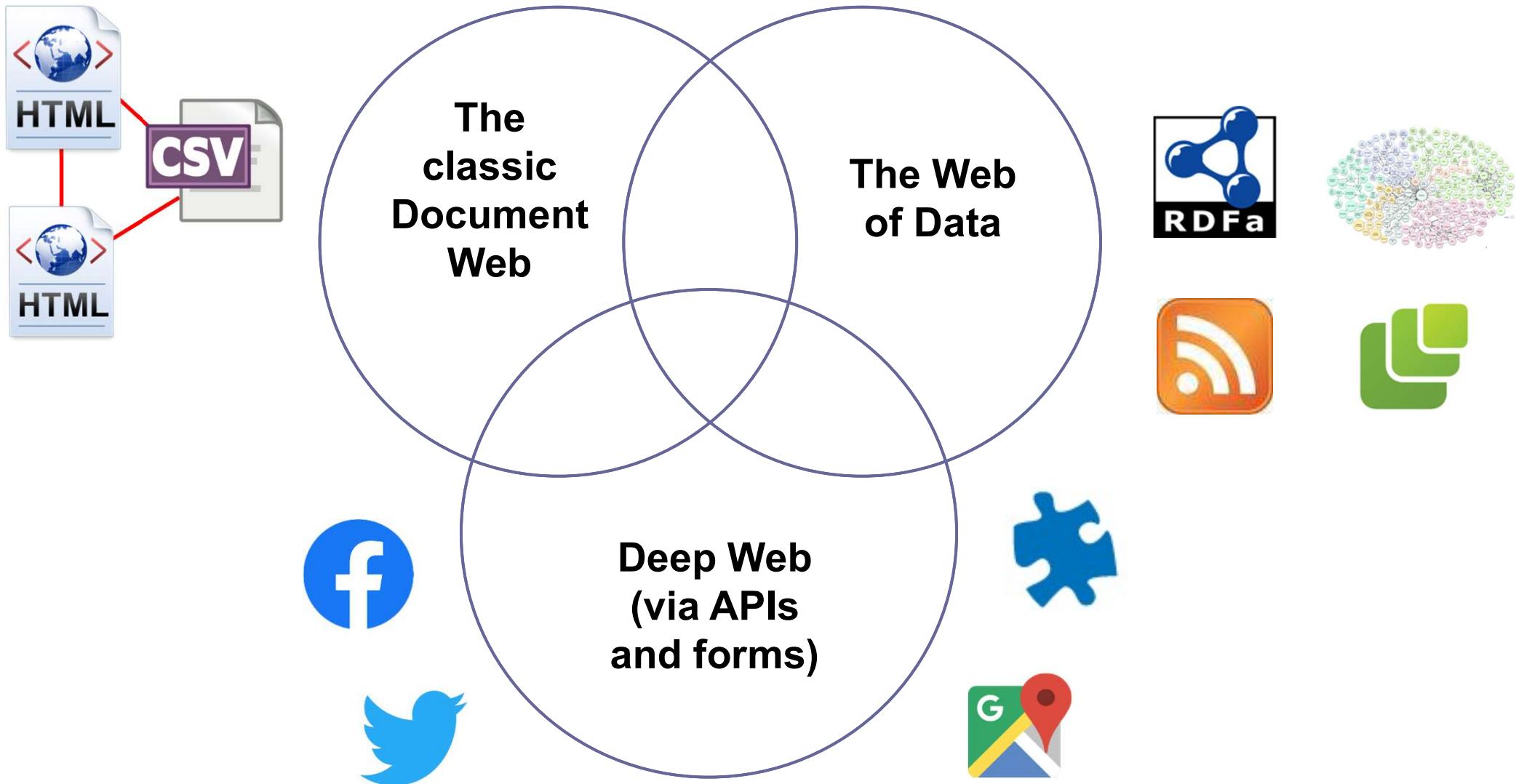


Web Data Integration

Types of Structured Data on the Web



Topology of the Web Today

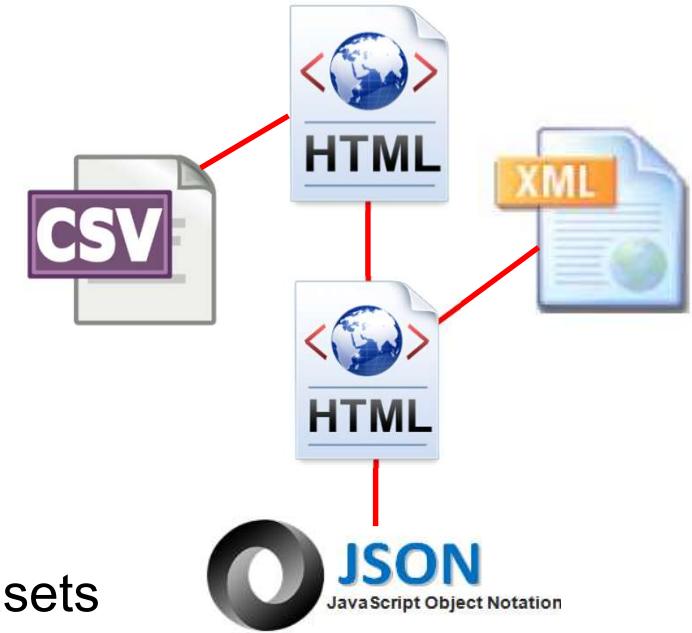


Outline

1. Data Portals
2. Web APIs
3. Linked Data
4. HTML-embedded Data
 1. RDFa, Microdata, JSON-LD
 2. HTML Tables and Templates
 3. Wikipedia as Data Source

1. Data Portals

- The Web traditionally contains structured data in various formats:
 - CSV files, Excel worksheets
 - XML documents, JSON, SQL dumps
- Data Portals and Data Marketplaces
 - collect and host datasets
 - collect and generate metadata describing the datasets
 - provide for data search and exploration
 - provide free or payment-based access to data
- List of Data Catalogs
 - <http://data.wu.ac.at/portalwatch/portalslist>



Main Types of Shared Data

- Public Sector Data
 - Goal: Make data publicly accessible which has been generated by public sector institutions.
 - Laws in many western countries require institutions to publish data
 - Types of data: maps, population statistics, economic data, health data
- Research Data
 - Goal: Accelerate innovation by sharing research data and making research results reproducible
 - Institutional repositories, national research data infrastructures, topical portals
- Commercial Data
 - Goal: Earn money by collecting, cleaning, and integrating data
 - Types of data: data about consumers, business partners, locations
 - Examples of commercial providers: data.world, Foursquare

Example: Government Data Portal

Daten **Showroom** **SPARQL** **Informationen** **Blog**

Das Datenportal für Deutschland

Open Government: Verwaltungsdaten transparent, offen und frei nutzbar

Nach Datensätzen suchen

Suchen

Erweiterte Suche

Kartensuche

62465
Datensätze

24
Anwendungen

106
Blogbeiträge

Govdata
@govdata_de 03. Aug. 2022
Die @OKFN sucht einen
Projektmanager, der dann unter
anderem auch unsere Arbeit kritis...

DCAT-AP.de Version 2.0 veröffentlicht

Die Geschäfts- und Koordinierungsstelle GovData ist verantwortlich für den allgemeinen deut:
Verwaltungsdaten DCAT-AP.de. Ab sofort ist die neueste Version von DCAT-AP.de verfügbar

<https://www.govdata.de/>

Mannheim

Suchen

Erweiterte Suche

Kartensuche

616 Treffer



Erweiterte Suche

Alles 616

Daten 616

Anwendungen 0

Informationen 0

Blog-Beiträge 0

Sortieren nach: Relevanz absteigend ▾

Filtermöglichkeiten

[zurücksetzen](#)

KATEGORIEN

- Bevölkerung und Gesellschaft 190
- Bildung, Kultur und Sport 8
- Gesundheit 22
- Landwirtschaft, Fischerei, Forstwirtschaft und Nahrungsmittel 2
- Regierung und öffentlicher Sektor 67
- Regionen und Städte 5
- Umwelt 65
- Verkehr 22
- Wirtschaft und Finanzen 265



Datensatz

Geologische Übersichtskarte der Bundesrepublik Deutschland 1:200.000 (GÜK200) - CC 7110 Mannheim

Auf Blatt Mannheim ist der nördliche Oberrheingraben mit seinen mesozoischen Flanken dargestellt. Die dominierende Baueinheit im Kartenausschnitt ist der Oberrheingraben. Er durchzieht von Südsüdwest...

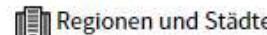
Dateiformate:
[PDF](#)

Letzte Änderung:
17.08.2022

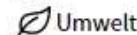
Zeitraum:
-

Veröffentlichende Stelle: Bundesanstalt für Geowissenschaften und Rohstoffe

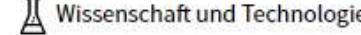
Kategorie:



Regionen und Städte



Umwelt

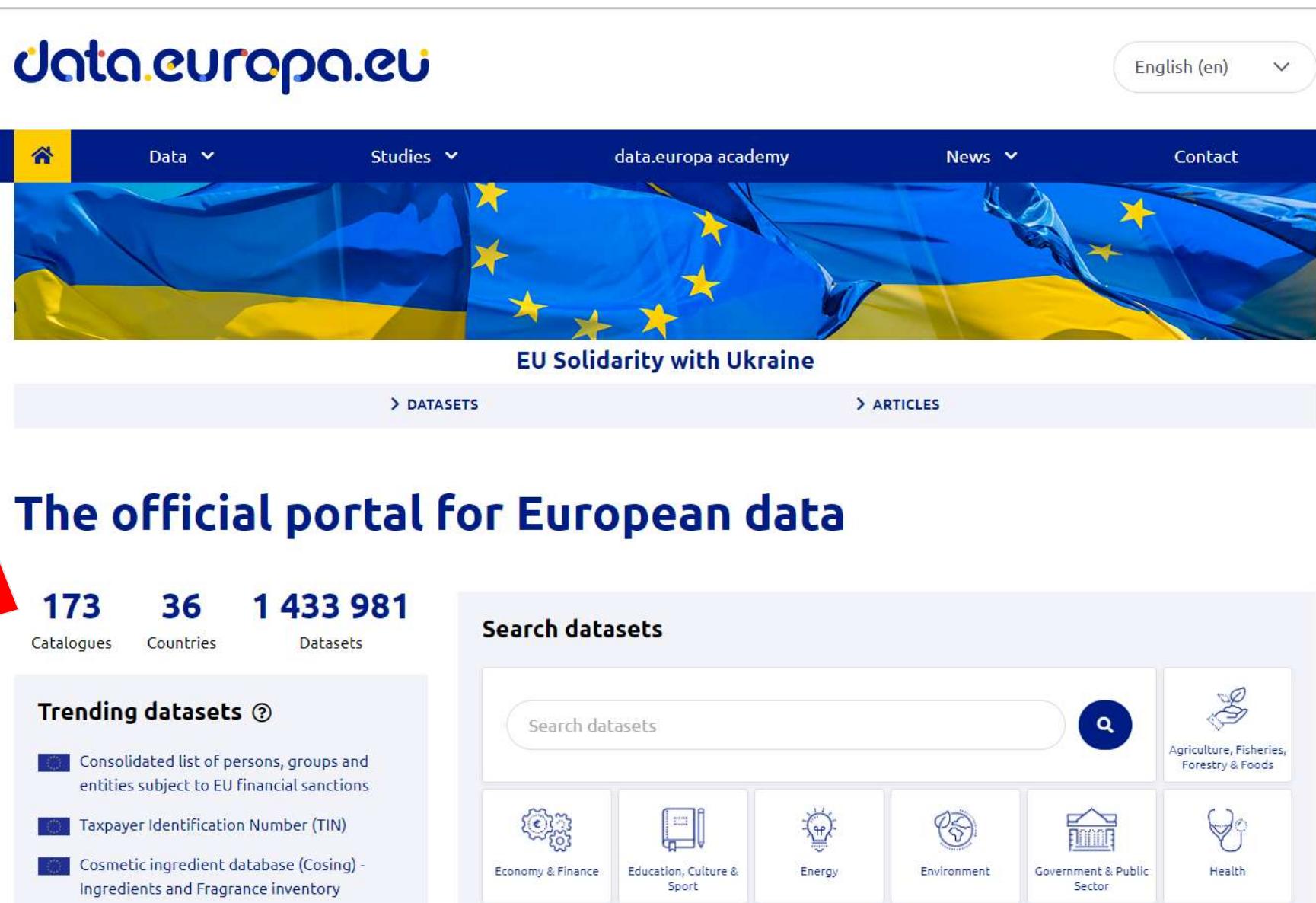


Wissenschaft und Technologie

Offenheit der Lizenz: Freie Nutzung

[zum Datensatz](#)

Example: Portal aggregating Metadata from other Portals



The official portal for European data

173 Catalogues 36 Countries 1 433 981 Datasets

Trending datasets

- Consolidated list of persons, groups and entities subject to EU financial sanctions
- Taxpayer Identification Number (TIN)
- Cosmetic ingredient database (CosIng) - Ingredients and Fragrance inventory

Search datasets

Search datasets

Economy & Finance Education, Culture & Sport Energy Environment Government & Public Sector Health

Agriculture, Fisheries, Forestry & Foods

<https://data.europa.eu/en>

Example: Institutional Research Data Repository

The screenshot shows the homepage of the MADATA Mannheim Research Data Repository. At the top left is the logo 're:search' with a magnifying glass icon. Next to it is the 'UNIVERSITÄT MANNHEIM' logo. In the center, the title 'MADATA' and 'Mannheim Research Data Repository' is displayed. At the top right is the 'UB' logo for Universitätsbibliothek Mannheim. Below the header, there is a navigation bar with links: Home, Publish Data, Browse Repository, Search Repository, About this Repository, and Statistics. A search input field and a 'Search' button are also present. A 'Login' link is located on the left side of the main content area. The main content area displays a title 'Product Datasets from the MWPD2020 Challenge at the ISWC2020 Conference (Task 1)'. Below this, there is a detailed table of metadata:

Item Type:	Dataset
Title:	Product Datasets from the MWPD2020 Challenge at the ISWC2020 Conference (Task 1)
Alternative Title:	Product Data Matching Task derived from the WDC Product Data Corpus Large-Scale Product Matching - Version 2.0 used for the MWPD2020 Challenge at the ISWC2020 Conference
Date:	November 2020
Creator :	Bizer, Christian ; Peeters, Ralph ; Primpeli, Anna
Divisions:	School of Business Informatics and Mathematics > Wirtschaftsinformatik V (Bizer)
DDC Classification:	004 Computer science, internet
Keywords:	schema.org ; product matching ; entity matching ; identity resolution ; record linkage ; e-commerce
The goal of Task 1 of the Mining the Web of Product Data Challenge (MWPD2020) was to compare the performance of methods for identifying offers for the same product from different e-shops. The datasets that are provided to the participants of the competition contain product offers from different e-shops in the form of binary product pairs (with corresponding label "match" or "no	

<https://madata.bib.uni-mannheim.de/>

Example: Focused Research Data Portal

COVID-19 Data Portal

About ▾ Tools ▾ FAQ Related Resources Bulk Downloads Submit Data

Viral Sequences Host Sequences Expression Proteins Networks Samples Cohorts Imaging Literature

COVID-19 Data

Accelerating research through data sharing

Search Examples: ACE2 , Severe acute respiratory syndrome 2 ... Advanced search

Viral sequences →
Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses.
14,296,968 records >

Host sequences →
Raw and assembled sequence and analysis of human and other hosts.
30,605 records >

Expression →
Gene and protein expression data of human genes implicated in the virus infection of the host cells. Identifying cell types and genes with highest expression in SARS-CoV-2 infections.

Proteins →
Curated functional and classification data on the SARS-CoV-2 protein entries and associated protein receptors.
3,639 records >

Latest news →

COVID-19 Data Accelerating research through data sharing

Viral sequences → Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses. 14,296,968 records >

Host sequences → Raw and assembled sequence and analysis of human and other hosts. 30,605 records >

Latest news → Global Search now available on the COVID-19 Portal

23 Aug 2022

<https://www.embl.org/news/science/embl-ebi-launches-covid-19-data-portal/>



Example: Data Portal facilitating the Replication of Research

Search 

Browse State-of-the-Art Datasets Methods More ▾

Datasets

6,785 machine learning datasets

Share your dataset with the ML community!

6785 dataset results

Search for datasets 

Best match 

Filter by Modality

Images	2011
Texts	1790
Videos	645
Audio	420
Medical	238

Filter by Task

Question Answering	292
--------------------	-----

CIFAR-10
The CIFAR-10 dataset (Canadian Institute for Advanced Research) consists of 60000 32x32 color images in 10 classes, with 5000 images per class. It has been widely used for benchmarking computer vision algorithms.
9,345 PAPERS • 58 BENCHMARKS

ImageNet
The ImageNet dataset contains 14,197,122 annotated images. Since 2010 the dataset is used in the ImageNet Large Scale Visual Recognition Competition (ILSVRC).
8,910 PAPERS • 87 BENCHMARKS

COCO (Microsoft Common Objects in Context)
The MS COCO (Microsoft Common Objects in Context) dataset is a large-scale dataset for image captioning, image segmentation, key-point detection, and caption generation.
6,289 PAPERS • 77 BENCHMARKS

MNIST
The MNIST database (Modified National Institute of Standards and Technology digit database) is a large collection of handwritten digits. It has a training set of 55,000 examples and a test set of 5,000 examples.

Entity Resolution on Abt-Buy

Leaderboard Dataset

View F1 (%) by Date

F1 (%)

2018 2020 2022

DeepMatcher - Hybrid Ditto RoBERTa-SupCon

Other models Models with highest F1 (%)

Filter: Splits: Deepmatcher Splits: Own zero-shot untagged

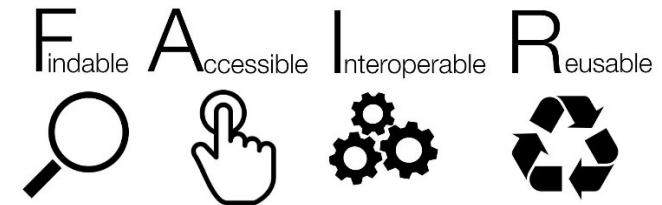
Edit Leaderboard

Rank	Model	F1 (%)	Paper	Code	Result	Year	Tags
1	RoBERTa-SupCon	94.29	Supervised Contrastive Learning for Product Matching			2022	Splits: Deepmatcher

<https://paperswithcode.com/datasets>

The FAIR Data Principles

- **Findable**
 - (Meta)data are assigned a globally unique identifier
 - Data are described with rich metadata
 - (Meta)data are registered or indexed in a searchable resource
- **Accessible**
 - (Meta)data are retrievable by their identifier using a standardised communications protocol
 - Metadata are accessible, even when the data are no longer available
- **Interoperable**
 - (Meta)data use a formal, broadly applicable language for knowledge representation
 - (Meta)data use vocabularies that follow FAIR principles
 - (Meta)data include qualified references to other (meta)data
- **Reusable**
 - (Meta)data are released with a clear data usage license
 - (Meta)data are associated with detailed provenance
 - (Meta)data meet domain-relevant community standards



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

Example: Dataset Search Engine

The screenshot shows the Google Dataset Search interface. At the top left is the "Google Dataset Search" logo. In the center is a search bar with the query "Mannheim". To the right of the search bar is a close button (an "X"). Below the search bar, a message says "100+ results found". The results are listed in a grid format:

- Straßentypen in Mannheim**
mannheim.opendatasoft.com
Updated 10.10.2016
- Straßennamen in Mannheim**
mannheim.opendatasoft.com
Updated 16.11.2016
- Entwicklung der Einwohnerzahl in Mannheim bis 2017**
de.statista.com

On the right side of the screenshot, there is a detailed view of the first result, "Bevölkerungsbestand in Mannheim 2013-2018". It includes a blue "Explore at mannheim.opendatasoft.com" button. Below it, the dataset was last updated on 15.07.2019. It also lists the license as "dl-de-by-2.0" and available download formats as "excel, csv, json". The "Description" section contains a long block of text about population statistics in Mannheim.

Crawls dataset metadata from the Web.

schema.org

2. Web 2.0 Applications and Web APIs

- A multitude of **Web-based applications (platforms)** enable users to share information.
- These applications form separate data spaces that might be **partly accessible** from the Web via
 - HTML interfaces
 - Web APIs



Example: Size of Facebook Social Graph

- Users (September 2018)
 - 2.3 billion monthly active users
 - including 1 billion mobile users
- 740 billion friend connections
- 4 million likes every minute
- 250 billion photos uploaded
- Data Volume
 - 4 Petabyte of new data generated every day
 - over 300 Petabyte in Facebook's data warehouse

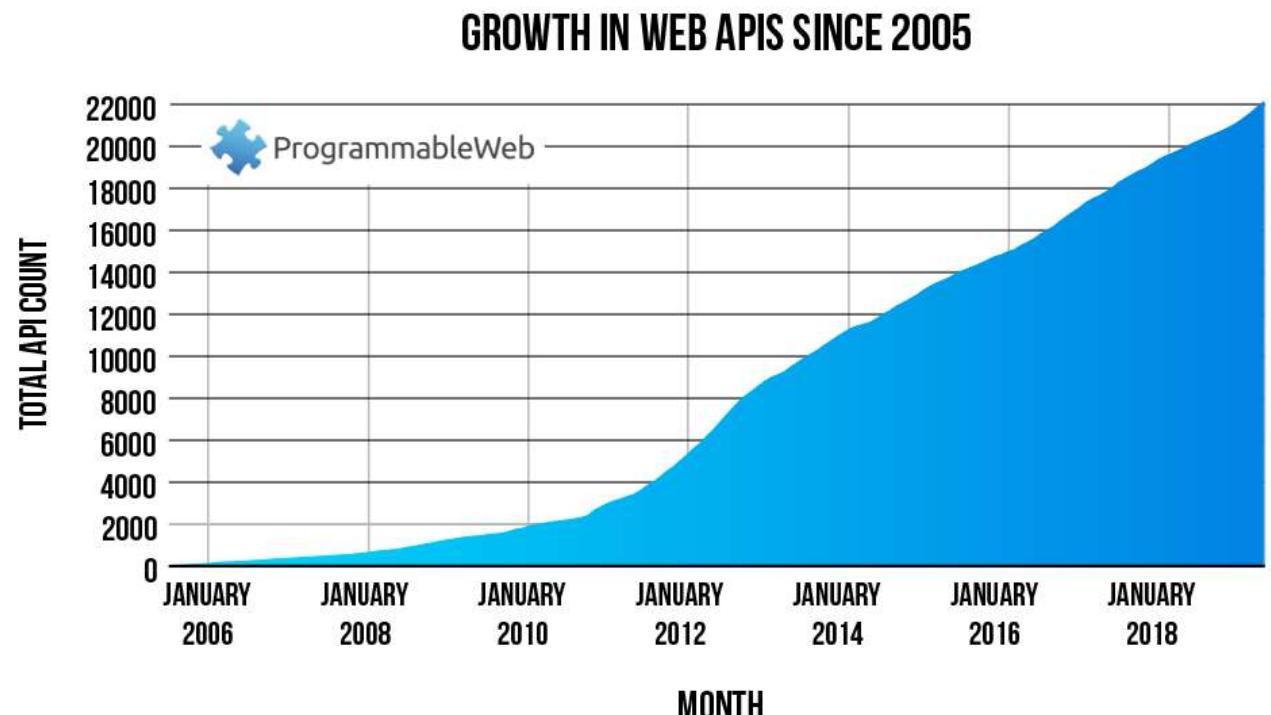


<https://www.brandwatch.com/blog/facebook-statistics/>

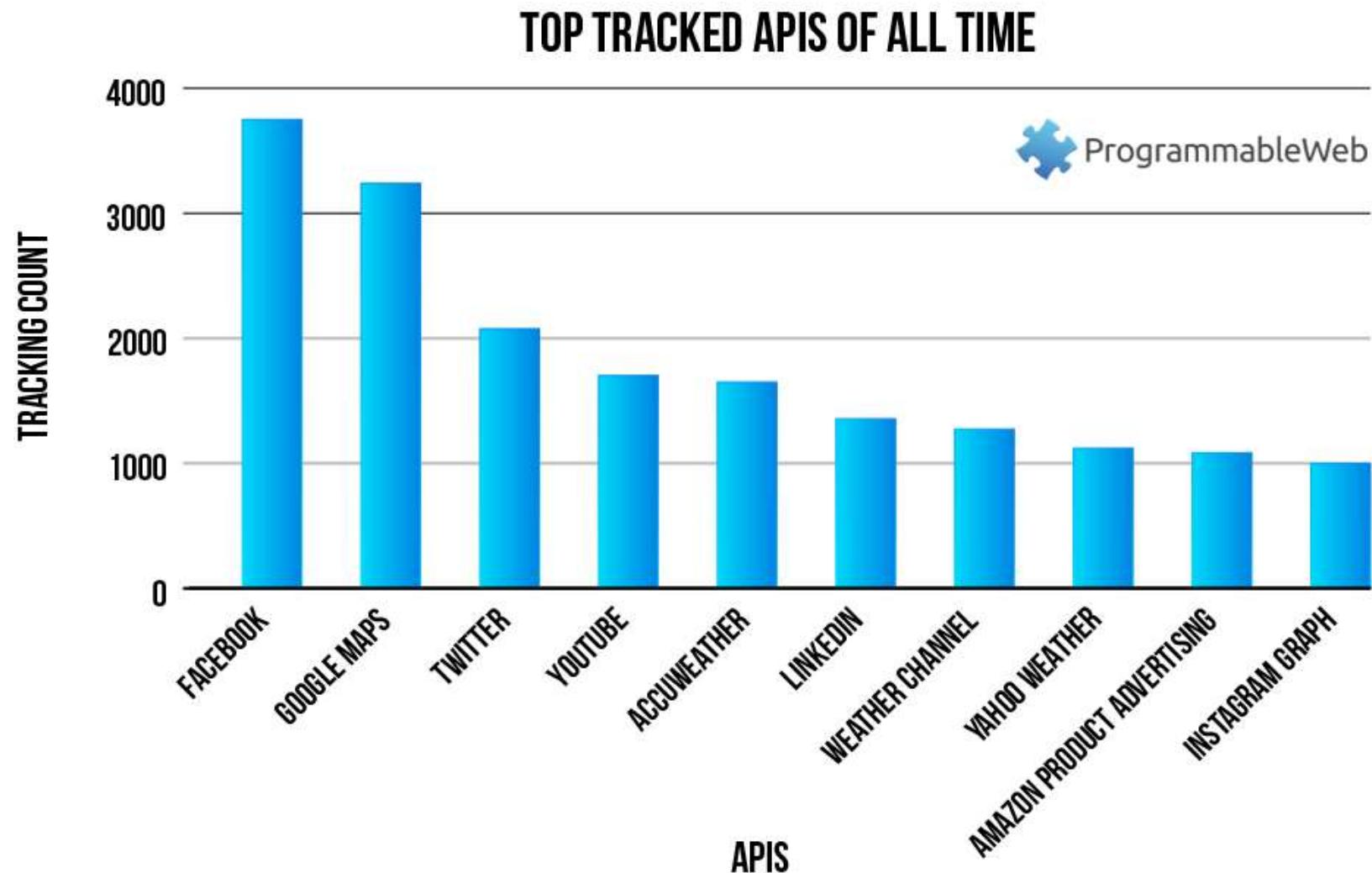
<http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/>

Web APIs

- Provide limited access to the collected data
 - restricted to specific queries (canned queries)
 - restricted by number of queries / number of results
- ProgrammableWeb API Catalog
 - lists over 24,000 Web APIs
 - lists over 6,800 mashups

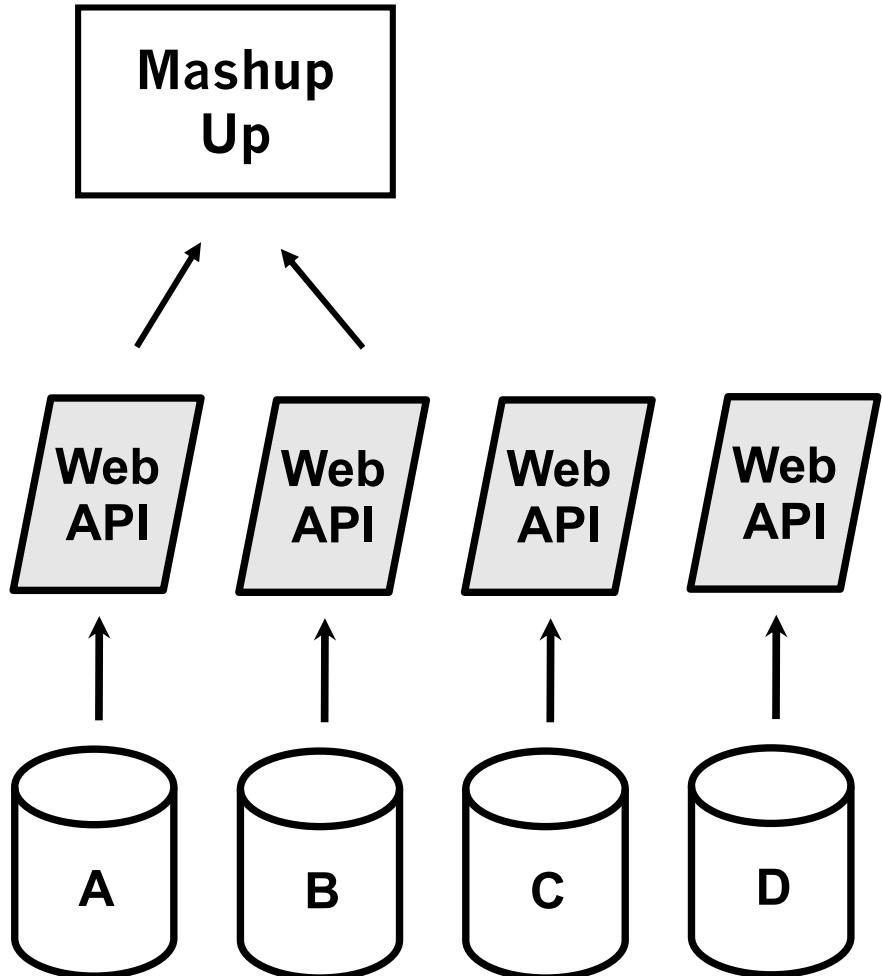


Most Popular APIs



<https://www.programmableweb.com/news/which-are-developers-favorite-apis/research/2019/10/24>

Mashups are based on a fixed set of data sources

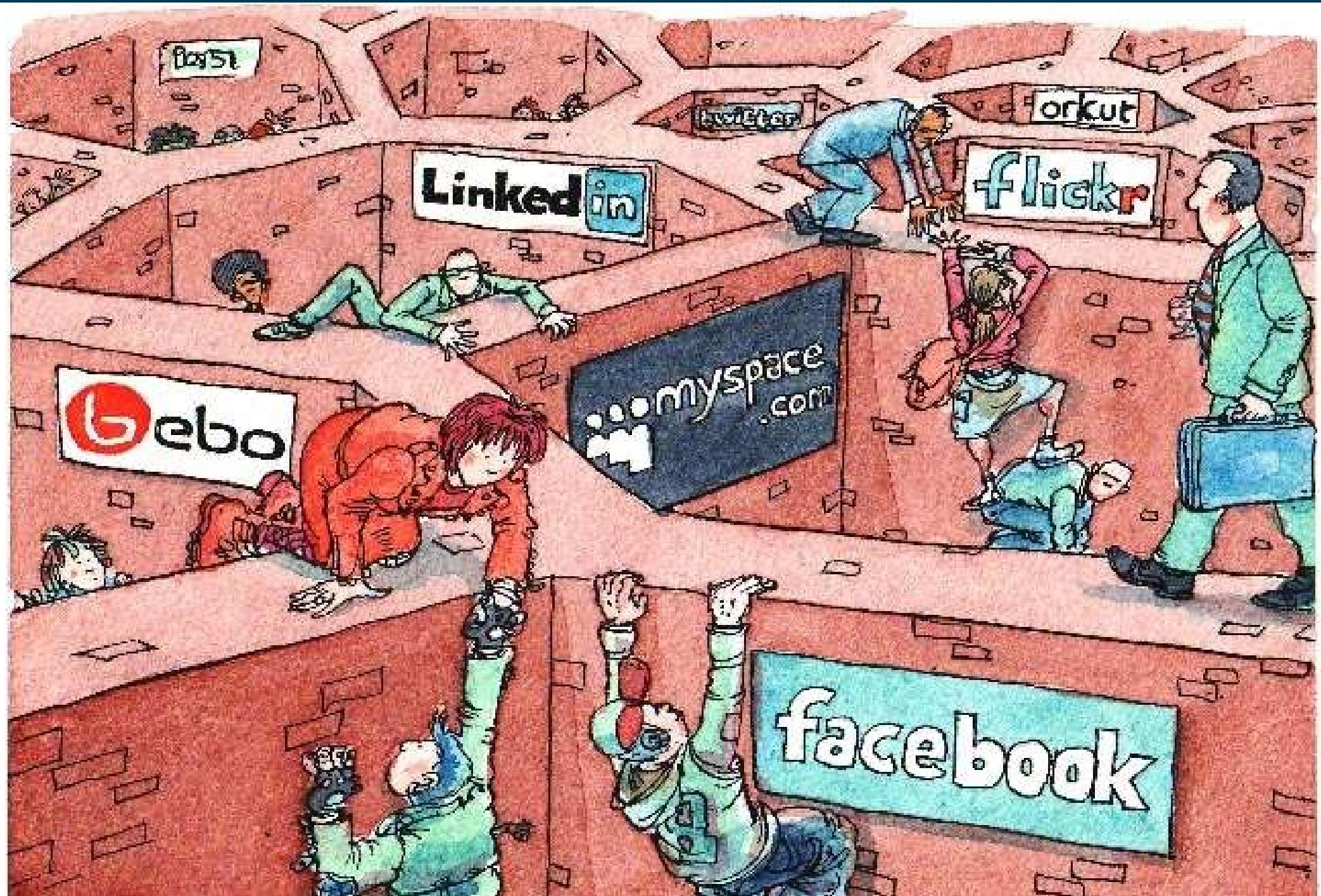


Web APIs expose proprietary interfaces

- Not indexable by generic web crawlers
- No automatic discovery of additional data sources
- No single global data space



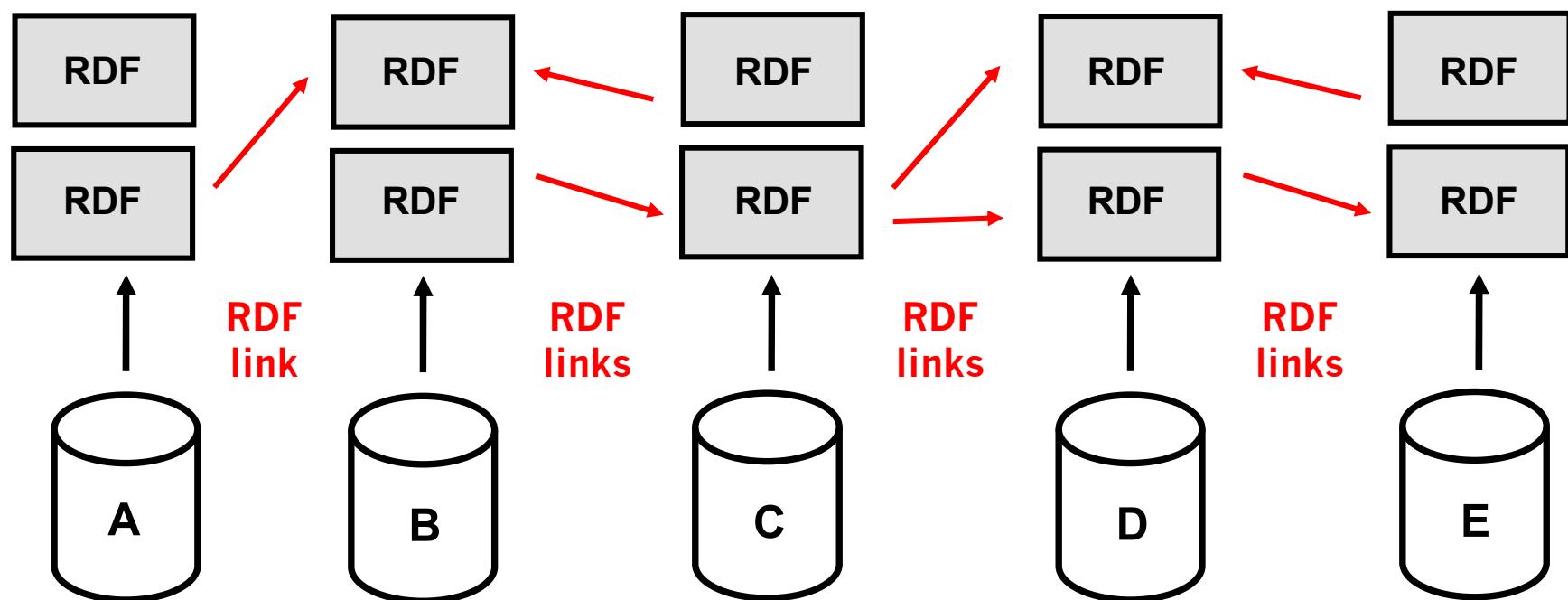
Web APIs slice the Web into Data Silos



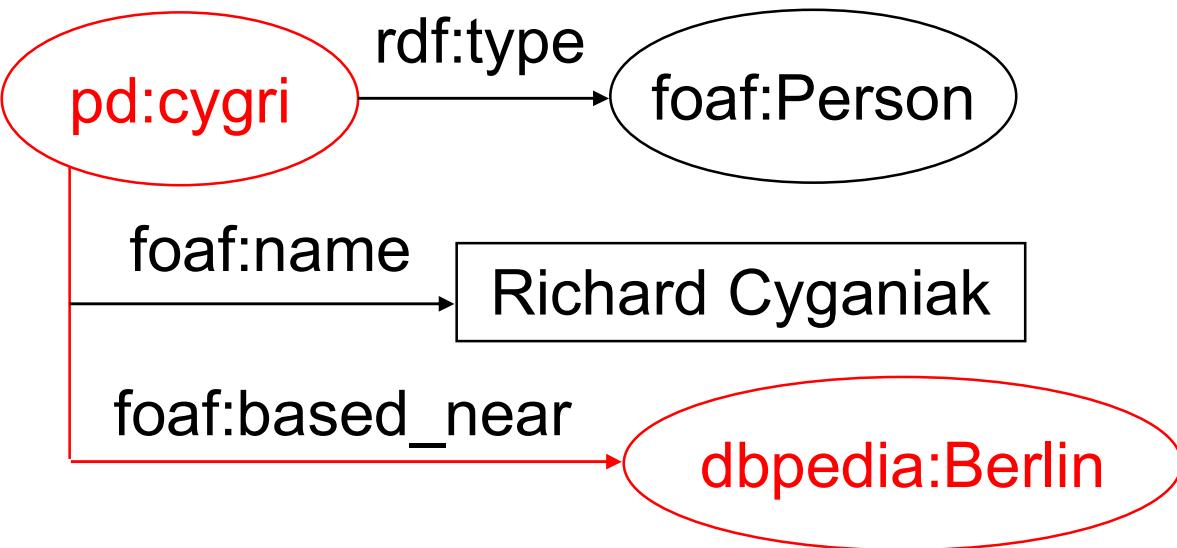
3. Alternative Approach: Linked Data



- Extend the Web with a single global data graph
 - by using RDF to publish structured data on the Web
 - by setting links between data items within different data sources



Entities are identified with HTTP URIs

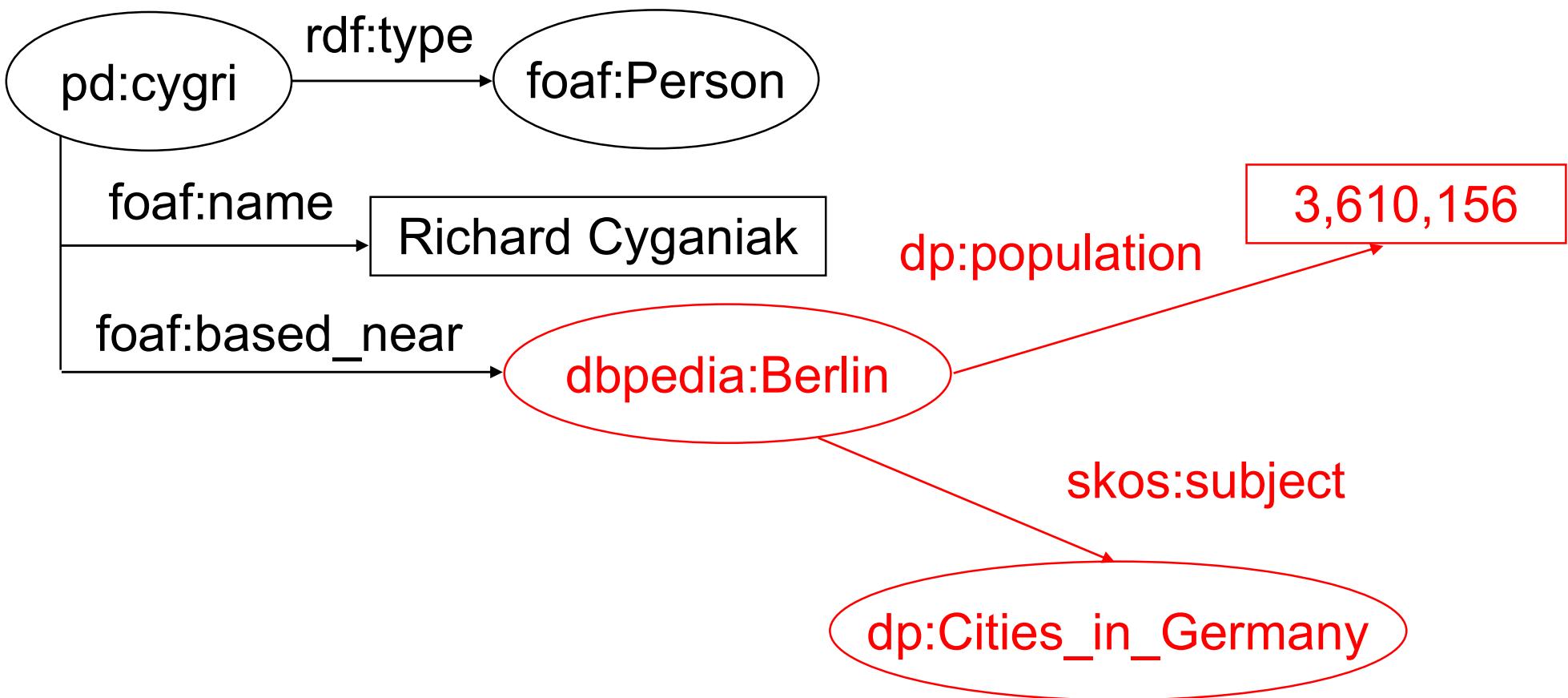


HTTP URIs take the role of global primary keys.

pd:cygri = <http://richard.cyganiak.de/foaf.rdf#cygri>

dbpedia:Berlin = <http://dbpedia.org/resource/Berlin>

URIs can be looked up on the Web



- By following RDF links applications can
 - navigate the global data graph
 - discover new data sources
- Linked Data is a specific technical realization of the FAIR principles (F1, A1, I1, I2, I3)

The Marbles Hyperdata Browser

<http://www.w3.org/People/Berners-Lee/card#i>

Open



Tim Berners-Lee

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

- Person
- <http://www.w3.org/2000/10/swap/pim/contact#Male>

label

- Tim Berners-Lee

sameAs

- Tim Berners-Lee (also at www4.wiwiiss.fu-berlin.de)

image



Weblinks

<http://www.w3.org/People/Berners-Lee/>

name

- Tim Berners-Lee
- Timothy Berners-Lee
- Tim Berners Lee

Given name

- Timothy

family name

- Berners-Lee

sha1sum of a personal mailbox URI name

• 985c47c5a70db7407210cef8e4e8f5374a525c5c

workplace homepage

- <http://www.w3.org/>

nickname

- TimBL

nickname

- TimBL

personal mailbox

- <mailto:timbl@w3.org>

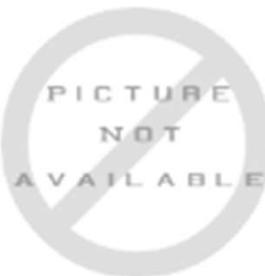
The SigMa Linked Data Search Engine

[Help](#)[About](#)[Forum](#)

Chris Bizer

[Add More Info](#)[Start New](#)[Options](#)[Order](#)[Permalink](#)

picture:

PICTURE
NOT
AVAILABLE

[5]

[16]

given name: Chris [3,5,9,10,16]

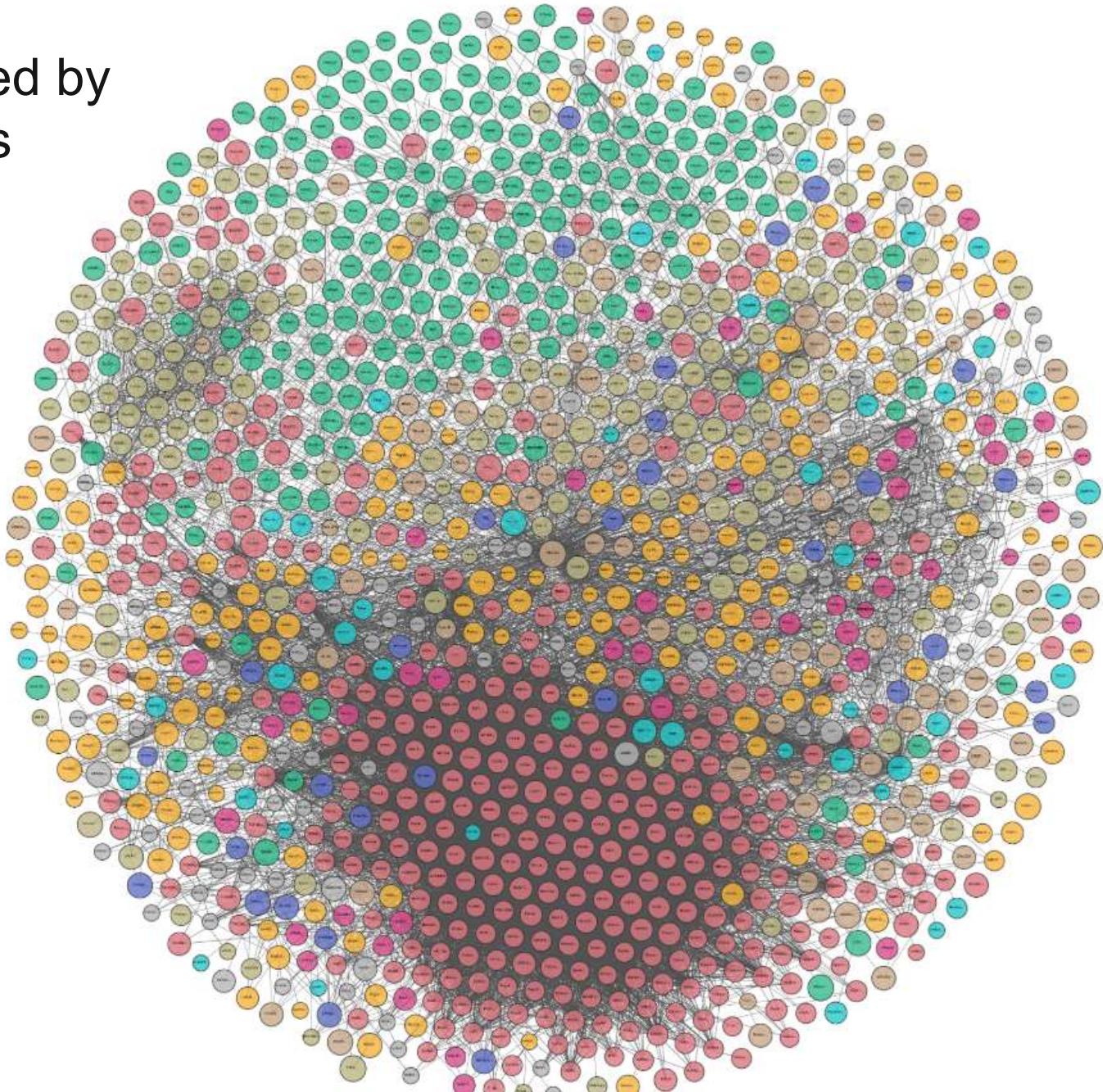
family name: Bizer [3,5,9,10,16]

is creator of: [DBpedia: A Nucleus for a Web of Open Data | Semantic Web Dog Food](#) [6,18]<http://data.semanticweb.org/conference/eswc/2007/demo-3> [9][The TriQL.P Browser: Filtering Information using Context-, Content- and Rating-Based Trust Policies.](#) [16][D2R Server - Publishing Relational Databases on the Semantic Web.](#) [16][Named Graphs, Provenance and Trust](#) [16][hide value](#)[just this value](#)[which sources](#)[reject sources](#)Sources (20) Approve1 [Chris Bizer - Free University Berlin](#)
http://videolectures.net/chris_bizer/2 [Chris Bizer - semanticweb.org](#)
http://ontoworld.org/wiki/Chris_Bizer3 [Untitled document](#) 6 facts
BOSS <http://www.facebook.com/ChrisBizer>4 [Chris Bizer - semanticweb.org](#)
http://semanticweb.org/wiki/Chris_Bizer5 [Chris Bizer - LinkedIn](#)
BOSS <http://www.linkedin.com/in/chrisbizer>6 [Chris Bizer 10 facts | 20 triples](#)
<http://data.semanticweb.org/people/chrisbizer>7 [Chris Bizer - semanticweb.org](#)
http://semanticweb.org/index.php?title=Chris_Bizer8 [Flickr: Chris Bizer's Photo Stream](#)
BOSS <http://flickr.com/photos/chrisbizer/>9 [Untitled document](#) 8 facts
<http://data.semanticweb.org/people/chrisbizer>10 [Chris Bizer 6 facts | 20 triples](#)
BOSS <http://ebiquity.umbc.edu/~chrisbizer/>

The Linked Open Data Cloud

1,255 datasets connected by
16,174 sets of RDF links
(as of May 2020)

Legend

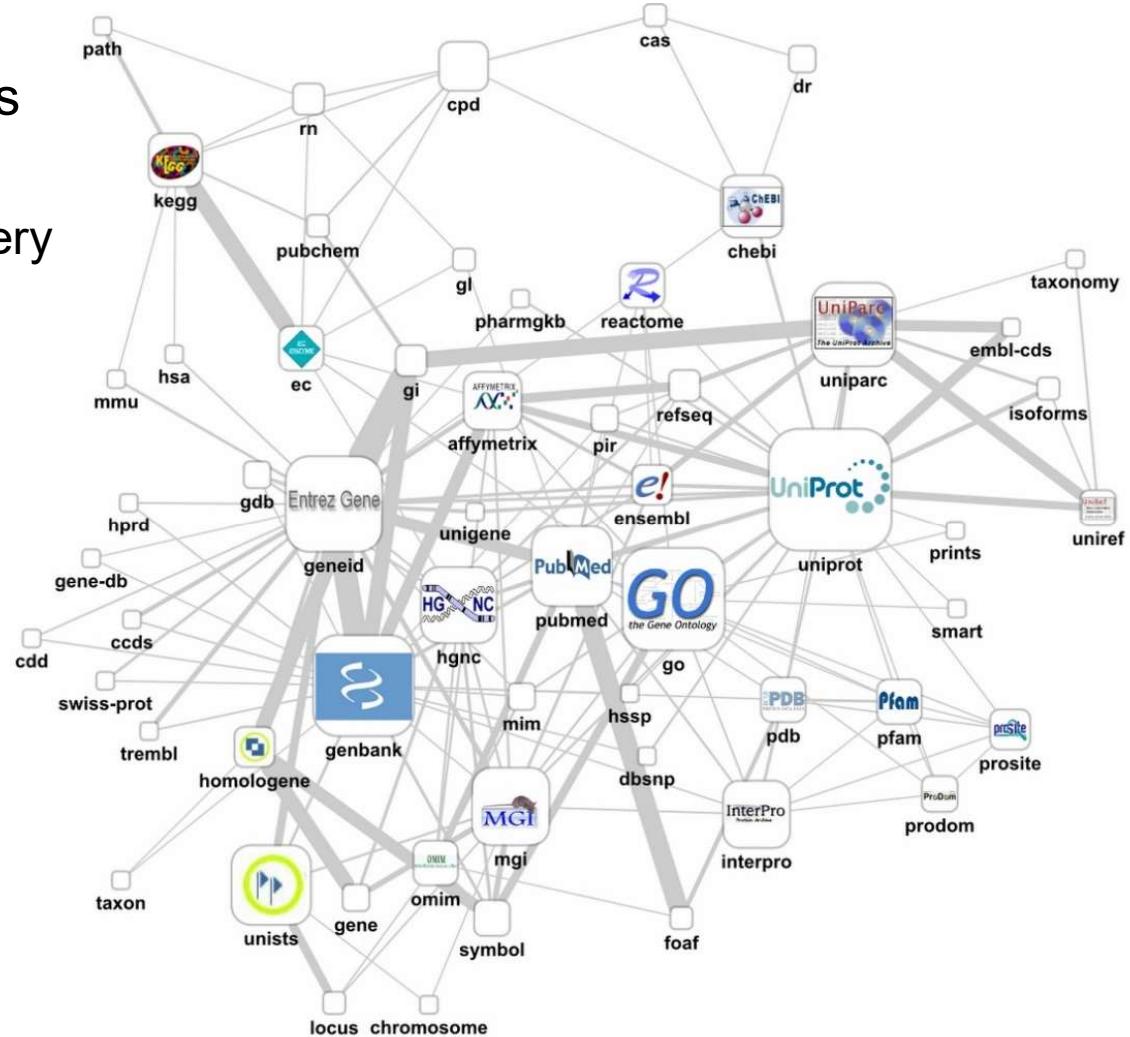


Uptake in the Life Science Domain

– Goals:

1. Connect life science datasets in order to support
 - biological knowledge discovery
 - drug discovery
2. Reuse results of previous integration efforts

BIO  **RDF**



Uptake in the Libraries Community

- Goals:
 1. interconnect resources between repositories
(by topic, by author, by location, by historical period, by ...)
 2. enable integration of library catalogs on global scale
- Institutions publishing Linked Data
 - Library of Congress (subject headings and catalog)
 - German National Library (PND dataset and subject headings)
 - Swedish National Library (Libris catalog)
 - Europeana Digital Library (catalog)
 - TIB Hannover (Open Research Knowledge Graph)
 - Springer Nature (publications, researchers, projects)



LIBRARY OF
CONGRESS



Jinfang Niu: **Diffusion and adoption of Linked Data among libraries**. In:
Information Science and Technology, 2020.

Hands-on: How to get the Data?

- Download the LOD-a-lot Dataset
 - 28 billion RDF triples (524 GB zipped)
 - crawled from the public Web of Linked Data in 2017
 - <http://lod-a-lot.lod.labs.vu.nl/>
- Download the Billion Triples Challenge Dataset
 - 4 billion RDF triples (52 GB gzipped, 1.1 TB uncompressed)
 - crawled from the public Web of Linked Data in 2014
 - <http://km.aifb.kit.edu/projects/btc-2014/>
- Use SPARQL endpoints of individual data sets
 - Endpoint list: <http://sparqles.ai.wu.ac.at/availability>

4. HTML-embedded Data

1. Webpages traditionally contain structured data in the form of **HTML tables** as well as **template data**
2. More and more websites semantically markup the content of their HTML pages using **standardized markup formats**

Microformats



Microdata



RDFa



JSON-LD



4.1 Microformats



- Microformat effort dates back to 2003
- Small set of fixed formats
 - hcard : people, companies, organizations, and places
 - XFN : relationships between people
 - hCalendar : calendaring and events
 - hListing : small-ads; classifieds
 - hReview : reviews of products, businesses, events
- Shortcoming of Microformats
 - can not represent any kind of data.
- indexed by Google and Yahoo since 2009



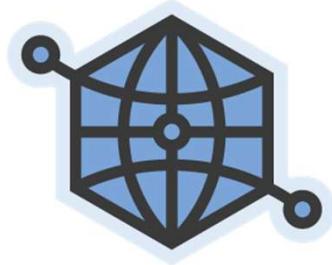
- serialization format for embedding RDF data into HTML pages
- W3C Recommendation in 2008
- can be used together with any vocabulary
- can assign URIs as global primary keys to entities

```
1 <html xmlns="http://www.w3.org/1999/xhtml"
2   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3   xmlns:foaf="http://xmlns.com/foaf/0.1/">
4 ...
5   <div about="http://example.com/Peter" typeof="foaf:Person">
6     <span property="foaf:name">Peter Smith</span> knows
7     <a rel="foaf:knows" href="http://example.com/Paula">Paula
8       Jones</a>.
9   </div>
10 ...
```

Open Graph Protocol



- allows site owners to determine how entities are described in Facebook
- relies on **RDFa** for embedding data into HTML pages
- available since April 2010



A screenshot of a Facebook news feed. It shows a post from the official "The Rock" Facebook page, which links to a specific IMDb movie page for "The Rock". The post indicates that "Francis Luu likes The Rock on IMDb." and was made "5 minutes ago".

Microdata



- alternative technique for embedding structured data
- proposed in 2009 by WHATWG as part of HTML5 work
- tries to be simpler than RDFa (5 new attributes instead of 8)

```
<div itemtype="http://schema.org/Hotel">  
  <span itemprop="name">Vienna Marriott Hotel</span>  
  <span itemprop="address" itemscope="" itemtype="http://schema.org/PostalAddress">  
    <span itemprop="streetAddress">Parkring 12a</span>  
    <span itemprop="addressLocality">Vienna</span>  
  </span>  
  <div itemprop="aggregateRating" itemscope itemtype="http://schema.org/AggregateRating">  
    <span itemprop="ratingValue"> 4 </span> stars-based on  
    <span itemprop="reviewCount"> 250 </span> reviews.  
  </div>
```

JSON-LD

- used for embedding data into the HEAD of HTML pages
- putting data into the HEAD is recommended by Google as it is empirically less error prone than annotations in BODY



```
<script type="application/ld+json">
{  "@context": "http://schema.org",
  "@type": "Product",
  "description": "Has six preset cooking ....",
  "name": "Kenmore White 17\" Microwave",
  "offers": {
    "@type": "Offer",
    "availability": "http://schema.org/InStock",
    "price": "55.00",
    "priceCurrency": "USD"
  },
}
</script>
```



- ask site owners since 2011 to annotate data for enriching search results
- 675 Types: Event, Local Business, Product, Review, Job Offer
- Encoding: Microdata, RDFa, JSON-LD

schema.org Search

Home Schemas Documentation

Thing > Organization > LocalBusiness

A particular physical business or branch of an organization. Examples of LocalBusiness include a restaurant, a particular branch of a restaurant chain, a branch of a bank, a medical practice, a club, a bowling alley, etc.

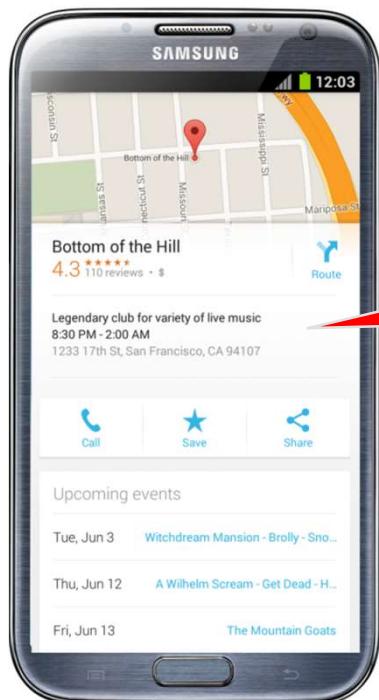
Property	Expected Type	Description
Properties from Thing		
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
url	URL	URL of the item.
Properties from Place		
address	PostalAddress	Physical address of the item.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
containedIn	Place	The basic containment relation between places.



Usage of Schema.org Data @ Google

Gramercy Tavern - Flatiron - New York, NY | Yelp
www.yelp.com › Restaurants › American (New) ▾
★★★★★ Rating: 4.5 - 1,288 reviews - Price range: \$\$\$
Jeff C and I were in New York for vacation, and I wanted to treat him to a nice dinner for Gramercy Tavern is certainly a legendary NY dining establishment.

Gramercy Tavern Restaurant - New York, NY | OpenTable
www.opentable.com › ... › Gramercy restaurants ▾
★★★★★ Rating: 4.7 - 508 reviews - Price range: \$50 and over
Book now at Gramercy Tavern in New York, explore menu, see photos and read 508 reviews: "The menu was so limited but it was worth trying, food was deli..."



Data snippets
within
search results

Local businesses
on maps

Data snippets
within
info boxes



The Black Keys

Band

The Black Keys is an American rock duo formed in Akron, Ohio in 2001. The group consists of Dan Auerbach and Patrick Carney. [Wikipedia](#)

Origin: Akron, Ohio, United States

Members: Dan Auerbach, Patrick Carney

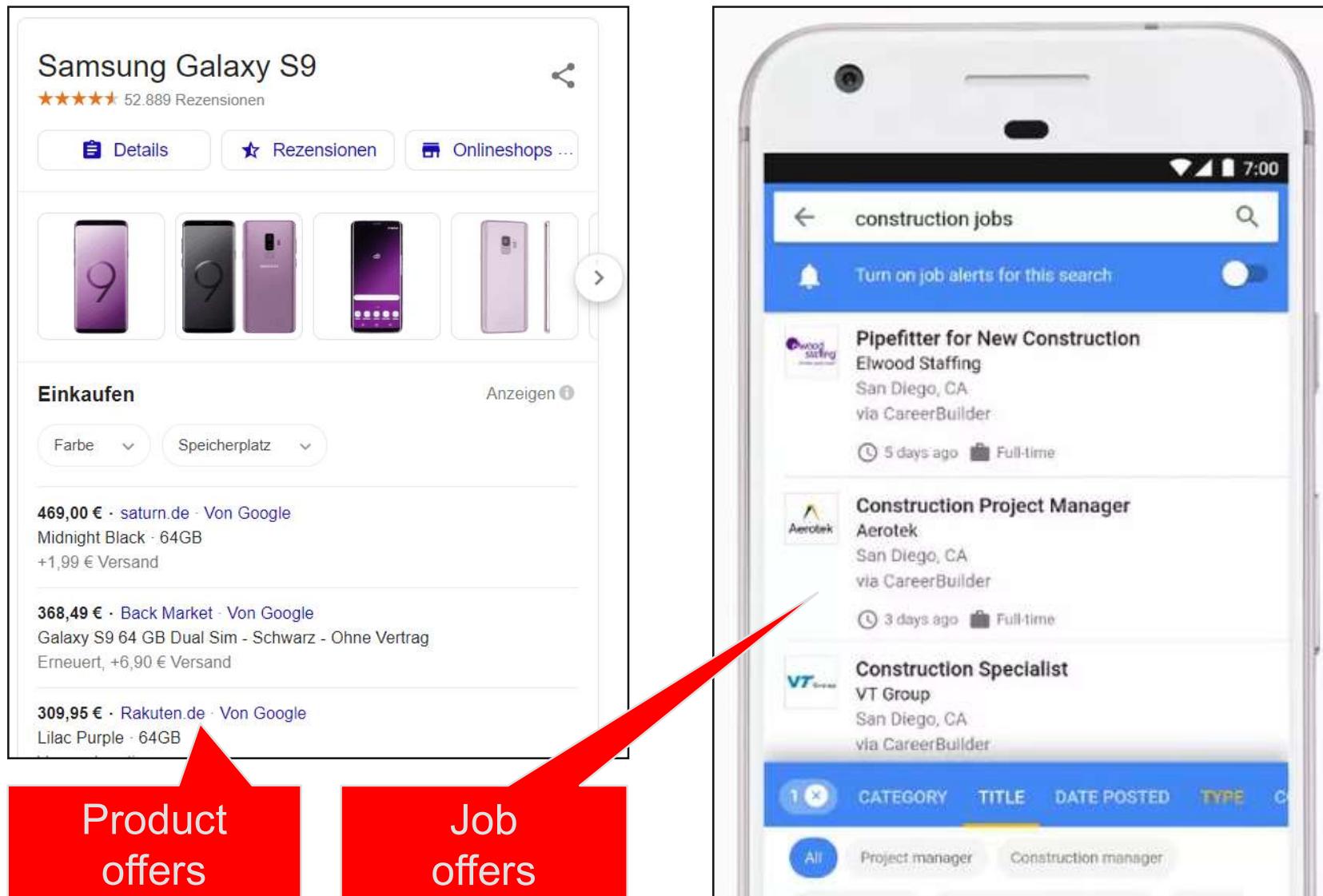
Record labels: Fat Possum Records, Nonesuch Records, V2 Records, Alive Natural sound Records

Awards: Grammy Award for Best Rock Album, more

Upcoming events

Jun 20	The Black Keys
Fri	Neuhausen ob Eck (near you)
May 16	The Black Keys
Fri	Gulf Shores, AL
Jun 22	The Black Keys
Sun	Scheeßel

Usage of Schema.org Data @ Google



<https://developers.google.com/search/docs/guides/search-gallery>

The Web Data Commons Project

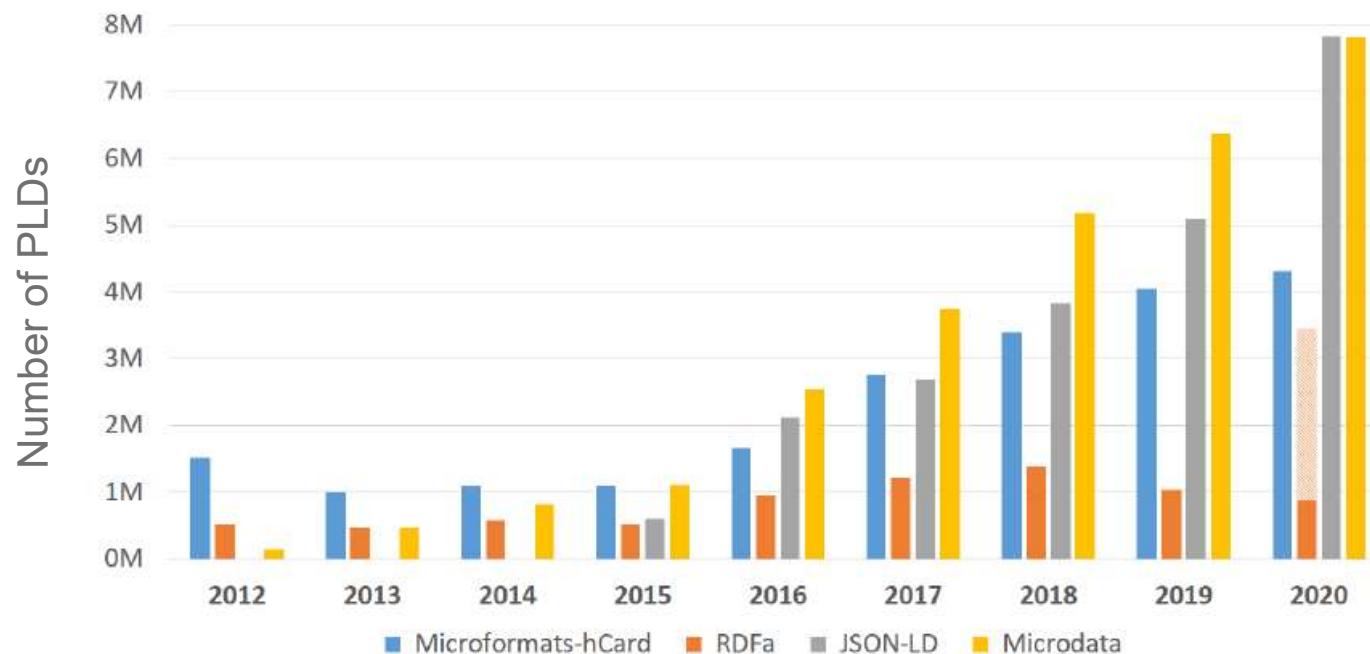
- extracts all Microformat, Microdata, RDFa, JSON-LD data from the **Common Crawl**
- analyzes and provides the extracted data for download
- statistics about some extraction runs
 - 2020 CC Corpus: 3.4 billion HTML pages → **86.3** billion RDF triples
 - 2017 CC Corpus: 3.1 billion HTML pages → 38.2 billion RDF triples
 - 2013 CC Corpus: 2.2 billion HTML pages → 17.2 billion RDF triples
 - 2010 CC Corpus: 2.8 billion HTML pages → **5.1** billion RDF triples
- uses 100 machines on **Amazon EC2**
 - approx. 2000 machine/hours → 500 Euro
- <http://webdatacommons.org/structureddata/>



Overall Adoption 2020

1.7 billion HTML pages out of the 3.4 billion pages provide semantic annotations (50.0%).

15.3 million pay-level-domains (PLDs) out of the 34.5 million PLDs (websites) provide semantic annotations (44.3%).



<http://webdatacommons.org/structureddata/2020-12/stats/stats.html>

Frequently used Schema.org Classes (2020)

Class	# Websites (PLDs)	
	JSON-LD	Microdata
schema:WebPage	4,484,026	1,339,999
schema:Person	3,151,809	514,990
schema:BreadcrumbList	1,688,820	924,991
schema:Article	1,327,578	627,303
schema:Product	1,234,972	1,059,149
schema:Offer	1,182,855	946,725
schema:PostalAddress	863,243	585,417
schema:BlogPosting	529,020	552,338
schema:LocalBusiness	363,843	280,338
schema:AggregateRating	432,014	315,253
schema:Place	255,139	93,124
schema:Event	194,115	77,722
schema:Review	181,097	158,333
schema:JobPosting	28,759	8,520

http://webdatacommons.org/structureddata/2020-12/stats/schema_org_subsets.html

Adoption by Travel Websites

Top 15 Travel Websites	schema:Hotel
Booking.com	✓
TripAdvisor	✓
Expedia	✓
Agoda	✓
Hotels.com	✓
Kayak	✓
Priceline	✓
Travelocity	✓
Orbitz	✓
ChoiceHotels	✓
HolidayCheck	✓
ChoiceHotels	✓
InterContinental Hotels Group	✓
Marriott International	✓
Global Hyatt Corp.	✗

Adoption:
93 %

Properties used to Describe Products (2020)

Attribute	% of PLDs
schema:Product/name	99 %
schema:Product/offers	94 %
schema:Offer/price	95 %
schema:Offer/priceCurrency	95 %
schema:Product/description	84 %
schema:Offer/availability	72 %
schema:Product/sku	56 %
schema:Product/brand	30 %
schema:Product/image	26 %
schema:Product/aggregateRating	17 %
schema:Product/mpn	6.3 %
schema:Product/productID	4.7 %
...	...



<http://webdatacommons.org/structureddata/schemaorgtables/>

Hands-on: How to get the Data?

- as RDF quads: <http://webdatacommons.org/structureddata/>
- as JSON for pandas: <http://webdatacommons.org/structureddata/schemaorgtables/>

Class-Specific Subsets of the Schema.org Data

Class Name	Total Number of	Top Classes (Entity Count)	Total File Size	Quad File
http://schema.org/AdministrativeArea	Quads: 1,724,857 URLs: 85,625 Hosts: 63	http://schema.org/AdministrativeArea (100,671) http://schema.org/GeoCoordinates (84,152) http://schema.org/Country (83,851) http://schema.org/Continent (83,567)	23 MB	schemaorgAdministrativeArea.nq.gz (sample)
http://schema.org/Hotel	Quads: 148,211,253 URLs: 3,136,152 Hosts: 5,337	http://schema.org/Rating (7,007,590) http://schema.org/Hotel (6,335,124) http://schema.org/Review (4,408,551) http://schema.org/AggregateRating (3,936,372)	2,994 MB	schemaorgHotel.nq.gz (sample)
http://schema.org/JobPosting	Quads: 234,475,135 URLs: 2,011,332 Hosts: 3,962	http://schema.org/JobPosting (22,804,279) http://schema.org/Place (16,321,339) http://schema.org/Organization (12,164,867) http://schema.org/Postaladdress (7,516,387)	5,078 MB	schemaorgJobPosting.nq.gz (sample)
http://schema.org/PostalAddress	Quads: 776,573,609 URLs: 13,475,055 Hosts: 131,064	http://schema.org/PostalAddress (48,086,763) http://schema.org/LocalBusiness (16,641,260) http://schema.org/GeoCoordinates (12,345,942) http://schema.org/Place (9,071,774)	14,364 MB	schemaorgPostalAddress.nq.gz (sample)
http://schema.org/Product	Quads: 2,829,523,589 URLs: 48,314,143 Hosts: 104,118	http://schema.org/Product (287,815,069) http://schema.org/Offer (221,781,710) http://schema.org/AggregateRating (38,398,548) http://schema.org/Review (26,209,678)	62,179 MB	schemaorgProduct.nq.gz (sample)

- Only tip of the iceberg, as each website is only partly crawled.

4.2 HTML Tables

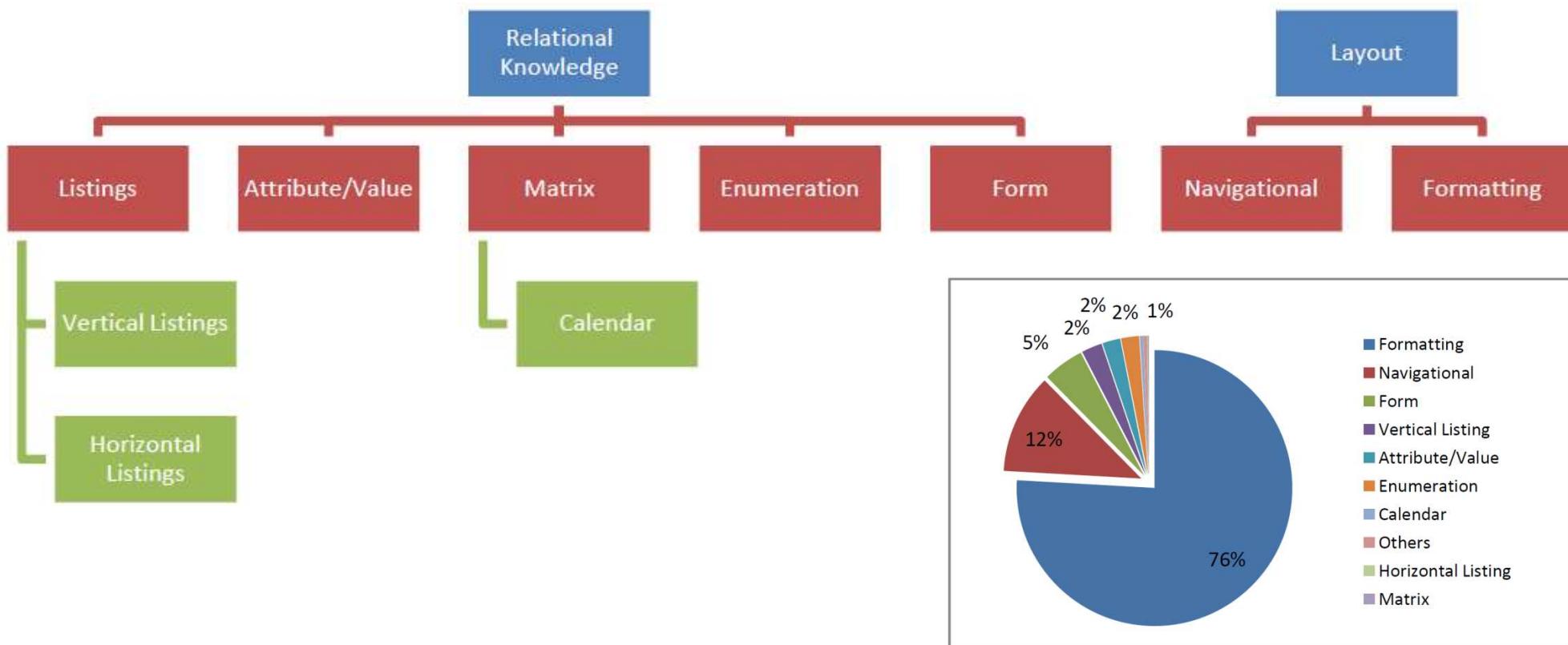
There are hundreds of millions of high-quality HTML tables on the Web and in Wikipedia.

Germany - Largest Cities				150 INTERNATIONAL AFFAIRS						Most Requested Songs						
Name	Population	Latitude/Longitude		WEBSITE		HUB		AIRLINE CODE		2016	2017	2018	2019	2020	Published December 28, 2011	Here are the most requested songs of the past year:
				www.aeroflot.ru/eng/	www.airfrance.fr	Moscow	SU	Paris	AF							
1 Berlin	3,426,354	52.524 / 13.411														
2 Hamburg	1,739,117	53.575 / 10.015														
3 Munich			150 INTERNATIONAL AFFAIRS			2016	2017	2018	2019	2020						
4 Cologne			8. End Funding for the United Nations Development Program (UNDP)			81	81	82	83	85						
5 Frankfurt am Main			9. End Funding for the U.N. Intergovernmental Panel on Climate Change (IPCC)			10	10	10	10	11						
6 Essen			10. Eliminate the U.S. Trade and Development Agency (USTDA)													
7 Stuttgart			11. Reform Food Aid Programs													
8 Dortmund			12. Eliminate Export-Import Bank													
9 Düsseldorf			13. Eliminate the Overseas Private Investment Corporation (OPIC)													
10 Bremen			14. Eliminate Funding for the United Nations Population Fund (UNFPA)													
Contestant		Age	SUBTOTALS													
Kelly Louise Maguire	24	1.70 m (5 ft 7 in)	MASSAU													
Aquelle Plakaris	24	1.70 m (5 ft 7 in)	EVERGEM													
Jessica van Moorlegh	18	1.70 m (5 ft 7 in)	SANTA CRUZ													
Yovana O'Brien	19	1.80 m (5 ft 11 in)														

Types of Web Tables

In corpus of 14B raw tables, 154M are “good” relations (1.1%).

Cafarella (2008)



Cafarella, et al.: **WebTables: Exploring the Power of Tables on the Web.** VLDB 2008.
Crestan, Pantel: **Web-Scale Table Census and Classification.** WSDM 2011.

Attribute/Value Table together with schema.org Annotations

s:breadcrumb
s:name

Attribute/Value
HTML Table

The screenshot shows a Walmart product page for an HP Ultrabook EliteBook Folio G1 12.5" Laptop. The page includes a breadcrumb trail, a product title, a star rating, a specifications table, and a product image.

Breadcrumb: Electronics > Computers > Laptops > Shop Laptops by Type > All Laptop Computers

Title: HP Ultrabook EliteBook Folio G1 12.5" Laptop, Windows 10 Pro, Intel Core m5-6Y54 Processor, 8GB RAM, 256GB Solid State Drive

Price: \$1,060⁷⁹ + FREE shipping
Only 2 left!

Sold & Shipped by: BidDeal
Shipping: FREE Standard shipping | Shipping options
Pickup: Pickup not available from this seller

Qty: 1 **Add to Cart**

Specifications:

Aspect Ratio:	16:9
Graphics Information:	HD Graphics 515
Processor Type:	Intel Core M5-6Y54 Dual-Core Processor
Hard Drive Capacity:	256 GB
Color:	Silver
Display Technology:	LED Backlight, Full HD Display
Resolution:	1080p
Form Factor:	Laptop
Processor Speed:	1.10 GHz
Color Category:	Silver
Contained Battery Type:	Lithium Ion
Maximum RAM Supported:	8 GB

Image: An image of the laptop showing the HP logo on the screen.

Qui, et al.: **DEXTER: Large-Scale Discovery and Extraction of Product Specifications on the Web.** VLDB 2015.
Petrovski, et al: **The WDC Gold Standards for Product Feature Extraction and Product Matching.** ECWeb 2016.

Hands-on: Web Data Commons – Web Tables Corpus

- Large public corpus of relational Web tables
- extracted from Common Crawl 2015 (1.78 billion pages)
- 90 million relational tables
 - selected out of 10.2 B raw tables (0.9%)
 - download includes the HTML pages of the tables (1TB zipped)
 - <http://webdatacommons.org/webtables/>

Common Crawl



Web Data Commons – Web Tables Corpus

■ Attribute Statistics

Attribute	#Tables
name	4,600,000
price	3,700,000
date	2,700,000
artist	2,100,000
location	1,200,000
year	1,000,000
manufacturer	375,000
country	340,000
isbn	99,000
area	95,000
population	86,000

28,000,000 different attribute labels

■ Subject Attribute Values

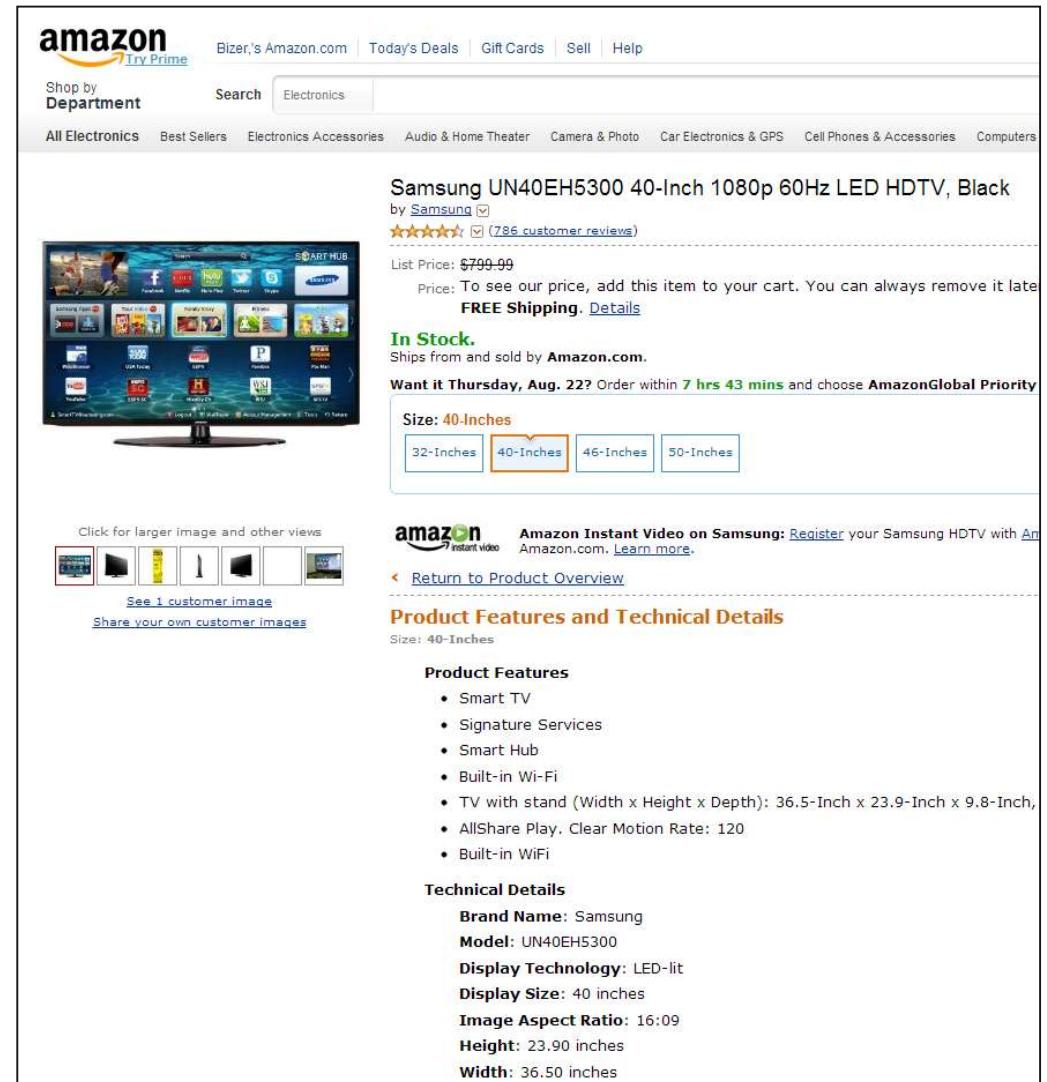
Value	#Rows
usa	135,000
germany	91,000
greece	42,000
new york	59,000
london	37,000
athens	11,000
david beckham	3,000
ronaldinho	1,200
oliver kahn	710
twist shout	2,000
yellow submarine	1,400

1.74 billion rows

253,000,000 different subject labels

Exploiting the Template-Structure of HTML Pages

- Most webpages are generated from databases using HTML-templates
- Approaches to extract the data:
 - hand-written wrappers using Xpath or regexes
 - wrapper induction using machine learning techniques
(see Bing Liu: Web Data Mining book)
- Problem:
 - wrappers are site-specific
 - thus, the approach does not scale to large numbers of websites
 - possible way out: Distant supervision in the form of knowledge bases



4.3 Wikipedia as Data Source

The screenshot shows the Bristol Wikipedia page. A red box highlights the title "Bristol". Another red box highlights the main content area, which includes a brief description, a map, and a coat of arms. A third red box highlights the "Boundaries" section at the bottom, which contains a map and a table of administrative details.

Bristol

From Wikipedia, the free encyclopedia

This article is about the English city. For other uses, see Bristol (disambiguation).

Bristol (pronunciation (help · info); IPA: /brɪstəl/) is a city, unitary authority and ceremonial county in South West England, 105 miles (169 km) west of London, and 44 miles (71 km) east of Cardiff.

With an approximate population of 410,950, and urban area of 550,200, it is England's sixth, and the United Kingdom's ninth most populous city, one of England's core cities and the most populous city in South West England. It received a royal charter in 1155 and was granted county status in 1373. For half a millennium it was the second or third largest English city, until the rapid rise of Liverpool, Birmingham and Manchester in the Industrial Revolution in the later part of the 18th century. It borders on the Counties of Somerset, and Gloucestershire, between the cities of Bath, Gloucester and Newport, and has a short coastline on the estuary of the River Severn, which flows into the Bristol Channel.

Bristol is one of the centres of culture, employment and education in the region. From its earliest days, its prosperity has been linked to that of the Port of Bristol, the commercial port, which was in the city centre but has now moved to the Severn estuary coast at Avonmouth and Portbury. In more recent years the economy has been built on the aerospace industry, and the city centre docks have been regenerated as a centre of heritage and culture.^[2]

Contents

- 1 Boundaries
- 2 History
- 3 Economy and industry
- 4 Culture
 - 4.1 Arts
 - 4.2 Sport and leisure
 - 4.3 Media
 - 4.4 Dialect
- 5 Politics and government
- 6 Demographics
- 7 Physical geography
- 8 Education, science and technology
- 9 Transport
- 10 Twin cities
- 11 See also
- 12 References
- 13 External links

Boundaries

There are a number of different ways in which Bristol's boundaries are defined, depending on whether the boundaries attempt to define the city, the built-up area, or the wider "Greater Bristol". The narrowest definition of the city is the city council boundary; although this definition does include a large portion of the Severn Estuary, west as far as the islands of Steep Holm and Flat Holm.^[3] A slightly less narrow definition is used by the Office for National Statistics; this includes built-up areas which adjoin Bristol but are not within the city council boundary, such as Whitchurch village, Filton, Patchway, Bradley Stoke, and excludes non-built-up areas within the city council boundary.^[4] The ONS has also defined an area which it calls the "Bristol Urban Area" which includes Kingswood, Mangotsfield, Stoke Gifford, Winterbourne, Frampton Cotterell, Almondsbury and Easton-in-Gordano.^[5] The term "Greater Bristol" (used for example by the Government Office of the South West)^[6] is most usually used to refer to the area covered by the city and its three neighbouring local authorities.

Coordinates: 51°28'N 2°35'W

Bristol

Bridge and the Avon Gorge

Coat of Arms of the City Council

Coordinates: 51°27'14"N 2°35'48"W

Sovereign state	United Kingdom
Constituent country	England
Region	South West England
Ceremonial county	Bristol
Historic counties	County corporate (Gloucestershire and Somerset)
Admin HQ	Avon
Royal Charter	Bristol
County status	1155
Government	1373
- Type	Unitary authority, City
- Governing body	Bristol City Council
- Leadership	Leader & Cabinet
- Executive	Labour
- MPs	Roger Berry (L) Kerry McCarthy (L) Doug Naysmith (L)/(Co-op) Dawn Primarolo (L) Stephen Williams (LD)



WIKIPEDIA
The Free Encyclopedia

Title

Description

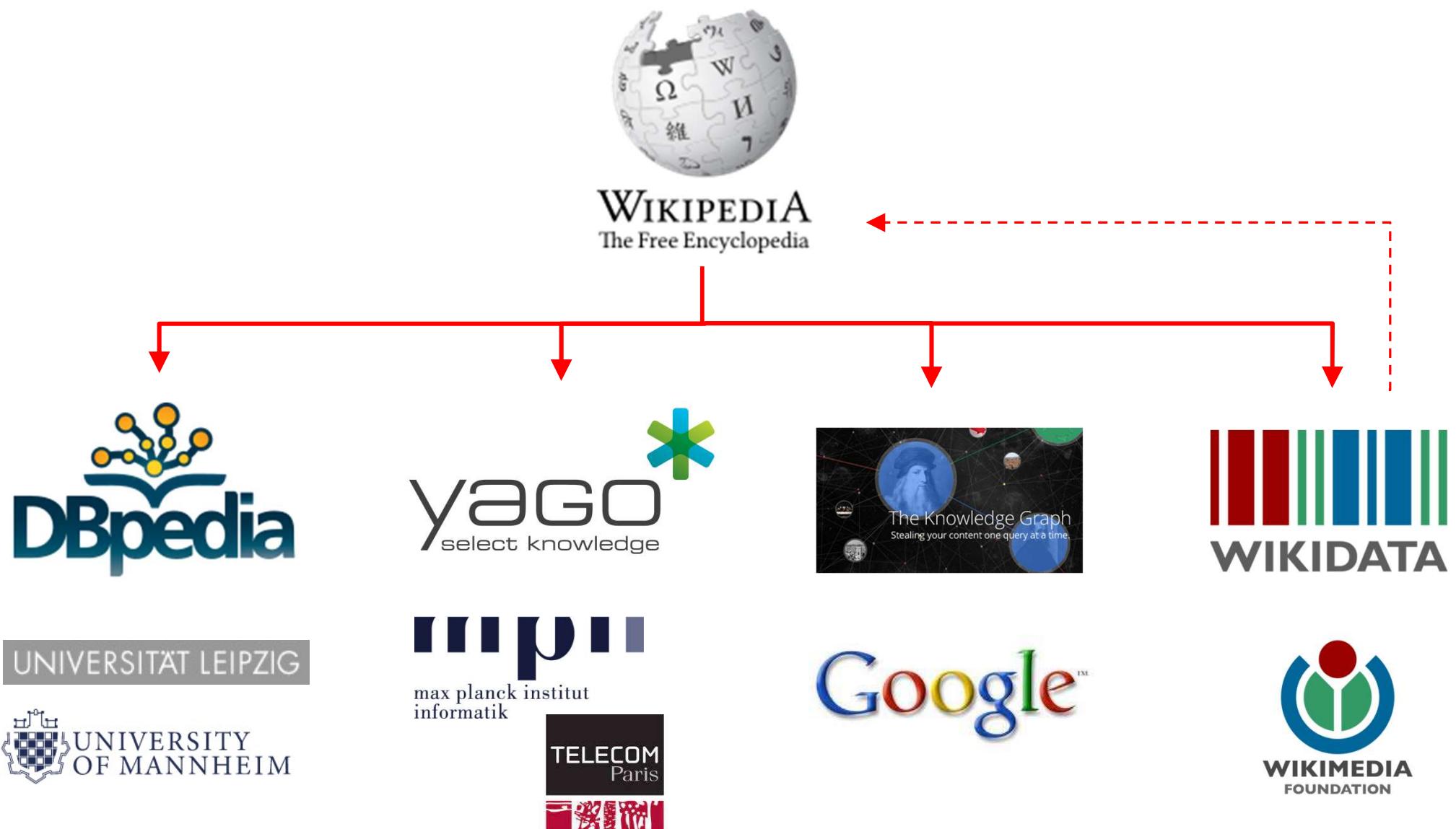
Cross
Language
Links

Geo-
Coordinates

Images

Infoboxes

Extracting Knowledge from Wikipedia



Ringler, Paulheim: **Analyzing the Differences Between DBpedia, YAGO, Wikidata**. KI 2017.

The DBpedia Knowledge Graph - Release 2022

- Describes **7.6 million things**, out of which 6.5 million are classified in a consistent ontology using 760 classes and 1377 different properties
 - 1,790,000 persons
 - 748,000 places
 - 345,000 organizations
 - 139,000 music albums
- Altogether 20 billion pieces of information (RDF triples)
 - 850 million were extracted from the English edition of Wikipedia
 - 29,000,000 links to external web pages
 - 139,000,000 external RDF links into 179 other RDF datasets
- DBpedia Internationalization
 - provides data from 125 Wikipedia language editions for download
 - for 28 popular languages DBpedia provides cleaned infobox data




[First](#) | [Previous](#) | [Next](#) | [Last](#)
▼ item type

[Skyscraper \(12\)](#)
[Place \(12\)](#)
[Building \(12\)](#)
[more](#)
▼ location

[Hong Kong \(12\)](#)
[China \(3\)](#)
[Sham Tseng \(1\)](#)
[more](#)
▼ building started in year

[2000 \(5\)](#)
[1977 \(1\)](#)
[1997 \(1\)](#)
[more](#)
▼ building completed in year
Your Filters
[Reset Filters](#)
Results 7 to 12 of 12
[item type Skyscraper](#)  [floor count 50 and up](#)  [building completed in year up to 2000](#)  [location Hong Kong](#) 


Highcliff

Highcliff is a 252.4-metre (828-foot) tall skyscraper located on a south slope of Happy Valley on the Hong Kong Island in Hong Kong. The 75 storey (70 floors of which are livable space) building's construction began in 2000 and was completed in 2003 under a design by DLN Architects & Engineers. It was the Silver Winner of the 2003 Emporis Skyscraper Award, coming in second to 30 St Mary Axe in London.



The Harbourside

The Harbourside is a 255 m (836.6 ft) tall residential skyscraper located at 1 Austin Road West, in Union Square complex on Kowloon peninsula. The building is erected on the West Kowloon Reclamation west of Kwun Chung. Construction of the 74 storey building began in 2000 and was completed in 2003 under the design by P & T Architects & Engineers. The building is, in fact, three towers joined at the base, middle



Hands-on: How to get DBpedia Data?

- Download Data Dumps
- Use SPARQL endpoint

SPARQL Explorer for http://dbpedia.org

SPARQL:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dbo: <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?name ?birth ?death ?person WHERE {
  ?person dbo:birthPlace :Berlin .
  ?person dbo:birthDate ?birth .
  ?person foaf:name ?name .
  ?person dbo:deathDate ?death .
  FILTER (?birth < "1900-01-01"^^xsd:date) .
}
```

Results: [Browse](#) [Go!](#) [Reset](#)

SPARQL results:

name	birth	death
"Helene" Ellen Franz"@en	"1839-05-30"^^xsd:date	"1923-03-24"^^xsd:date
"()"@en	"1811-10-29"^^xsd:date	"1873-06-06"^^xsd:date
(Carl Heinrich) Eduard Knoblauch Knoblauch"@en	"1801-09-25"^^xsd:date	"1865-05-29"^^xsd:date
Achim von Arnim"@en	"1781-01-26"^^xsd:date	"1831-01-21"^^xsd:date
Adalbert Of Prussia"@en	"1811-10-29"^^xsd:date	"1873-06-06"^^xsd:date
Adam Heinrich Müller"@en	"1779-06-30"^^xsd:date	"1829-01-17"^^xsd:date
Adam Müller"@en	"1779-06-30"^^xsd:date	"1829-01-17"^^xsd:date
Adolf Christen"@en	"1811-08-07"^^xsd:date	"1883-07-13"^^xsd:date
Adolf Heinrich von Arnim-Boitzenburg"@en	"1803-04-10"^^xsd:date	"1868-01-08"^^xsd:date
Adolf Heinrich von Arnim-Boitzenburg"@en	"1876-12-25"^^xsd:date	"1959-06-09"^^xsd:date
Adolf von Baeyer"@en	"1835-10-31"^^xsd:date	"1917-08-20"^^xsd:date
Agnes of Brandenburg"@en	"1584-07-17"^^xsd:date	"1629-03-26"^^xsd:date
Albert Frederick of Brandenburg-Schwedt"@en	"1672-01-24"^^xsd:date	"1731-06-21"^^xsd:date

DATABUS

About News Report Issue Sparql Endpoint Collections Login Register OPEN BETA

Snapshot of core data from en.wikipedia.org 2021-12
dbpedia » collections » dbpedia-snapshot-2021-12

SUMMARY

Label	Snapshot of core data from en.wikipedia.org 2021-12
Collection URI	https://databus.dbpedia.org/dbpedia/collections/dbpedia-snapshot-2021-12
Files	155
Size	15.58 GB
License(s)	http://purl.oclc.org/NET/rdflicense/cc-by3.0 http://creativecommons.org/licenses/by-sa/3.0/ http://purl.oclc.org/NET/rdflicense/cc-by-sa3.0

Previous Snapshot 2021-09

About

Releases of essential data of DBpedia quality-controlled since 2007. The collection focuses on English in this selection, but over 140 Wikipedia languages are available

<https://databus.dbpedia.org>

<https://dbpedia.org/snorql>

Knowledge Graphs

Large cross-domain knowledge bases which aim to cover all “relevant” entities in the world.

- Google Knowledge Graph
 - development started 2012, builds on Freebase
 - 570 million objects described by over 18 billion facts (2012)
 - 1500 classes, 35,000 properties
- Microsoft Satori Knowledge Base
 - revealed to the public in mid-2013
- Yahoo Knowledge Graph
 - revealed to the public early-2014
- Knowledge Graphs employ RDF-style graph data models

See also: IE650
Knowledge Graphs

Data Sources used to Build Knowledge Graphs

1. Wikipedia

- infoboxes, category system, information extraction from text

2. Open license sources

- e.g. CIA World Factbook, MusicBrainz, ...

3. Commercial third-party data

- e.g. IMDB, company listings, ...

4. schema.org annotations in web pages

- e.g. contact information for companies
- e.g. logos of companies

Lots of effort is spent on data integration and manual curation

Application of the Google Knowledge Graph

- Enrich search results with **knowledge cards** and lists
- Goal: Fulfil information need without having users navigate to other websites

The image displays two screenshots of Google search results. The top screenshot shows a search for "marie curie". The results page includes a sidebar with filters like "Everything", "Images", "Maps", etc., and a location dropdown set to "Mountain View, CA". The main results list includes links to Wikipedia and Nobel Prize biographies, with snippets of text and small thumbnail images. A detailed knowledge card for Marie Curie is highlighted with a blue border, showing her portrait, birth and death dates (November 7, 1867, Warsaw; July 4, 1934, Sancellemoz), her spouse (Pierre Curie), children (Irène Joliot-Curie, Ève Curie), and her discovery of Radium and Polonium. Below the card, a section titled "People also search for" lists Albert Einstein, Pierre Curie, Ernest Rutherford, Louis Pasteur, and John Dalton. The bottom screenshot shows a search for "things to do in paris". It features a sidebar for "Tourist attractions" with a list of sites. The main results list shows cards for the Eiffel Tower, Louvre, Notre Dame, Arc de Triomphe, Musée d'Orsay, Basilica of the Sacred Heart, and Centre Pompidou, each with a thumbnail image and a brief description.

Behind-the-Scenes Applications of KGs

Various tasks become easier, if you know all entities in the world.

- Google
 - uses its knowledge graph to identify entities in web pages (Entity Linking)
 - Hummingbird ranking algorithm (deployed in 2013) uses knowledge graph as background knowledge for ranking search results
- Yahoo
 - uses its knowledge graph to “support applications across the company:
 - Web Search, Content Understanding
 - Recommendation, Personalization, Advertisement
- Data Integration
 - becomes matching data sources against knowledge graphs as intermediate schemata (see Table Annotation)



SEO Topic: How to influence Knowledge Graphs?

J.Crew

Specialty retailer company

J.Crew Group, Inc., is an American multi-brand, multi-channel, specialty retailer. The company offers an assortment of women's, men's and children's apparel and accessories, including swimwear, outerwear, ... [Wikipedia](#)

Customer service: 1 (800) 562-0258

Headquarters: New York City, NY

CEO: Mickey Drexler

Founder: Emily Scott

Founded: 1983



Logo: can be specified by using Organization Schema Mark-up

Company type: can be influenced by Wikidata or Wikipedia

Company details: can be influenced by Wikidata, Wikipedia, organization and local business schema mark-up

Profiles



Facebook



Instagram



Twitter



LinkedIn



Google+

Social profiles: can be influenced by organization schema mark-up with social links specified

Recent posts on Google+



J.Crew

1,488,424 followers • Shared publicly



Because THIS is the summer you finally learn how to surf. And THESE are the board shorts you'll be wearing when you catch the first wave.
<http://jcrew.co/LVNQ1> ... 6 hours ago

Google + feed: can be influenced by Rel Publisher linking

People also search for



Banana Republic



Nordstrom



Anthropol...



ANN TAYLOR
Ann Inc.



Vineyard Vines

View 15+ more

Related companies/brands: cannot be influenced, entirely controlled by Google

<http://searchengineland.com/leveraging-wikidata-gain-google-knowledge-graph-result-219706>

Feedback

Summary

- Web data integration might involve millions of data sources
 - integration in corporate data lakes often remains below 20 sources
- The topics of the published data partly correlate with the publication methods used:
 - Data Portals: public sector data, statistical data, research data
 - Web APIs: user generated content, location-related data, weather data
 - Schema.org data: e-commerce, local business, event, job data
 - Linked Data: library data, research data, government data
 - Wikipedia, HTML tables: General knowledge
- The Web is the perfect playground for researching and applying Big Data Integration techniques
 - tough challenges concerning heterogeneity, volume, and data quality
 - rewarding if challenges can be handled, e.g. web-scale queries and mining

5. References

- Linked Data
 - Max Schmachtenberg, Christian Bizer, Heiko Paulheim: Adoption of the Linked Data Best Practices in Different Topical Domains. In: 13th International Semantic Web Conference, 2014.
- RDFa, Microdata and Microformats
 - Christian Bizer, et al.: Deployment of RDFa, Microdata, and Microformats on the Web – A Quantitative Analysis. 12th International Semantic Web Conference, 2013.
- Extracting HTML Table Data
 - Michael Cafarella, Alon Halevy, et al.: Ten Years of Web Tables. Proceedings of the VLDB Endowment, 2018.
- Wrapper Induction
 - Furche, Gottlob, et al: The Ontological Key: Automatically Understanding and Integrating Forms to Access the Deep Web. The VLDB Journal, 2013.
 - Lockard, Dong, et al: CERES: Distantly Supervised Relation Extraction from the Semi-Structured Web. PVLDB, 2018.
- Knowledge Graphs
 - Jens Lehmann, et al: DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal, 2014.