa Thesaurus (DET)

The Digital Europa Thesaurus (DET) is a multilingual thesaurus covering the main subject matters of the European Commission's public communications.

[Learn more >](#)

#### ELI (European Legislation Identifier)

The technical specifications of the European Legislation Identifier (ELI) can be found here.

[Learn more >](#)

69 Data Publishers

419,000 Opportunities openly published every month

1,200+ Locations covered by open data sessions



OpenActive Vocabulary Terms

The terms in the OpenActive Vocabulary, listed below, are defined within the [Modelling Opportunity Data](#) and [Open Booking API](#) specifications.

These specifications are being developed by the [OpenActive W3C Community Group](#).

This vocabulary is also available in JSON-LD format via a GET of the URL "https://openactive.io/" using an Accept header of application/ld+json, and additionally via a CDN at the URL "https://openactive.io/ns/oa.jsonld" for production use.

For more information, see the [developer documentation](#).

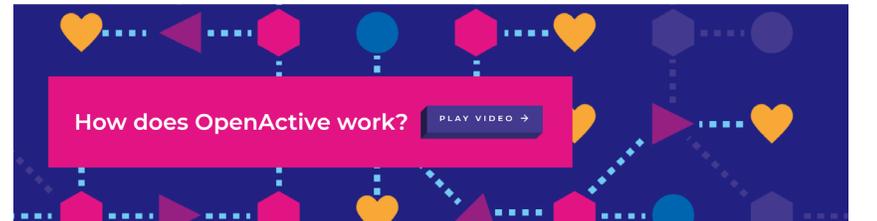
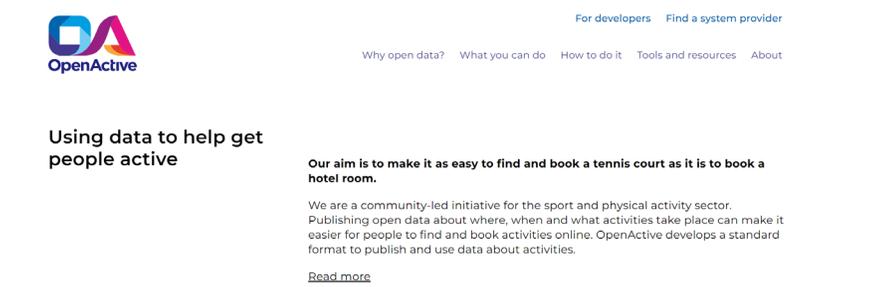
Classes

- [AgentBroker](#)
- [AttendeeAbsent](#)
- [AttendeeAttended](#)
- [AwaitingSellerConfirmation](#)
- [BabyChanging](#)

Example data types



Example data properties



# Unlocking the value of fragmented public data

Recommendations for policy makers

Alex Parsons and Anna Powell-Smith

## Use cases

### Filter by

Country

- Any -

Region

- Any -

Sector

- Any -

Type of Use Case

Re-use, Release

Clear filters

linked data

Search results (2)

Sort by

Publication date

Type of use case: Re-use Release

Government & Public Sector | 21 May 2019

[DublinDashboard](#) PDF

Dublinked is the open data portal for the city of Dublin. At the website, users can download information on, among others, transport, energy and the environment, public health and recreation...

Show more

Science & Technology | 26 March 2018

[Linked Data Center](#) PDF

The Linked Data Center helps any business in extracting value from data, by converting them into smart and connected data.

Items per page 10

### Rating

Bronze

Silver

Gold

Diamond

### Description

Available on the web, in any format, with an open licence.

The same as for one star, plus available as machine-readable structured data.

The same as two stars, but available in an agreed format for the data (i.e. matching a schema, or using a standard template).

All of the above, but the file, or a reference to the file, is also published in a central repository.

Google search for "football dataset" showing results from DataHub.io, Kaggle, and Data.world.

Google  
football dataset

All Images News Videos Books More Tools

About 4,670,000 results (0.35 seconds)

**DataHub.io**  
https://datahub.io › collections › football  
**Football Data - DataHub.io**  
13 Dec 2017 — Football Data. A collection of awesome **football datasets** including national teams, clubs, match schedules, players, stadiums, etc.

**Kaggle**  
https://www.kaggle.com › datasets › hugomathien › so...  
**European Soccer Database - Kaggle**  
The ultimate Soccer database for data analysis and machine learning. What you get: +25,000 matches; +10,000 players; 11 European Countries with their lead ...  
https://www.kaggle.com › datasets › martj42 › internat...  
**International football results from 1872 to 2022 | Kaggle**  
This **dataset** includes 44,341 results of international **football** matches starting from the very first official match in 1872 up to 2022.

**Data.world**  
https://data.world › datasets › soccer  
**There are 38 soccer datasets available on data.world.**  
There are 38 soccer **datasets** available on data.world. Find open data about soccer contributed

Datasets :

https://www.kaggle.com/hugomathien/soccer  
**European Soccer Database**  
Oct 23, 2016 — The ultimate Soccer database for data analysis and machine learning What you get: +25,000 matches +10,000 players 11 European Countries with their lead championship Seasons 2008 ...  
Licence: Open Database License (ODbL) v1.0      Format(s): zip

https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats  
**2021-2022 Football Player Stats**  
May 29, 2022 — Context This dataset contains 2021-2022 football player stats per 90 minutes. Only players of Premier League, Ligue 1, Bundesliga, Serie A and La Liga are listed. 2022-2023 Football ...  
Licence: CC BY 4.0      Format(s): zip

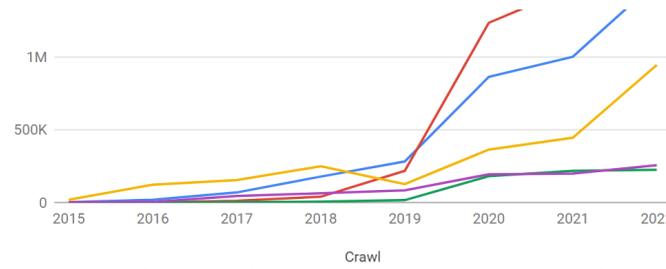
https://data.world/data-society/european-soccer-data  
**European Soccer Data**  
Feb 1, 2023 — The ultimate Soccer database for data analysis and machine learning What you get: +25,000 matches +10,000 players 11 European Countries with their lead championship Seasons 2008 ...  
Format(s): zip

45 million  
datasets  
(DCAT,  
schema.org)

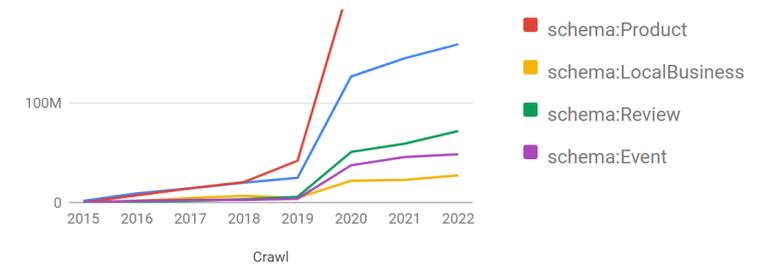
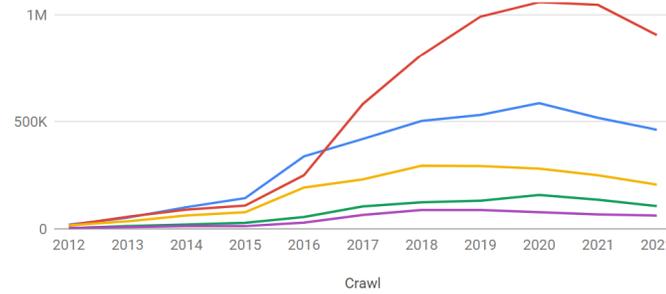
# ENDORSE

ific Subsets of the Schema.org Data

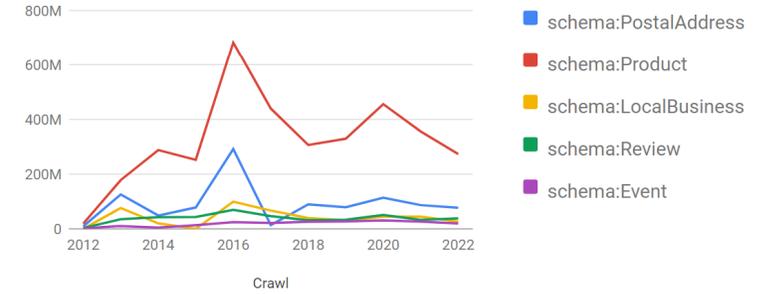
Subset	General Stats	Related Classes	Size (# Files)	Download (Sample)
eArea	Quads: 77,960,962 URLs: 381,624 Hosts: 2,695	<a href="http://schema.org/City">http://schema.org/City</a> (2,154,382) <a href="http://schema.org/ImageObject">http://schema.org/ImageObject</a> (1,260,633) <a href="http://schema.org/AdministrativeArea">http://schema.org/AdministrativeArea</a> (932,550) <a href="http://schema.org/Person">http://schema.org/Person</a> (819,487) <a href="http://schema.org/ListItem">http://schema.org/ListItem</a> (810,411)	923.8 MB (1)	<a href="#">AdministrativeArea (sample)</a>
	Quads: 54,033,331 URLs: 160,852 Hosts: 538	<a href="http://schema.org/Airport">http://schema.org/Airport</a> (4,569,542) <a href="http://schema.org/GeoCoordinates">http://schema.org/GeoCoordinates</a> (2,461,017) <a href="http://schema.org/Flight">http://schema.org/Flight</a> (1,593,849) <a href="http://schema.org/Airline">http://schema.org/Airline</a> (1,473,676) <a href="http://schema.org/Offer">http://schema.org/Offer</a> (1,244,616)	415.83 MB (1)	<a href="#">Airport (sample)</a>
	Quads: 1,485,993,805 URLs: 13,279,199 Hosts: 251,442	<a href="http://schema.org/Answer">http://schema.org/Answer</a> (53,728,963) <a href="http://schema.org/Question">http://schema.org/Question</a> (45,624,470) <a href="http://schema.org/ImageObject">http://schema.org/ImageObject</a> (30,813,788) <a href="http://schema.org/ListItem">http://schema.org/ListItem</a> (30,093,835) <a href="http://schema.org/Person">http://schema.org/Person</a> (18,278,578)	26.69 GB (15)	<a href="#">Answer (sample)</a>
	Quads: 308,456,812 URLs: 4,999,213 Hosts: 21,623	<a href="http://schema.org/Book">http://schema.org/Book</a> (13,434,918) <a href="http://schema.org/Country">http://schema.org/Country</a> (7,685,986) <a href="http://schema.org/Offer">http://schema.org/Offer</a> (7,381,043) <a href="http://schema.org/Person">http://schema.org/Person</a> (6,792,930) <a href="http://schema.org/ListItem">http://schema.org/ListItem</a> (3,755,829)	7.67 GB (4)	<a href="#">Book (sample)</a>
	Quads: 213,841,120 URLs: 1,152,893 Hosts: 11,786	<a href="http://schema.org/City">http://schema.org/City</a> (5,832,337) <a href="http://schema.org/ImageObject">http://schema.org/ImageObject</a> (3,890,159) <a href="http://schema.org/Person">http://schema.org/Person</a> (3,482,668) <a href="http://schema.org/ListItem">http://schema.org/ListItem</a> (3,132,617) <a href="http://schema.org/PostalAddress">http://schema.org/PostalAddress</a> (2,928,984)	2.25 GB (3)	<a href="#">City (sample)</a>
	Quads: 154,890,333	<a href="http://schema.org/CollegeOrUniversity">http://schema.org/CollegeOrUniversity</a> (5,643,657)	2.15 GB	<a href="#">CollegeOrUniversity (sample)</a>



Development of Selected schema.org Classes by #PLDs - Microdata format



Development of Selected schema.org Classes by #Entities - Microdata format



## ded JSON-LD, and Microformat

and Microformats data sets which have been extracted

5 billion pages contained in the crawl (46.88%). These data sets consist of 86 billion [RDF quads](#).

Microformats data sets are given on the page [how to get](#)

ta and JSON-LD corpora.

domains are not counted as separate domains.

There is a lot of annotated data online, especially about locations, products, businesses



(Source: Walker & Simperl, 2017)

# The ten guidelines

Co-located  
Documentation

Promote use

Organise for use of the datasets - rather than simply for publication

Promote use through data storytelling and community building, borrowing from open-source communities and other peer-production systems

Invest in discoverability best practices, borrowing from e-commerce and web search

Publish good quality metadata - to enhance reuse

Adopt standards to ensure interoperability

Co-locate tools so that a wider range of users can be engaged with

Link datasets to enhance value

Be accessible by offering options for from APIs to CSV downloads

Co-locate documentation - users should not need to be domain experts to understand the data;

Be measurable - as a way to assess how well they are meeting users' needs.

Publish metadata

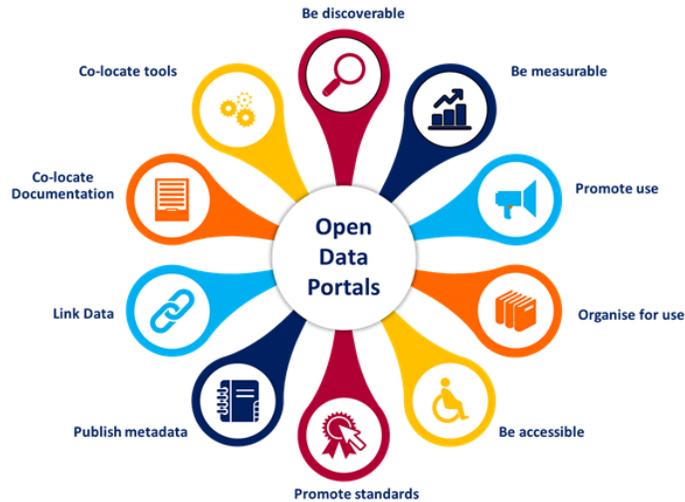
Organise for use

Be accessible

# Operationalising the guidance

Literature review to develop 5\* schemes to operationalise indicators.

Application of the schemes on 10 open data portals at different maturity level.



(Walker & Simperl, 2017)

**LINKED DATA**  
 ★ On the web, open license  
 ★ Machine-readable data  
 ★ Non-proprietary format  
 ★ RDF standards  
 ★ Linked RDF  
**IS YOUR DATA 5 ★ ?**

	Bronze	Silver	Gold	Platinum
<b>Legal</b>				
Openly licensed & legally reusable (= 'open')	✓	✓	✓	✓
Clear rights statement, detailing any copyrights		✓	✓	✓
Privacy issues addressed		✓	✓	✓
Machine readable rights statement			✓	✓
<b>Practical</b>				
Accessible on the web	✓	✓	✓	✓
Discoverable (linked to from other web pages)		✓	✓	✓
Data is timestamped or up to date			✓	✓
Data will be available for at least a year		✓	✓	✓
Guaranteed <i>timeliness</i> (data always up to date)			✓	✓
Regular backups of data			✓	✓
Quality issues documented			✓	✓
<b>Technical</b>				
Data uses a machine readable format		✓	✓	✓
Data published in content appropriate formats		✓	✓	✓
Data uses open standard machine readable formats			✓	✓
Single consistent URL for downloading data			✓	✓
Machine readable provenance documentation				✓
URLs used as identifiers within data				✓
<b>Social</b>				
Data is documented		✓	✓	✓
Contact details for people to provide feedback and ask questions		✓	✓	✓
Machine readable metadata (documentation)			✓	✓
Social media accounts used to promote data			✓	✓
Forum or mailing list for users			✓	✓
Dedicated comms team building user community				✓

# Example: Organise for use

---

Each dataset is accompanied by a comprehensive descriptive record (going beyond a collection of structured metadata)

An extract of the data can be previewed (for sense making)

The portal provides recommendations for related datasets

The portal enables users to review/rate the datasets

Keywords from datasets are linked to other published datasets



# Example: Co-locate documentation

---

Supporting documentation does not exist.



Supporting documentation exists, but as a document found separately from the data.



Supporting documentation is found at the same time as the data (e.g. the link to the document is next to the link to the data in the search).



Supporting documentation can be immediately accessed from within the dataset but it is not context sensitive (e.g. a link to the documentation or text contained within the dataset).



Supporting documentation can be immediately accessed from within the dataset and it is context sensitive so that users can immediately access information about a specific item of concern (e.g. a link to a specific point in the documentation or the text contained within the dataset).



ENDORSE



LONDON DATASTORE



Varying open data maturity levels  
*Be discoverable , co-locate documentation, be measurable* are universally challenging

# User-centric principles in data.europa.eu

**Dataset** Football pitches

data.gov.uk

**Dataset** Quality Similar datasets

Data showing the usage of football pitches in Plymouth by organisation, as recorded in 2016.

Dataset feed Linked data Cite

**Publisher** Name: Plymouth City Council

**Catalogue Record** Added to data.europa.eu: 12.10.2021 07:01  
Updated on data.europa.eu: 01.07.2022 05:04

**URIref** <http://data.europa.eu/88u/dataset/football-pitches>

Show less ^

**Distributions (1)**

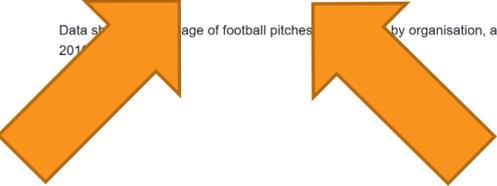
Link to the data	Format	Distribution added	Actions
<a href="#">Plymouth Football Pitches - Plymouth 2016</a>	CSV		Options Download Linked data

**Plymouth Football Pitches - Plymouth 2016**  
Data showing the usage of football pitches in Plymouth by organisation. Details include which days each pitch is used and the structure of the teams which use it. Correct as of 2016. Source: Plymouth City Council\_

**Licence** [Open Government Licence 3.0](#) Licensing Assistant

Show less ^

Name	Sub Area	Ground	Club ID	Team N...	Adult m...	Adult fe...	Youth m...	Youth fe...	Youth m...	Youth fe...	MS 7v7 (...)	MS 5v5 (...)	Team ID	Day Pla...	Adult m...	Adult m...	Adult m...	Adult m...	Youth m...	Y...	
AFC Ply...			218,936	1st		1							11,217,915	Sunday				1			
Edgecu...			80,836	1st	1								11,024,550	Saturday		1					
Ocean C...			238,765	First	1								11,698,063	Saturday		1					
SB Fran...			80,922	U15				1					11,635,564	Saturday							1
Signal B...			228,872	U13				1					11,660,036	Saturday							1
Ivybridge...			230,316	second	1								11,466,317	Wednes...							1
Ivybridge...			230,316	Thirds	1								11,467,870	Wednes...							1
Plymout...			234,026	First	1								11,464,314	Tuesday							1
Ivybridge...			230,316	First	1								11,466,316	Wednes...							1
Ivybridge...			230,316	Women				1					11,467,871	Wednes...							1
HMS Ral...			81,190	1st	1								11,024,985	Wednes...							1
Devon a...			80,635	1st	1								11,022,893	Saturday		1					
SB Fran...			80,922	U14				1					11,651,902	Saturday							1
Argyle L...			222,779	First	1								11,272,293	Sunday			1				
Horrabri...			115,414	U11						1			11,633,096	Saturday							
Plymout...			232,072	U16				1					11,645,394	Saturday							1
Plymout...			232,072	U18	1								11,596,957	Sunday				1			



# SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES

- » Research data
- » Publication characteristics

## Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson et al.<sup>1\*</sup>

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its use by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them and some essential implementation in the community.

### Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of “long-term care” of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes “good data management” is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to “data” in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other’s data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtis.nl).

\*A full list of authors and their affiliations appears at the end of the paper.

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016

# A lot of guidance available already

A curated, informative and educational resource on data and metadata standards, inter-related databases and data policies.

HOW CAN WE HELP?

Journal editors & publishers  
Create and maintain an interrelated list of citable standards, databases and repositories to recommend to your authors, users or their community, and revise this recommendation over time... (read more)

Researchers Developers & Curators **Journal Publishers** Librarians & Trainers Societies & Alliances Funders

Editorial

## Ten Simple Rules for the Care and Feeding of Scientific Data

Alyssa Goodman<sup>1</sup>, Alberto Pepe<sup>1\*</sup>, Alexander W. Blocker<sup>1</sup>, Christine L. Borgman<sup>2</sup>, Kyle Cranmer<sup>3</sup>, Merce Crosas<sup>3</sup>, Rosanne Di Stefano<sup>4</sup>, Yolanda Gil<sup>5</sup>, Paul Groth<sup>6</sup>, Margaret Hedstrom<sup>7</sup>, David W. Hogg<sup>3</sup>, Vinay Kashyap<sup>1</sup>, Ashish Mahabal<sup>1</sup>, Aneta Siemiginowska<sup>1</sup>, Aleksandra Slavkovic<sup>8</sup>

<sup>1</sup> Harvard University, Cambridge, Massachusetts, United States of America, <sup>2</sup> University of California, Los Angeles, Los Angeles, California, United States of America, <sup>3</sup> New York University, New York, New York, United States of America, <sup>4</sup> University of Southern California, Los Angeles, Los Angeles, California, United States of America, <sup>5</sup> Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, <sup>6</sup> University of Michigan, Ann Arbor, Michigan, United States of America, <sup>7</sup> California Institute of Technology, Pasadena, California, United States of America, <sup>8</sup> Pennsylvania State University, State College, Pennsylvania, United States of America

### Introduction

In the early 1600s, Galileo Galilei turned a telescope toward Jupiter. In his log book each night, he drew to-scale schematic diagrams of Jupiter and some oddly moving points of light near it. Galileo labeled each drawing with the date. Eventually he used his observations to conclude that the Earth orbits the Sun, just as the four Galilean moons orbit Jupiter. History shows Galileo to be much more than an astronomical pro, though. His clear and accurate journal publication style—on a log-let Galileo understood the system, key cosine law to let astronomers know Galileo did it, Galileo’s notes on the phases of the data (drawing of Jupiter and its moons), key metadata (timing of each observation, weather, and telescope properties), and text (descriptions of method, analysis, and conclusion). Critically, when Galileo included the information from those notes in *Sideways News* [1], this integration of text, data, and metadata was preserved, as shown in Figure 1. Galileo’s work advanced the “Scientific Revolution,” and his approach to observation and analysis contributed significantly to the shaping of today’s modern “scientific method” [2,3].

Today, most research projects are considered complete when a journal article based on the analysis has been written and published. The trouble is, unlike Galileo’s report in *Sideways News*, the amount of real data and data description in modern publications is almost never sufficient to repeat or even statistically verify a study being presented. Worse, researchers wishing to build upon and extend work presented in the literature often have trouble recovering data associated with an article after it has been published. More often than scientists would like to admit, they cannot even recover the data associated with their own published works.

Complicating the modern situation, the words “data” and “analysis” have a wider variety of definitions today than at the time of Galileo. Theoretical investigations can create large “data” sets through simulations (e.g., The Millennium Simulation Project: <http://www.mpa-garching.mpg.de/galform/virgo/millennium/>). Large-scale data collection often takes place as a community-wide effort (e.g., The Human Genome project: <http://www.genome.gov/10001772>), which leads to gigantic online “databases” (ongoing examples of these include GenBank and the Protein Data Bank) and in the reusing of experimental observations (e.g., data, the “code” in Python to do a drawing line between data and “analysis” or “code”) when discussing the care and feeding of “data.” Sometimes, a copy of the code used to create or process data is so essential to the use of those data that the code should almost be thought of as part of the “metadata” description of the data. Other times, the code used in a scientific study is more separable from the data, but even then, many preservation and sharing principles apply to code just as well as they do to data.

So how do we go about caring for and feeding data? Extra work, no doubt, is associated with nurturing your data, but care up front will save time and increase insight later. Even though a growing number of researchers, especially in large collaborations, know that conducting research with

sharing and reuse in mind is essential, it still requires a paradigm shift. Most people are still motivated by piling up publications and by getting to the next one as soon as possible. But, the more we scientists find ourselves wishing we had access to extant but now unfindable data [4], the more we will realize why bad data management is bad for science. How can we improve?

This article offers a short guide to the steps scientists can take to ensure that their data and associated analyses continue to be of value long after they are recognized. In just the past few years, hundreds of scholarly papers and reports have been written on questions of data sharing, data provenance, research reproducibility, licensing, attribution, privacy, and more—but our goal here is not to review that literature. Instead, we present a short guide intended for researchers who want to know why it is important to “care for and feed” data, with some practical advice on how to do that. The final section at the close of this work (Links to Useful Resources) offers links to the types of services referred to throughout the text. **Boldface lettering** below highlights actions one can take to follow the suggested rules.

### Rule 1. Love Your Data, and Help Others Love It, Too

Data management is a repeat-play game. If you take care to make your data

Citation: Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Comput Biol* 10(4): e1003542. doi:10.1371/journal.pcbi.1003542

Editor: Philip E. Bourne, University of California San Diego, United States of America

Published: April 24, 2014

Copyright: © 2014 Goodman et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors received no specific funding for writing this manuscript.

Competing Interests: The authors have declared that no competing interests exist.

\* Email: alberto.pepe@gmail.com

# Article

## Dataset Reuse: Toward Translating Principles to Practice

Laura Koesten,<sup>1,\*</sup> Pavlos Vougiouklis,<sup>2</sup> Elena Simperi,<sup>1</sup> and Paul Groth<sup>3,4,\*</sup>

<sup>1</sup>King's College London, London WC2B 4BG, UK

<sup>2</sup>Huawei Technologies, Edinburgh EH9 3BF, UK

<sup>3</sup>University of Amsterdam, Amsterdam 1090 GH, the Netherlands

<sup>4</sup>Lead Contact

\*Correspondence: [laura.koesten@kcl.ac.uk](mailto:laura.koesten@kcl.ac.uk) (L.K.), [p.groth@uva.nl](mailto:p.groth@uva.nl) (P.G.)

<https://doi.org/10.1016/j.patter.2020.100136>

**THE BIGGER PICTURE** The web provides access to millions of datasets. These data can have additional impact when it is used beyond the context for which it was originally created. We have little empirical insight into what makes a dataset more reusable than others, and which of the existing guidelines and frameworks, if any, make a difference. In this paper, we explore potential reuse features through a literature review and present a case study on datasets on GitHub, a popular open platform for sharing code and data. We describe a corpus of more than 1.4 million data files, from over 65,000 repositories. Using GitHub's engagement metrics as proxies for dataset reuse, we relate them to reuse features from the literature and devise an initial model, using deep neural networks, to predict a dataset's reusability. This work demonstrates the practical gap between principles and actionable insights that allow data publishers and tools designers to implement functionalities that provably facilitate reuse.

**1 2 3 4 5** Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

### SUMMARY

The web provides access to millions of datasets that can have additional impact when used beyond their original context. We have little empirical insight into what makes a dataset more reusable than others and which of the existing guidelines and frameworks, if any, make a difference. In this paper, we explore potential reuse features through a literature review and present a case study on datasets on GitHub, a popular open platform for sharing code and data. We describe a corpus of more than 1.4 million data files, from over 65,000 repositories. Using GitHub's engagement metrics as proxies for dataset reuse, we relate them to reuse features from the literature and devise an initial model, using deep neural networks, to predict a dataset's reusability. This demonstrates the practical gap between principles and actionable insights that allow data publishers and tools designers to implement functionalities that provably facilitate reuse.

### 1 INTRODUCTION

There has been a gradual shift in the last years from viewing datasets as byproducts of (digital) work to critical assets, whose value increases the more they are used.<sup>1,2</sup> However, our understanding of how this value emerges, and of the factors that demonstrably affect the reusability of a dataset is still limited.

At the same time, making data reusable to a diverse audience, in terms of domain, skill sets, and purposes, is an important way to realize its potential value (and recover some of the, sometimes considerable, resources invested in policy and infrastructure support). This is one of the reasons why scientific journals and research-funding organizations are increasingly calling for further data sharing<sup>3</sup> or why industry bodies, such as the Interna-

Feature	Description	References
<b>Access</b>		
License	(1) available, (2) allows reuse	W3C 1,20,45-47
Format/machine readability	(1) consistent format, (2) single value type per column, (3) human as well as machine readable and non-proprietary format, (4) different formats available	W3C2,22,48-50
Code available	for cleaning, analysis, visualizations	51-53
Unique identifier	PID for the dataset/ID's within the dataset	W3C2,55
Download link/API	(1) available, (2) functioning	W3C47,56

### Documentation: Methodological Choices

Methodology	description of experimental setup (sampling, tools, etc.), link to publication or project	3,13,54,60,63,66
Units and reference systems	(1) defined, (2) consistently used	54,67
Representativeness/Population	in relation to a total population	21,60
Details	change classifications, analyze special cases, sample size/coverage/rounding	65,66
Cleaning/pre-processing	(1) cleaning choices described, (2) are the raw data available?	3,13,21,68

Biases/limitations	different types of bias (i.e., sampling bias)	21,49,69
Data management	(1) mode of storage, (2) duration of storage	3,70,71

### Documentation: Quality

Missing values/null values	(1) defined what they mean, (2) ratio of empty cells	W3C22,48,49,59,60
Margin of error/reliability/quality control procedures	(1) confidence intervals, (2) estimates versus actual measurements	54,65
Formatting	(1) consistent data type per column, (2) consistent date format	W3C41,65
Outliers	are there data points that differ significantly from the rest	22
Possible options/constraints on a variable	(1) value type, (2) if data contains an "other" category	W3C72
Last update	information about data maintenance if applicable	21,62
Completeness of metadata	empty fields in the applied metadata structure?	41
Abbreviations/acronyms/codes	defined	49,54

### Connections

Relationships between variables defined	(1) explained in documentation, (2) formulae	21,22
Cite sources	(1) links or citation, (2) indication of link quality	21
Links to dataset being used elsewhere	i.e., in publications, community-led projects	21,59
Contact	person or organization, mode of contact specified	W3C41,73

### Provenance and Versioning

Publisher/producer/repository	(1) authoritativeness of source, (2) funding mechanisms/other interests that influenced data collection specified	21,49,54,59,74,75
Version indicator	version or modification of dataset documented	W3C50,66,76
Version history	workflow provenance	W3C50,76
Prior reuse/advice on data reuse	(1) example projects, (2) access to discussions	3,27,59,60

### Ethics

Ethical considerations, personal data	(1) data related to individually identifiable people, (2) if applicable, was consent given	21,57,71,75
---------------------------------------	--	-------------

### Semantics

Schema/Syntax/Data Model	defined	W3C47,67
Use of existing taxonomies/vocabularies	(1) documented, (2) link	W3C2

Is there any evidence that it works?

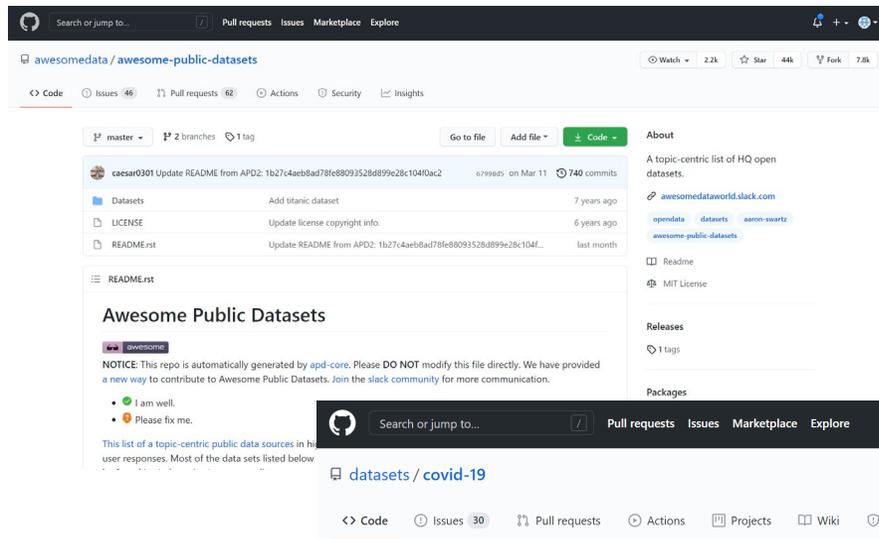
# Be measurable: GitHub as a data platform

~1.4 million datasets (e.g. CSV, excel) from ~65K repos

Map literature features to both dataset and repository features

Use engagement metrics as proxies for data reuse

Train a predictive model to see what publishing guidance leads to higher engagement values



Size

Attributes

actions-user Auto-update of the data packages Latest commit 5612b64 4 hours ago History

3 contributors

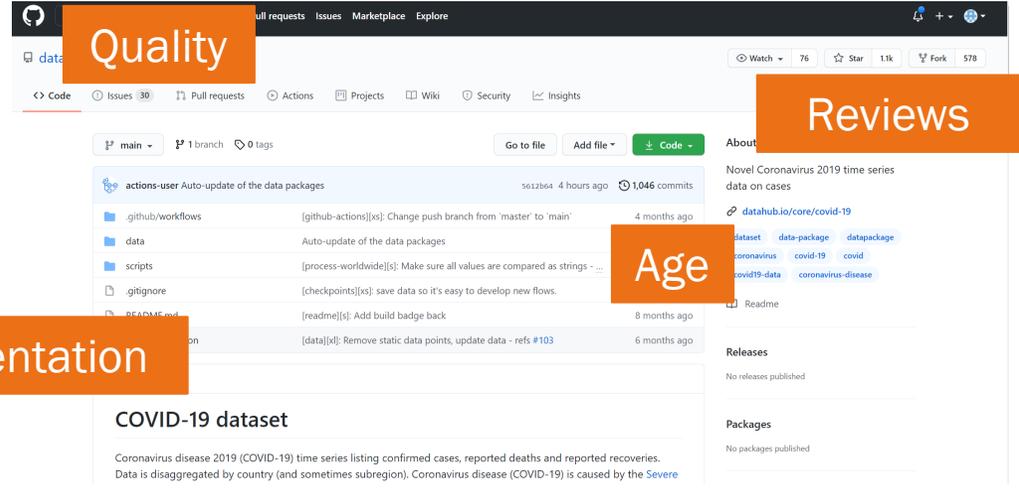
454 Lines (454 sloc) 23.7 KB

Search this file...

Date	Confirmed	Recovered	Deaths	Increase rate
2020-01-22	557	30	17	
2020-01-23	655	32	18	17.594254937163374
2020-01-24	941	39	26	43.66412213740458
2020-01-25	1433	42	42	52.28480340063763

Quality

Reviews



Documentation

Age

## COVID-19 dataset

Coronavirus disease 2019 (COVID-19) time series listing confirmed cases, reported deaths and reported recoveries. Data is disaggregated by country (and sometimes subregion). Coronavirus disease (COVID-19) is caused by the [Severe](#)

# Recommendations for publishers

## Co-locate documentation

- Informative, short text about the dataset
- Comprehensive README file in a structured form, links to further information

## Co-locate tools

- Standard processable file sizes for dataset distributions
- Openable with a standard configuration of a common library (such as Pandas)

**Table 6. Characteristics of the Dataset Corpus and for Four Groups of Reuse: 1 = Lowest Reuse, 4 = Highest Reuse**

Type	Characteristics	Mean G1	Mean G2	Mean G3	Mean G4	Quantile G1	Quantile G2	Quantile G3	Quantile G4
README	no. of words in README (non-code related) <sup>a</sup>	286.2 (± 963.8)	345.1 (± 835.6)	541.9 (± 1,509.7)	801.9 (± 1,808.7)	[6.0, 48.0, 287.0]	[15.0, 125.0, 389.8]	[63.0, 250.0, 626.0]	[151.5, 416.0, 869.0]
	no. of tables <sup>a</sup>	0.0 (± 0.5)	0.1 (± 0.6)	0.1 (± 1.6)	0.3 (± 2.2)	[0.0, 0.0, 0.0]	[0.0, 0.0, 0.0]	[0.0, 0.0, 0.0]	[0.0, 0.0, 0.0]
	no. of code blocks <sup>a</sup>	0.9 (± 3.5)	1.3 (± 4.2)	2.3 (± 6.1)	3.5 (± 8.1)	[0.0, 0.0, 1.0]	[0.0, 0.0, 1.0]	[0.0, 0.0, 2.0]	[0.0, 1.0, 4.0]
	no. of headers <sup>a</sup>	2.3 (± 4.1)	3.6 (± 5.6)	5.3 (± 7.9)	8.8 (± 54.6)	[0.0, 1.0, 3.0]	[1.0, 1.0, 5.0]	[1.0, 3.0, 7.0]	[2.0, 6.0, 10.0]
	no. of URLs <sup>a</sup>	6.0 (± 10.4)	8.1 (± 18.4)	12.8 (± 21.1)	25.2 (± 113.7)	[1.0, 2.0, 8.0]	[1.0, 4.0, 11.0]	[2.0, 8.0, 17.0]	[6.0, 15.0, 28.0]
	no. of images <sup>a</sup>	0.3 (± 1.7)	0.7 (± 5.5)	1.1 (± 4.8)	2.5 (± 6.1)	[0.0, 0.0, 0.0]	[0.0, 0.0, 0.0]	[0.0, 0.0, 1.0]	[0.0, 1.0, 3.0]
Repository	repository size <sup>a</sup>	33,689.8 (± 152,529)	50,916.3 (± 194,154)	70,511.1 (± 225,835)	133,307.1 (± 423,076)	[580.0, 5,386.5, 22,780.2]	[1,230.0, 7,667.0, 33,723.8]	[2,174.5, 14,557.0, 52,912.2]	[4,896.5, 27,393.0, 113,130.0]
	no. of open issues <sup>a</sup>	1.1 (± 10.8)	2.0 (± 13.2)	6.4 (± 21.8)	38.1 (± 163.7)	[0.0, 0.0, 0.0]	[0.0, 0.0, 1.0]	[0.0, 1.0, 4.0]	[0.0, 5.0, 25.0]
	no. of closed issues <sup>a</sup>	1.9 (± 13.5)	7.6 (± 31.7)	38.4 (± 130.8)	3,74.7 (± 1,823.4)	[0.0, 0.0, 0.0]	[0.0, 0.0, 3.0]	[0.0, 2.0, 19.0]	[2.0, 25.0, 175.5]
	description length <sup>a</sup>	6.2 (± 8.3)	7.7 (± 9.2)	8.9 (± 11.2)	9.6 (± 10.2)	[0.0, 4.0, 9.0]	[2.0, 6.0, 11.0]	[4.0, 7.0, 11.0]	[4.0, 7.0, 12.0]
	ratio of data files per repository <sup>a</sup>	8.2 (± 14.0)	7.1 (± 12.7)	5.4 (± 10.9)	3.6 (± 8.7)	[0.2, 2.3, 10.0]	[0.4, 2.2, 7.7]	[0.3, 1.4, 5.3]	[0.1, 0.7, 2.8]
	age of repository (days) <sup>a</sup>	1,467.9 (± 490.0)	1,513.4 (± 545.2)	1,627.7 (± 592.3)	1,725.3 (± 653.0)	[1,067.0, 1,448.0, 1,791.0]	[1,093.2, 1,453.0, 1,816.0]	[1,214.0, 1,562.0, 1,964.0]	[1,256.5, 1,628.0, 2,082.5]
Data File	ratio of problematic files for a standard config (Pandas) <sup>b</sup>	0.3 (± 2.7)	0.4 (± 2.8)	0.3 (± 2.6)	0.2 (± 1.5)	[0.0, 0.0, 0.0]	[0.0, 0.0, 0.0]	[0.0, 0.0, 0.0]	[0.0, 0.0, 0.0]
	average size of data files (csv) <sup>b</sup>	309,999.4 (± 4,314,537)	337,453.3 (± 2,901,912)	532,226.8 (± 3,595,252)	248,120.4 (± 2,268,705)	[1,732.0, 7,017.0, 33,942.0]	[1,419.0, 6,046.5, 53,402.0]	[1,692.0, 10,396.0, 79,279.0]	[4,763.8, 28,315.0, 73,671.0]
	average size of data files(xls(x)) <sup>b</sup>	426,555.6 (± 2,755,034.2)	528,439.2 (± 2,953,938)	360,737.8 (± 2,050,485.3)	330,846.9 (± 1,518,167.8)	[20,430.2, 30,511.0, 83,968.0]	[20,287.0, 45,568.0, 147,138.5]	[16,856.8, 45,056.0, 203,837.5]	[16,896.0, 34,462.0, 95,356.0]
	no. of rows (csv) <sup>b</sup>	3,845.2 (± 50,528)	4,324.6 (± 52,089)	6,221.6 (± 55,637)	3,087.6 (± 35,192.0)	[41.0, 85.0, 569.0]	[33.0, 79.0, 719.0]	[42.0, 147.0, 930.0]	[41.0, 118.0, 293.0]
	no. of columns (csv) <sup>b</sup>	23.3 (± 340.0)	16.3 (± 376.5)	23.7 (± 524.6)	14.7 (± 363.2)	[3.0, 7.0, 18.0]	[2.0, 4.0, 7.0]	[3.0, 6.0, 13.0]	[4.0, 11.0, 11.0]
	no. of rows (xls(x)) <sup>b</sup>	1,337.2 (± 22,013.9)	409.4 (± 10,184.4)	324.2 (± 8,992.9)	1,105.0 (± 16,615.8)	[26.0, 64.0, 141.0]	[64.0, 86.0, 122.0]	[19.0, 31.0, 52.0]	[20.0, 46.0, 176.0]
no. of columns (xls(x)) <sup>b</sup>	29.8 (± 397.2)	36.2 (± 531.0)	23.8 (± 155.0)	25.6 (± 423.3)	[5.0, 9.0, 16.0]	[19.0, 19.0, 19.0]	[9.0, 12.0, 16.0]	[6.0, 10.0, 15.0]	
missing values ratio (csv) <sup>b</sup>	8.7 (± 16.6)	7.2 (± 19.1)	10.5 (± 20.5)	13.0 (± 13.6)	[0.0, 0.0, 11.3]	[0.0, 0.0, 0.0]	[0.0, 0.0, 11.7]	[0.0, 19.0, 19.8]	

Quantiles values are reported in the [X<sub>25</sub>, X<sub>50</sub>, X<sub>75</sub>] format, where X<sub>25</sub>, X<sub>50</sub> and X<sub>75</sub> represent the 25th, 50th, and 75th quantile of a particular group's characteristic.

<sup>a</sup>Indicates statistically significant differences (p ≤ 0.05) of pairwise comparisons across all four groups.

<sup>b</sup>Denotes cases for which statistical significant differences are observed between the values of groups 1 and 4 but not necessarily between the rest of pairwise comparisons.

30 data sets found

Football-related arrests and banning orders, England and Wales: 2021 to 2022 season



Football-related arrests and banning orders, England and Wales: 2021 to 2022 season

Explore at: gov.uk

2 scholarly articles cite this dataset (View in Google Scholar)

Data set updated

Sep 22, 2022

Dataset provided by

GOV.UK

Statistics on football-related arrests and football banning...

# Can people find the data they need?

MAYOR OF LONDON LONDON ASSEMBLY

## LONDON DATASTORE

Login

- Data Analysis Collaboration COVID-19 Area Profiles Blog Guidance About

Home / Datasets

### Topics

(No further facets)

### Formats

(No further facets)

football

Did you mean: [footfall](#)

No datasets found

About 4,670,000 results (0.35 seconds)

DataHub.io Football Data - DataHub.io

13 Dec 2017 — Football Data. A collection of awesome football datasets including national teams, clubs, match schedules, players, stadiums, etc.

Kaggle European Soccer Database - Kaggle

The ultimate Soccer database for data analysis and machine learning. What you get: +25,000 matches; +10,000 players; 11 European Countries with their lead ...

International football results from 1872 to 2022 | Kaggle

Data.world There are 38 soccer datasets available on data.world

### Datasets

European Soccer Database

2021-2022 Football Player Stats

European Soccer Data

# Analysis of logs and data requests (2018)

- Four national open government data portals, 2.2 million queries (2013 – 2016), 1500 data requests.
- Shorter queries, include **temporal and location information**.
- Explorative search.
- **Native and external queries topically different**.
- Data requests offer more context to user intent.



# Analysis of logs (2020 - 21)

---

844k sessions from 04/2018 to 06/2020, web search as well as native search sessions from the European Data Portal

**Location**, provenance, format, licence, time frame and date, publishing date, location of publication and data schema

**Mostly web search**, web search and native search users have **different information needs** and **different success rates**

**Dataset preview page** is important in web search

**Linking to stories** and other content helps with traffic

## Latest Reports



### Report 18: Characterising dataset search on the European Data Portal

28/09/2020

This report illustrates a quantitative study on data search through more than two years of EDP search and interaction logs. Understanding data search behaviour is key to developing better search algorithms and improving the search experience. This study presents current findings from key literature in dataset search. [Read more](#)

## Latest Reports



### Analytical Report 19: Understanding supply and demand on the European Data Portal

24/02/2021

The European Data Portal's 19th analytical report provides a detailed analysis of the relationship between supply and demand of datasets on the EDP. [Read more](#)

# Recommendations for publishers

---

Two types of users

Spatial and temporal queries

Result presentation

Quality reviews

Data stories

More logs needed!

This forum provides a space to engage with the challenges of designing for intelligent algorithmic experiences. We invite articles that tackle the tensions between research and practice when integrating AI and UX design. We welcome interdisciplinary debate, artful critique, forward-looking research, case studies of AI in practice, and speculative design explorations. — Juho Kim and Henriette Cramer, Editors

# UX of Data: Making Data Available Doesn't Make It Usable

Laura Koesten, University of Vienna, Elena Simperl, King's College London

Data plays an important part in our daily lives. It shapes how we view the world and, for better or worse, informs the decisions we make. Despite controversies around who collects and does what with data, data-centric work is seen as a critical tool for solving the most complex problems of our time, from climate change and security to health crises and inequality. It spans many types of tasks in various professions, undertaken by multiple stakeholders who may or may not share common understandings of the domain or the task. It also raises questions of data literacy, inclusion, and fairness in ensuring that the value the data creates is shared as widely as possible.

In our research over the past five years, we have asked what makes interaction and working with data unique for different types of audiences.

Data is increasingly available online, for instance, in science and government. Organizations invest in infrastructure to share data internally and with their partners. Machine learning is heavily dependent on the availability of datasets to train algorithms, but creating these from scratch comes with costs. Overall, it means that more and more data is used outside the context in which or for which it was produced—reuse is often the best way to add value to the data and perhaps to recover some of the investment that went into publishing it and policy support.

But how comfortable are we with reading, interpreting, and working

with other people's data? Using a dataset beyond the context in which it originated remains challenging. Simply making data available, even when following existing guidelines and best practices, does not mean it can be easily used by others. Further, there is little evidence that such guidelines, including technical standards and community support, truly lead to greater data engagement and generate value.

This is not surprising if we consider the complex environment around data. Data-centric work is often carried out in diverse teams, drawing upon different skills, including domain knowledge, statistics, data science, engineering, and project management. While some data tasks are heavily automated, others depend on people's judgments and their decisions on how and what to record, document, and analyze, and what to leave out. Similar to other fields, there is a tension between the resources required to document each step in detail and

### Insights

- We need better ways to record and communicate that data represents only a perspective of reality.
- There is no typical data user, just typical data tasks. Data tools need to provide better support for each task in the data life cycle.
- Interaction challenges for making data usable include making data understandable, supporting collaboration, managing changes, and facilitating different modes of access and user engagement.

the ability at a later point in time to reproduce what happened. This is especially the case as deep, exploratory or data-curation work is often seen as secondary to model design or machine learning.

In our research, we learned a great deal about how different audiences go about using data, including how they choose what data to work with, and how they make sense of data that others have created. We explored scenarios where these tasks were carried out by individuals as well as in teams of data professionals. Using both qualitative and quantitative methods, we carried out several studies drawing upon a diverse range of data sources, including interviews, diaries, surveys, and digital traces of user engagement on data platforms such as Kaggle, GitHub, and open-data portals.

Our initial exploratory studies were framed in theories about information seeking: We asked people what their data tasks were—what they did with data in their jobs, how they searched for it, how they evaluated and selected datasets as search results, and how they explored and understood data with which they were not familiar.

We also used different forms of summarization (verbal and textual) to document the cognitive processes involved in data sensemaking. We asked people to describe and discuss data they knew or data that was new to them, and identified common activity patterns [1]. We could distinguish among three clusters: *inspecting*, *engaging with content*, and *placing data in context*. Our research suggests that

The csv file contains 225 records of stop and searches carried out by Greater Manchester police in February 2017. It contains 15 columns which give detail on the people stopped - including ethnicity, gender (male/female) and age (ranges), the stop itself (including type, location (long/lat), legislation, object of search and whether it was part police operation (true/false)), and the outcome of the stop. The majority of variables are complete for all records, however details on both gender and age range are missing for about 40 values each. In the majority of cases the outcome was that there was nothing found, in 55 cases the outcome was true and therefore action was taken (e.g. suspect arrested).

- 1. subtitle
- 2. format
- 3. headers
- 4. value types / ranges
- 5. provenance
- 6. temporal information
- 7. geospatial information
- 8. quality statements
- 9. analysis / highlights



# Conclusions

---

We are at a crucial moment in data availability and use, online and elsewhere.

There is an increasing body of evidence about what people's data needs and about how data is published on the web.

We don't have links and we don't always have great business cases for creating and maintaining them on the open, decentralised web. In fact, with the rise of generative AI we need to re-think business models for open data publishing yet another time.

Standards emerge and are adopted when they solve a real problem. Technology can help rather than hinder and should never take the centre stage.

arXiv:1803.09010v3 [cs.DB] 9 Jun 2018

Datasheets for Datasets

Timnit Gebru<sup>1</sup> Janie Morgenstern<sup>2</sup> Briana Veitch<sup>3</sup> Jennifer Wortman Vaughan<sup>3</sup> Hanna Wallach<sup>3</sup> Hal Daumé III<sup>1,4</sup> Kate Crawford<sup>1,5</sup>

**Abstract**

The machine learning community has no standardized way to document how and why a dataset was created, what information it contains, what tasks it should and should not be used for, and whether it might raise any ethical or legal concerns. To address this gap, we propose the concept of datasheets for datasets. In the electronics industry, it is standard to accompany every component with a datasheet providing standard operating characteristics, test results, recommended usage, and other information. Similarly, we recommend that every dataset be accompanied with a datasheet documenting its creation, composition, intended uses, maintenance, and other properties. Datasheets for datasets will facilitate better communication between dataset creators and users, and encourage the machine learning community to prioritize transparency and accountability.

We therefore propose the concept of datasheets for datasets. In the electronics industry, every component is accompanied by a datasheet describing standard operating characteristics, test results, and recommended usage. By analogy, we recommend that every dataset be accompanied with a datasheet documenting its motivation, creation, composition, intended uses, distribution, maintenance, and other information. We anticipate that such datasheets will increase transparency and accountability in the machine learning community.

Section 2 provides context for our proposal. Section 3 discusses the evolution of safety standards in other industries, and outlines the concept of datasheets in electronics. We give examples of questions that should be answered in datasheets for datasets in Section 4, and discuss challenges and future work in Section 5. The appendix includes a more complete proposal along with prototype datasheets for two well-known datasets: Labeled Faces in the Wild (Huang et al., 2007) and Pang and Lee’s polarity dataset (2004).

**2. Context**

A foundational challenge in the use of machine learning is the risk of deploying systems in unsuitable environments. A model’s behavior on some benchmark may say very little about its performance in the wild. Of particular concern are recent examples showing that machine learning systems can amplify existing societal biases. For example, Buolamwini & Gebru (2018) showed that commercial gender classification APIs have near perfect performance for lighter-skinned males, while error rates for darker-skinned females can be as high as 35%. Rotkiewicz et al. (2016) showed that word embeddings trained on news articles exhibit gender biases, finishing the analogy “man is to computer programmer as woman is to X” with “homemaker”, a stereotypical role for women. Caliskan et al. (2017) showed these embeddings also contain racial biases: traditional European-American names are closer to positive words like “joy”, while African-American names are closer to words like “agony.”

These biases can have dire consequences that might not be easily discovered. Much like a faulty resistor or a capacitor in a circuit, the effects of a biased machine learning

<sup>1</sup>Microsoft Research, New York, NY <sup>2</sup>Georgia Institute of Technology, Atlanta, GA <sup>3</sup>Cornell University, Ithaca, NY <sup>4</sup>University of Maryland, College Park, MD <sup>5</sup>AI Now Institute, New York, NY  
Correspondence to: Timnit Gebru (tgebru@gmail.com).

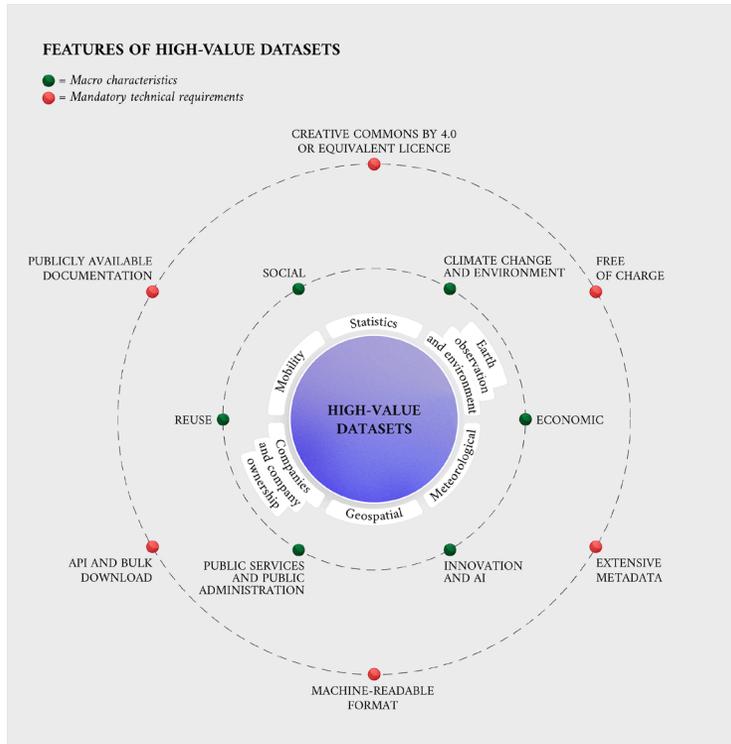
*Proceedings of the 25th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

Metadata vocabularies are used where there is a clear business case. More documentation needed to make them useful more widely.

# The Data Cards Playbook



A toolkit for transparency in AI dataset documentation.



(Source: data.europa.eu, Federica Fragapane, 2022)

### Smart Rural: The Open Data Gap

Johanna Walker<sup>1\*</sup>, Gefion Thuermer<sup>2</sup>, Elena Simper<sup>3</sup> and Les Carr<sup>1</sup>

<sup>1</sup>University of Southampton, <sup>2</sup>King's College London  
\* Johanna.Walker@soton.ac.uk

#### Abstract

The smart city paradigm has underpinned a great deal of the use and production of open data for the benefit of policy makers and citizens. This paper posits that this further enhances the existing urban-rural divide. It investigates the availability and use of rural open data along two parameters: pertaining to rural populations, and to key parts of the rural economy (agriculture, fisheries and forestry). It explores the relationship between key statistics of national / rural economies and rural open data, and the use and users of rural open data where it is available. It finds that although countries with more rural populations are not necessarily earlier in their Open Data Maturity journey, there is still a lack of institutionalisation of open data in rural areas; that there is an apparent gap between the importance of agriculture to a country's GDP and the amount of agricultural data published openly; and lastly, that the smart city paradigm cannot simply be transferred to the rural setting. It suggests instead the adoption of the emerging 'smart region' paradigm as that most likely to support the specific data needs of rural areas.

**Keywords** – Open data; agriculture; European Data Portal; rural open data.

#### 1 Introduction

The urban-rural divide is one of the most well-known foci of digital inequality (Vicente Cuervo & López Menéndez, 2006). In comparison to their urban or suburban counterparts, rural populations tend to experience gaps in education, income, device availability, and mobile and internet access. The Eurostat Rural Development Statistics<sup>1</sup> illuminate points of difference in underlying challenges in rural and urban areas across issues such as health, housing and education, which mean that the digital demands in each of these domains in rural areas differ from those experienced in the urban setting.

Twenty seven percent of EU citizens live in rural areas<sup>2</sup>. The more densely populated EU member states have a very high proportion of urban population, but the opposite is true for many countries, especially in Eastern Europe. Over half of Lithuanians and nearly half of Slovenians, Hungarians and

Croatians live in rural localities. However, even in some Western member states, such as Ireland and France over a third of the population are classified as living rurally. Although in Western and Northern member states rural living offers comparatively better housing and relatively available employment, in Eastern, Southern and Baltic member states, rural citizens tend to be among the poorest. Being on the wrong side of the digital divide therefore is exacerbated by rural poverty, and, we argue, potentially being excluded from the beneficial possibilities of open data.

Cities and urban areas across both the Global North and South have become avid users and sources of data in a variety of ways, but generally under the umbrella term of 'smart cities'. Conversely, the Rural Open Data Project<sup>3</sup> states that, "few if any rural local governments provide open data, and little is known about how open data affects rural communities. If there is a benefit to communities from open data policies, it is likely that rural communities are benefiting less than urban ones, if at all." Yet data is perceived to be so vital for cities and citizens that the Urban Data Platform<sup>4</sup> project has the aim of speeding up the adoption of common open urban data platforms and ensuring that "300 million European citizens are served by cities with competent urban data platforms, by 2025". Open data is at the very heart of the smart cities concept (Ojo et al., 2015).

In this short paper we attempt to define what rural data might consist of and identify some dimensions of its availability and use. We also engage with the smart city paradigm to explore its use for investigating 'smart rural' open data.

#### 2 Defining Rural

How to define 'rural' is virtually a research area of its own. As Wineman et al. (n.d. forthcoming) show, there are a variety of ways to define rural, and the different definitions affect certain indicators. Our approach is based on the quotidian dictionary definition of 'pertaining to the countryside, those living in the countryside, or agriculture' (Merriam-Webster, 2019). This therefore led us to explore both population (those living in the countryside as defined

<sup>1</sup> <https://ec.europa.eu/eurostat/web/rural-development-data/database>  
<sup>2</sup> [https://ec.europa.eu/eurostat/statistics-explained/images/4/49/Focus\\_on\\_rural\\_areas\\_RVR2017.xlsx](https://ec.europa.eu/eurostat/statistics-explained/images/4/49/Focus_on_rural_areas_RVR2017.xlsx)

<sup>3</sup> <https://www.ruralopendata.eu>  
<sup>4</sup> <https://en.smart-cities.eu/index.php/68/description>

Some data is missing, with serious consequences

 find me a dataset of bus timetables in berlin

 One possible source for a dataset of bus timetables in Berlin is the website of the Berliner Verkehrsbetriebe (BVG), the public transport operator in Berlin. They provide various datasets related to public transport, including timetables for buses, trains, and trams.

To access the BVG datasets, you can follow these steps:

1. Go to the BVG open data website at [https://daten.berlin.de/datensaetze?amp%3Bfield\\_categories\\_tid=237&field\\_categories\\_tid=237](https://daten.berlin.de/datensaetze?amp%3Bfield_categories_tid=237&field_categories_tid=237).
2. Scroll down to the "Verkehrsdaten" section and click on "Fahrplandaten Bus, Tram, U-Bahn, S-Bahn".
3. Scroll down to the "Daten" section and click on "Fahrplandaten".
4. You will be prompted to create an account or log in to your existing account on the Berlin Open Data platform. Follow the instructions to create an account if you don't have one already.
5. Once you are logged in, you can download the dataset for bus timetables in Berlin by clicking on the "Daten herunterladen" button.

 What are the main attributes in this CSV file: <https://data.london.gov.uk/download/fires-in-london---cause-of-ignition-is-white-goods/a2921da4-b1e4-4452-a54c-86b2b4f341a7/Fires%20in%20white%20goods%20from%202009.xlsx>

 The file you provided is in XLSX format, but assuming it has been converted to a CSV file with the same data, the main attributes in the file are likely to be as follows:

- Incident Number: A unique identifier for each fire incident.
- Date Of Incident: The date on which the fire incident occurred.
- Postcode District: The first part of the postcode for the location of the fire incident.
- Property Type: The type of property where the fire incident occurred (e.g., house, flat, commercial building, etc.).
- Property Category: The category of property where the fire incident occurred (e.g., dwelling, non-dwelling, vehicle, etc.).
- White Goods Brand: The brand of the white goods appliance that was involved in the fire incident (e.g., Hotpoint, Indesit, Beko, etc.).
- White Goods Model: The model of the white goods appliance that was involved in the fire incident.
- White Goods Type: The type of white goods appliance that was involved in the fire incident (e.g., fridge, freezer, washing machine, etc.).
- Location: The location within the property where the white goods appliance was located.
- Cause Of Ignition: The cause of ignition for the fire incident, which in this dataset is limited to white goods appliances.
- Total Incidents: The total number of fire incidents that involved white goods appliances for the given combination of attributes (i.e., brand, model, type, and cause of ignition).

## SPARQL Search

All metadata of data.europa.eu is stored as RDF triples (RDF) and can be queried using SPARQL query language at this endpoint. Queries can be executed via the query editor (below) or via the machine-readable endpoint (<https://data.europa.eu/sparql>).

To assist users in constructing useful SPARQL queries we provide a page with featured sample SPARQL queries that can be executed via our endpoint. Datasets and catalogues published as RDF triples on the portal follow the DCAT-AP specification. For further reading on the metadata ontology, please refer to the [data.europa.eu academy page on DCAT-AP](https://data.europa.eu/academy/dcat-and-dcat-ap) at: <https://data.europa.eu/academy/dcat-and-dcat-ap>.

Query

```

1 • PREFIX dcat: <http://www.w3.org/ns/dcat#>
2 PREFIX odp: <http://data.europa.eu/euodp/ontologies/ec-odp#>
3 PREFIX ddt: <http://purl.org/dc/terms/>
4 PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
5 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
6
7 SELECT * WHERE { ?d a dcat:Dataset } LIMIT 10
    
```

# Generative AI as a tool to support data sensemaking and reuse

# Thank you

Talking Datasets: understanding data sensemaking behaviours. L Koesten, K Gregory, P Groth, E Simperl. International Journal of Human-Computer Studies, 146:102562, 2021

Everything You Always Wanted to Know about a Dataset: Studies in Data Summarisation. L Koesten, E Simperl, E Kacprzak, T Blount, J Tennison. International Journal of Human-Computer Studies. 2019

Collaborative Practices with Structured Data: Do Tools Support what Users Need? L Koesten, E Kacprzak, E Simperl, J Tennison. ACM CHI Conference on Human Factors in Computing Systems, 2019.

Dataset search: a survey. A Chapman, E Simperl, L Koesten, G Konstantinidis, LD Ibáñez, E Kacprzak, P Groth. The International Journal on Very Large Data Bases, 2019.

Characterising dataset search — An analysis of search logs and data requests. E Kacprzak, L Koesten, LD Ibáñez, T Blount, J Tennison, E Simperl. Journal of Web Semantics, 2018

Making sense of numerical data-semantic labelling of web tables. Kacprzak, E., Giménez-García, J.M., Piscopo, A., Koesten, L., Ibáñez, L.D., Tennison, J. and Simperl, E. In European Knowledge Acquisition Workshop (pp. 163-178), Springer, 2018

The Trials and Tribulations of Working with Structured Data - a Study on Information Seeking Behaviour. L Koesten, E Kacprzak, J Tennison, E Simperl. ACM CHI Conference on Human Factors in Computing Systems, 2017

Dataset Reuse: Toward Translating Principles to Practice. L Koesten, P Vougiouklis, E Simperl, P Groth. Patterns, 2020

A comparison of dataset search behaviour of internal versus search engine referred sessions. L D Ibáñez, and E Simperl. ACM SIGIR Conference on Human Information Interaction and Retrieval (pp. 158-168), 2022

Characterising Dataset Search on the European Data Portal . L Ibáñez, L Koesten, E Kacprzak, E Simperl. European Data Portal Analytical Report 18, 2020

Understanding Supply and Demand on the European Data Portal. L Ibáñez, E Simperl. European Data Portal Analytical Report 19, 2020

The Future of Open Data Portals. J Walker, E Simperl. European Data Portal Analytical Report 8, 2017

Smart Rural: The Open Data Gap. J Walker, G Thuermer, E Simperl, L Carr. Data for Policy, 2020.