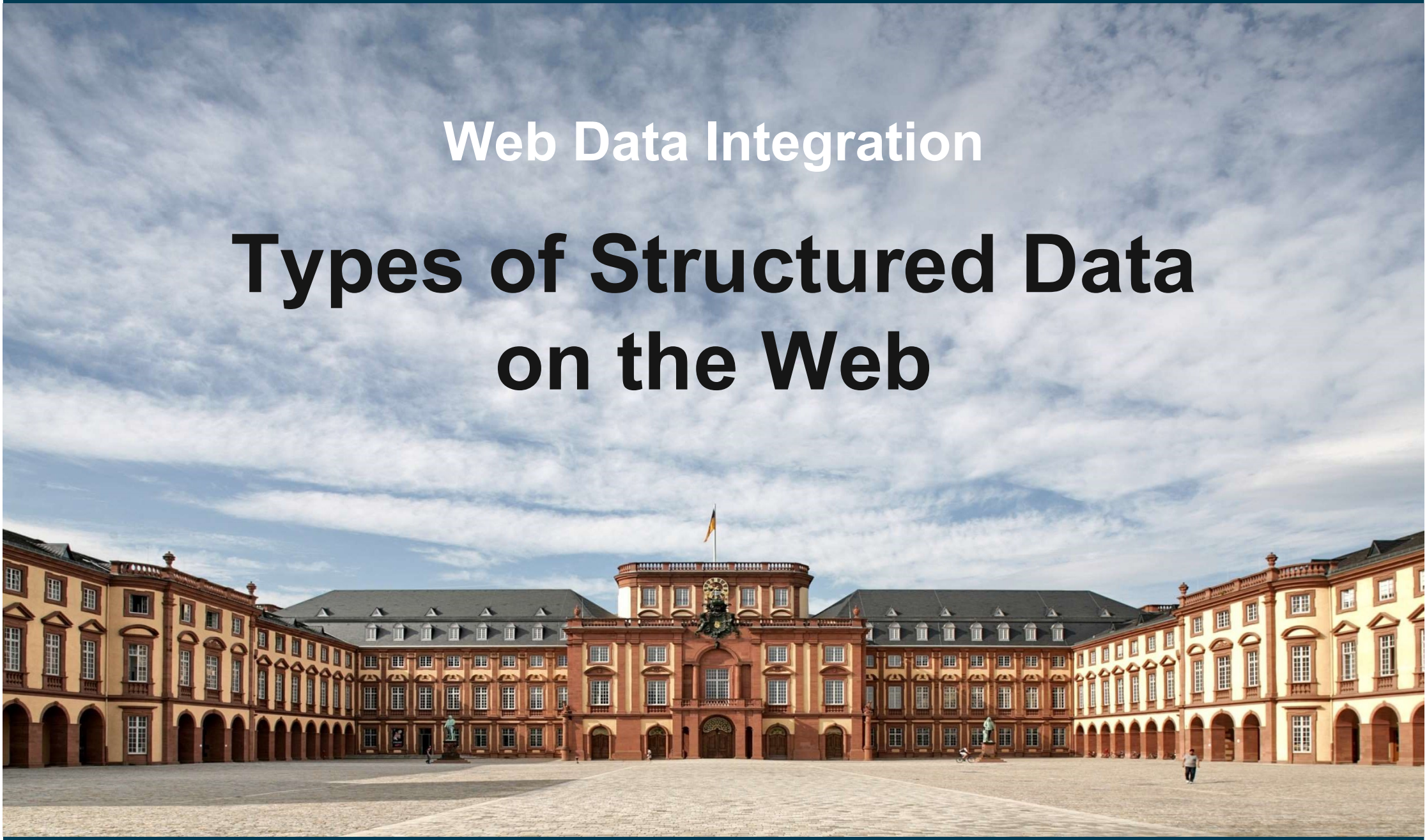
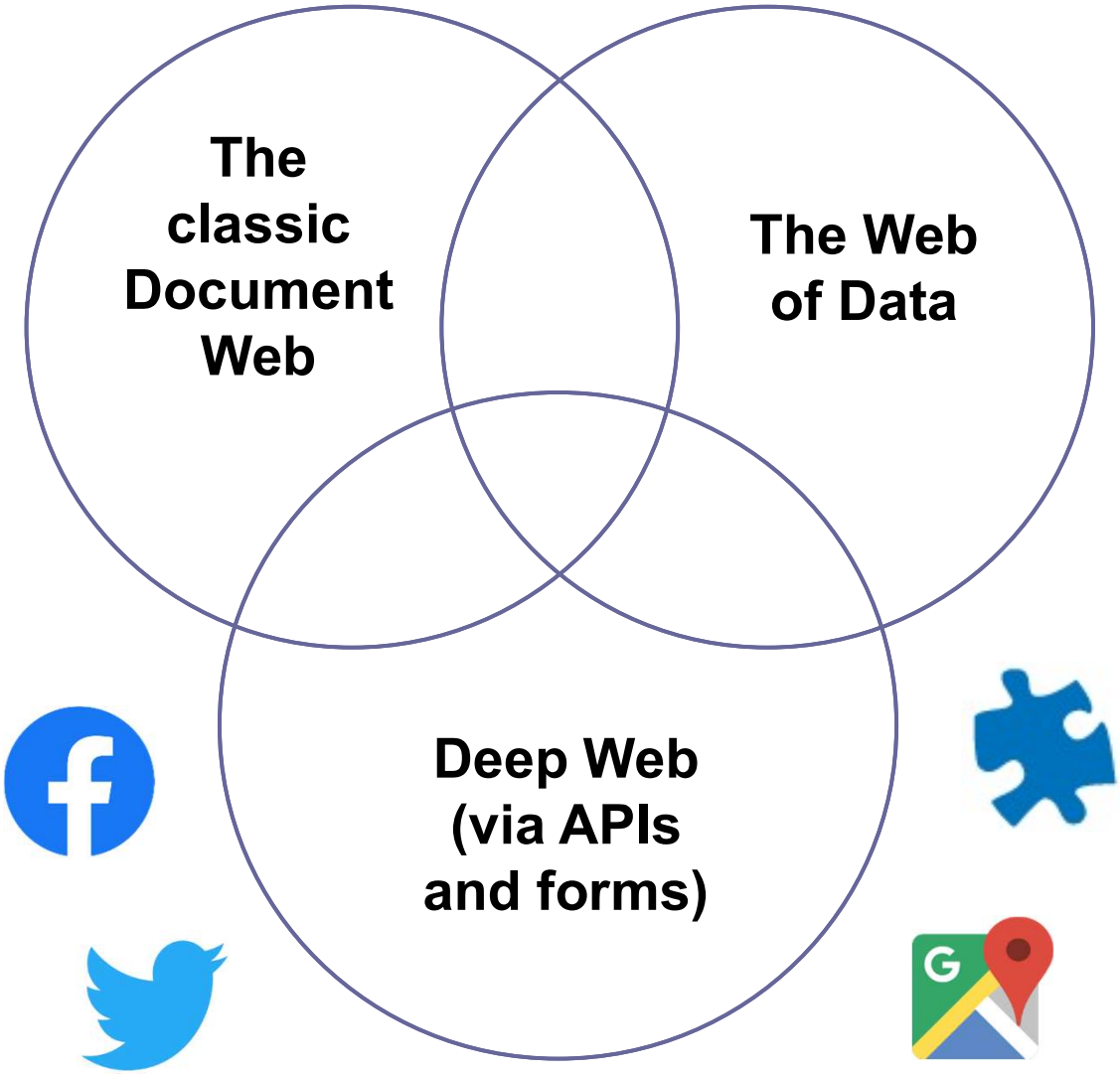
UNIVERSITY OF MANNHEIM

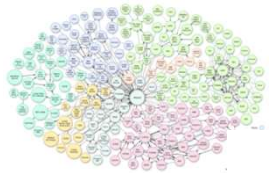
Web Data Integration

Types of Structured Data on the Web



Topology of the Web Today



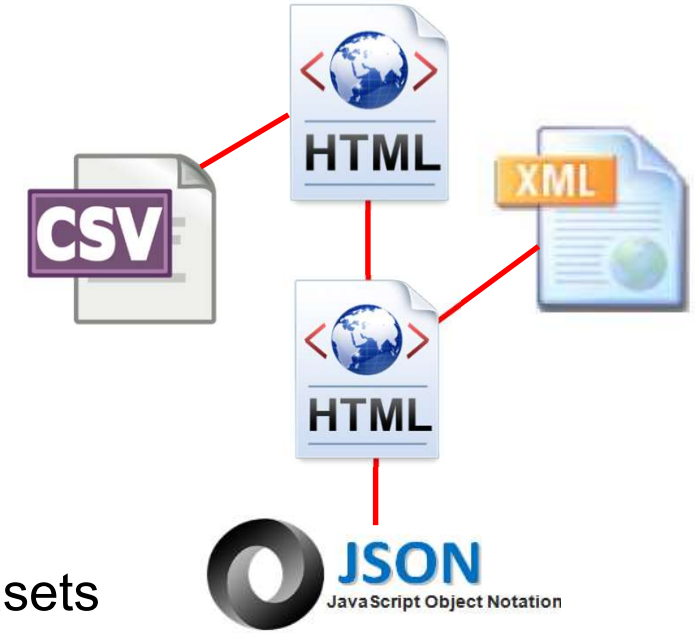


Outline

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

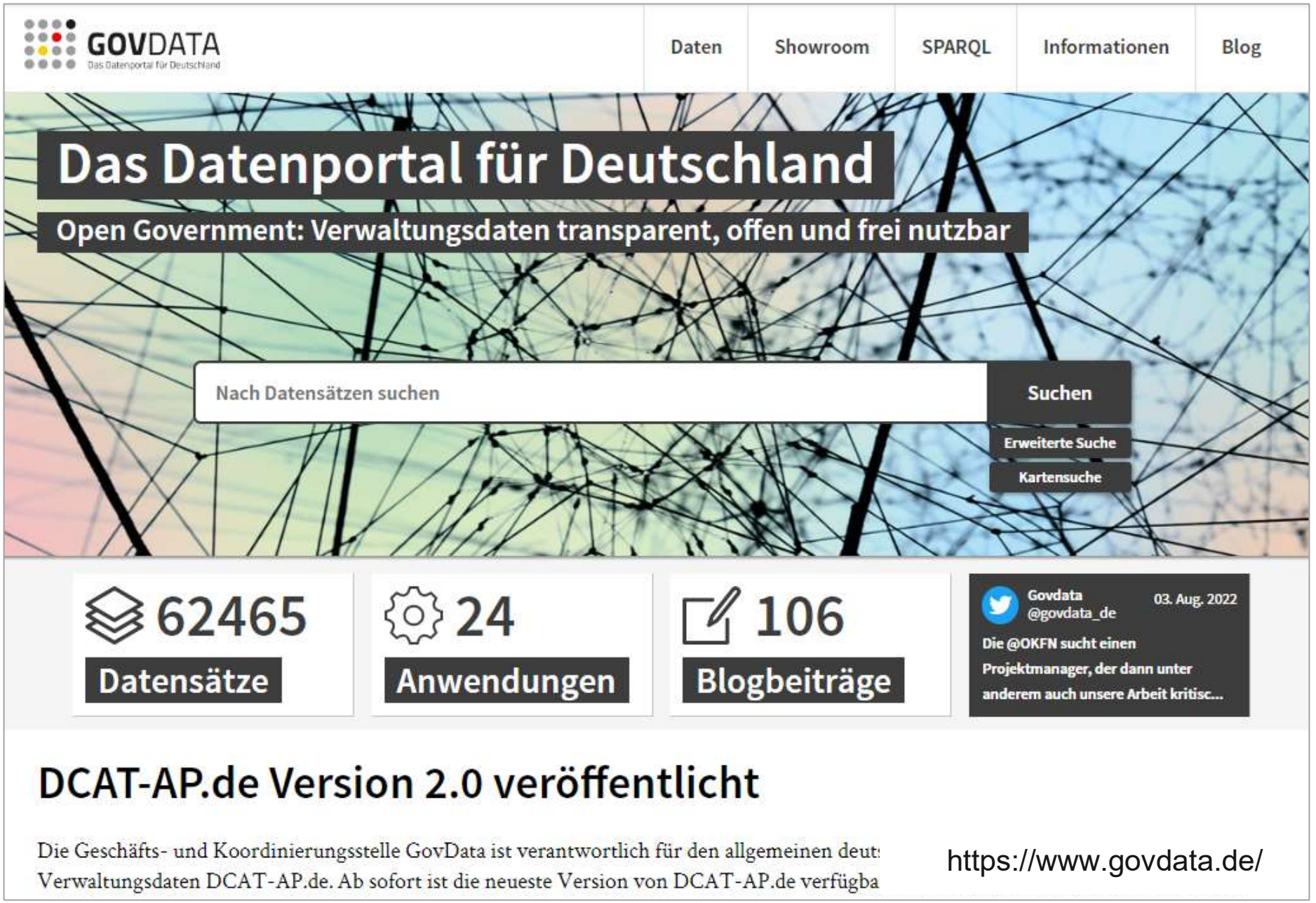


- Data Portals and Data Marketplaces
  - collect and host datasets
  - collect and generate metadata describing the datasets
  - provide for data search and exploration
  - provide free or payment-based access to data
- List of Data Catalogs
  - http://data.wu.ac.at/portalwatch/portalslist

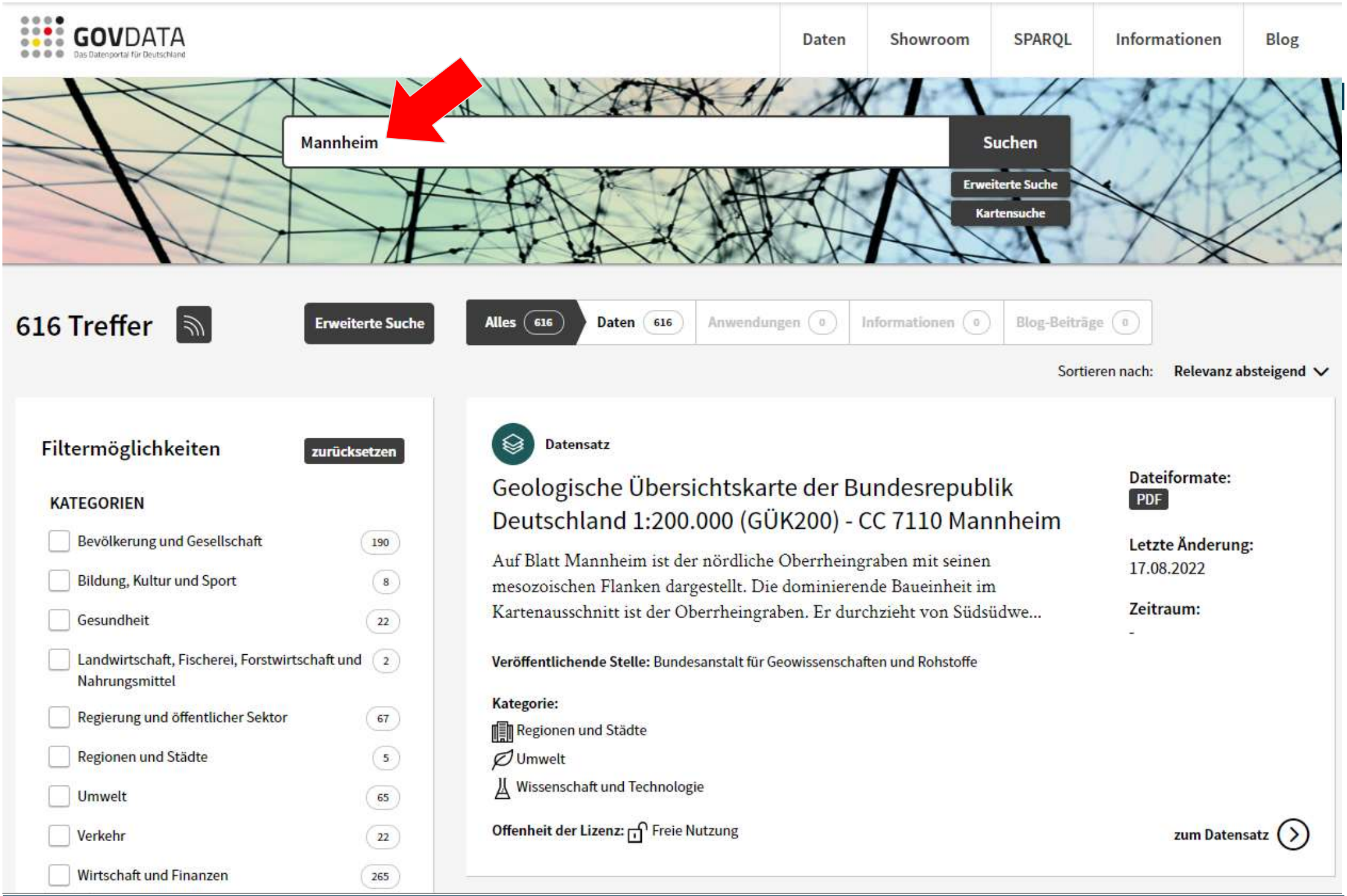
Main Types of Shared Data

- 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

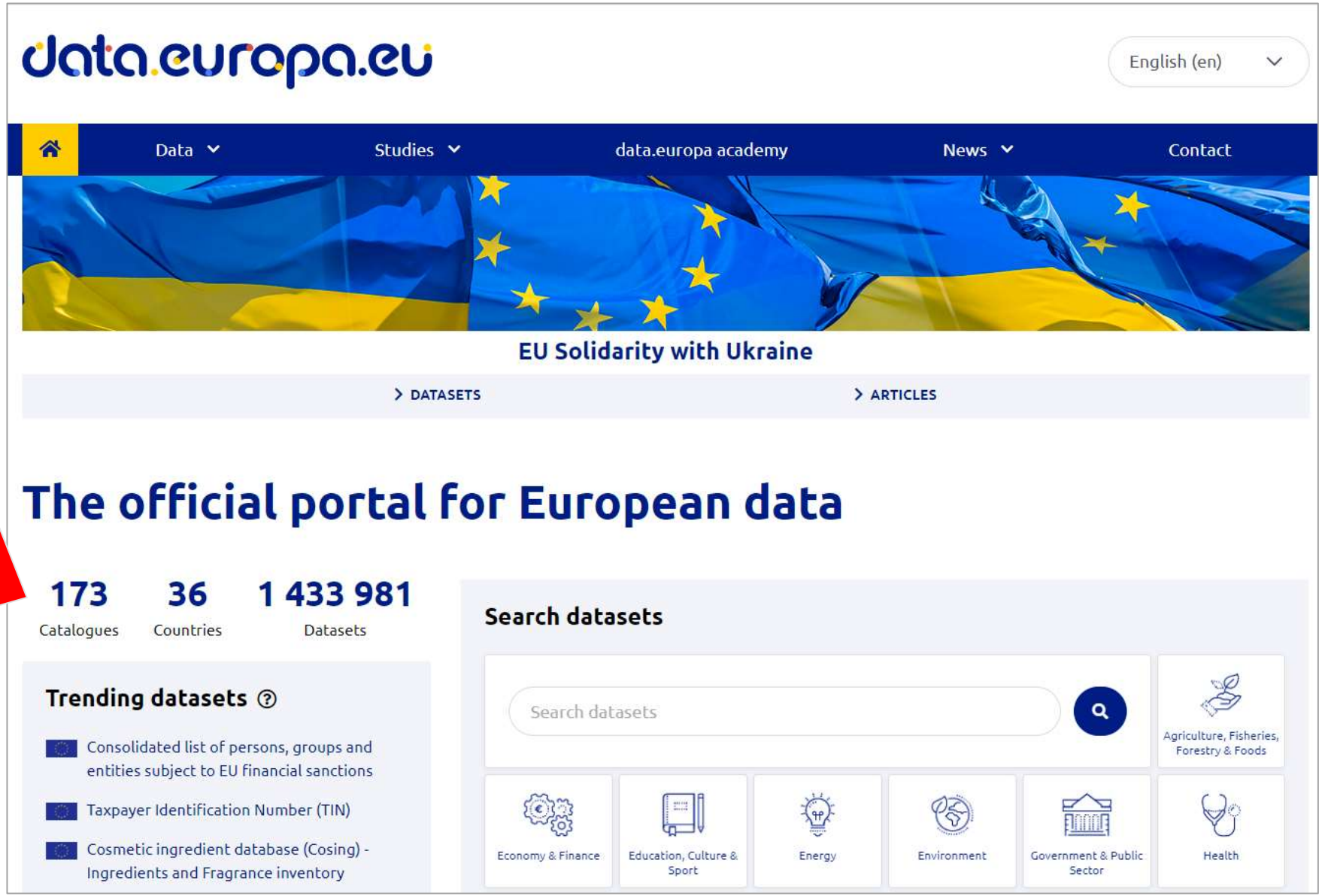
Example: Government Data Portal



https://www.govdata.de/

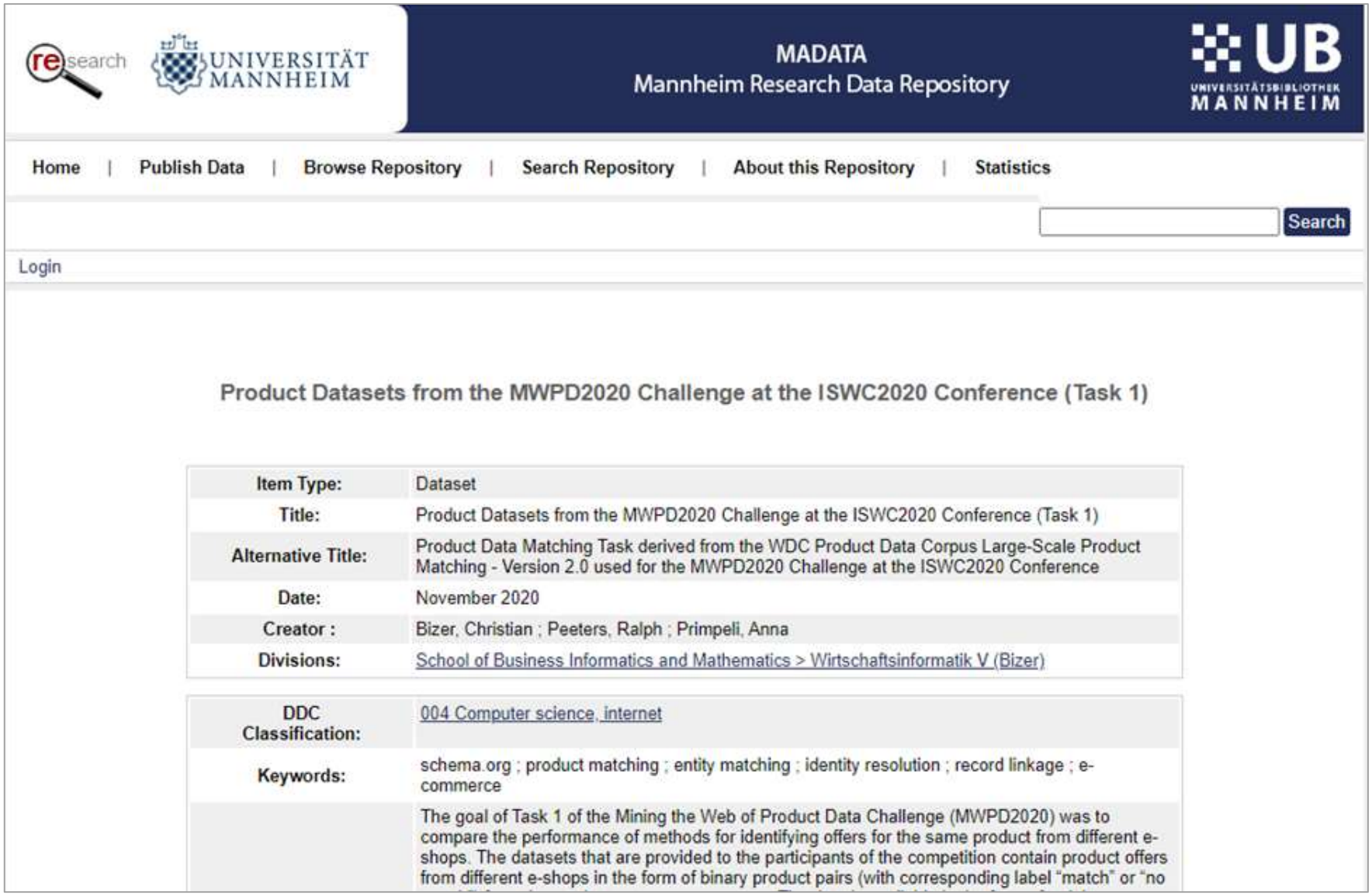

GOVDATA
Das Datenportal für Deutschland
Daten
Showroom
SPARQL
Informationen
Blog
Mannheim
Suchen
Erweiterte Suche
Kartensuche
616 Treffer
Erweiterte Suche
Alles 616
Daten 616
Anwendungen 0
Informationen 0
Blog-Beiträge 0
Sortieren nach: Relevanz absteigend
Filtermöglichkeiten
zurücksetzen
KATEGORIEN
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 Südsüdwe...
Veröffentlichende Stelle: Bundesanstalt für Geowissenschaften und Rohstoffe
Kategorie:
Regionen und Städte
Umwelt
Wissenschaft und Technologie
Offenheit der Lizenz: Freie Nutzung
Dateiformate:
PDF
Letzte Änderung:
17.08.2022
Zeitraum:
-
zum Datensatz

Example: Portal aggregating Metadata from other Portals



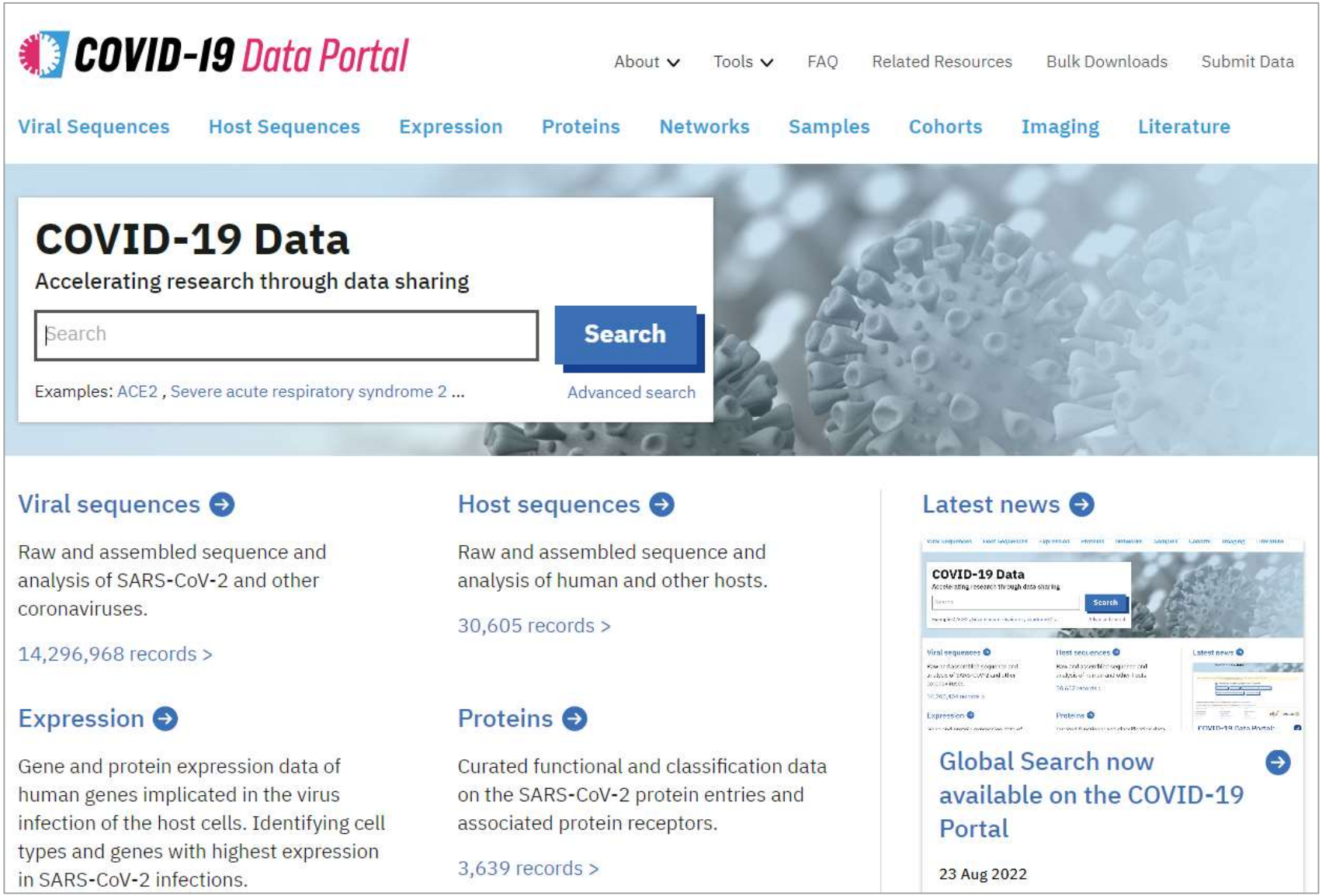
https://data.europa.eu/en

Example: Institutional Research Data Repository



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

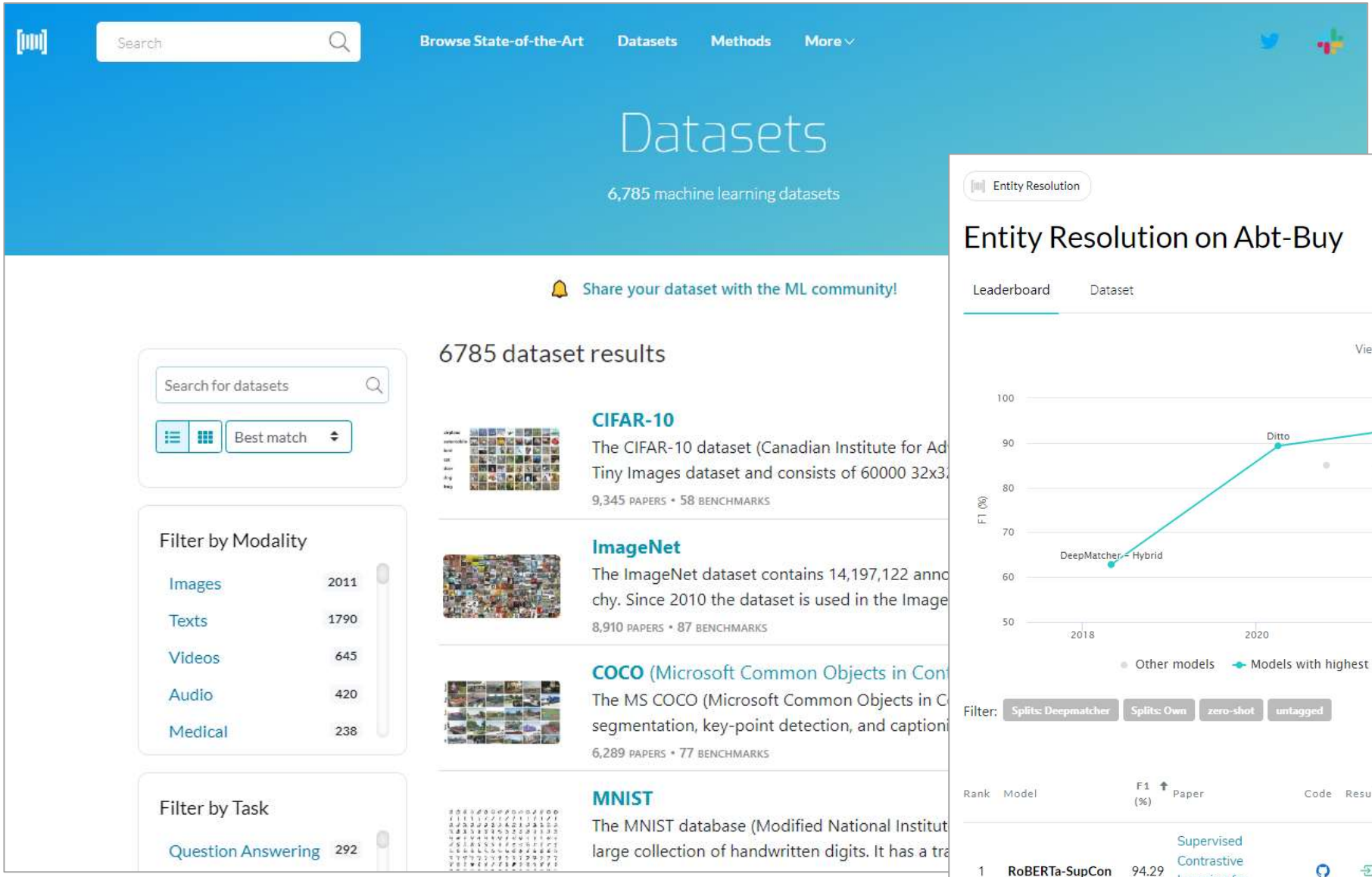
Example: Focused Research Data Portal



EMBL Heidelberg

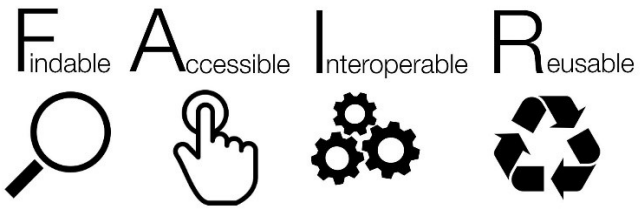
https://www.embl.org/news/science/embl-ebi-launches-covid-19-data-portal/

Example: Data Portal facilitating the Replication of Research



https://paperswithcode.com/datasets

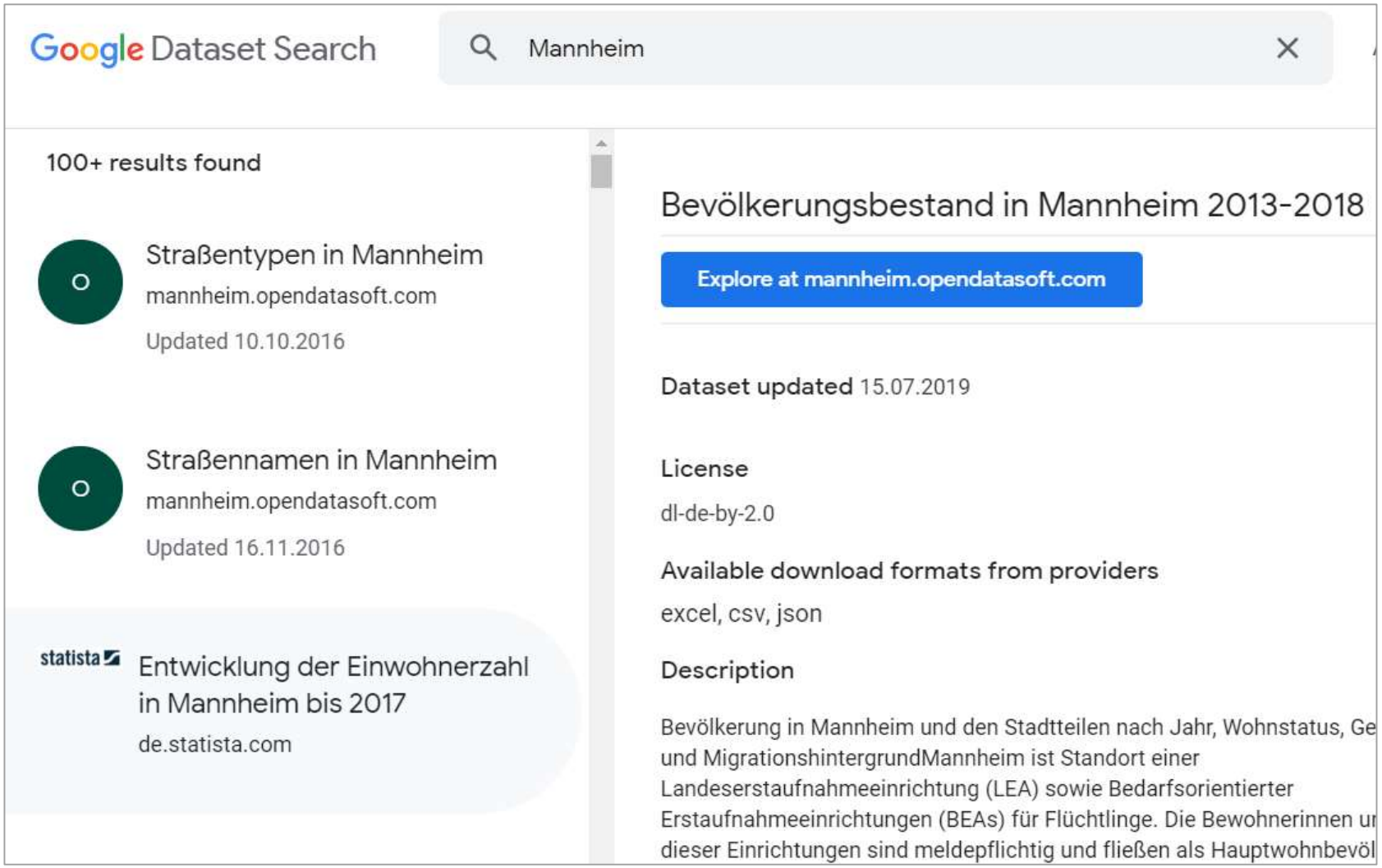
The FAIR Data Principles



- **F**indable
  - (Meta)data are assigned a globally unique identifier
  - Data are described with rich metadata
  - (Meta)data are registered or indexed in a searchable resource
- **A**ccessible
  - (Meta)data are retrievable by their identifier using a standardised communications protocol
  - Metadata are accessible, even when the data are no longer available
- **I**nteroperable
  - (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
- **R**eusable
  - (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



Crawls dataset metadata from the Web.

schema.org

https://datasetsearch.research.google.com/

2. Web 2.0 Applications and Web APIs

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

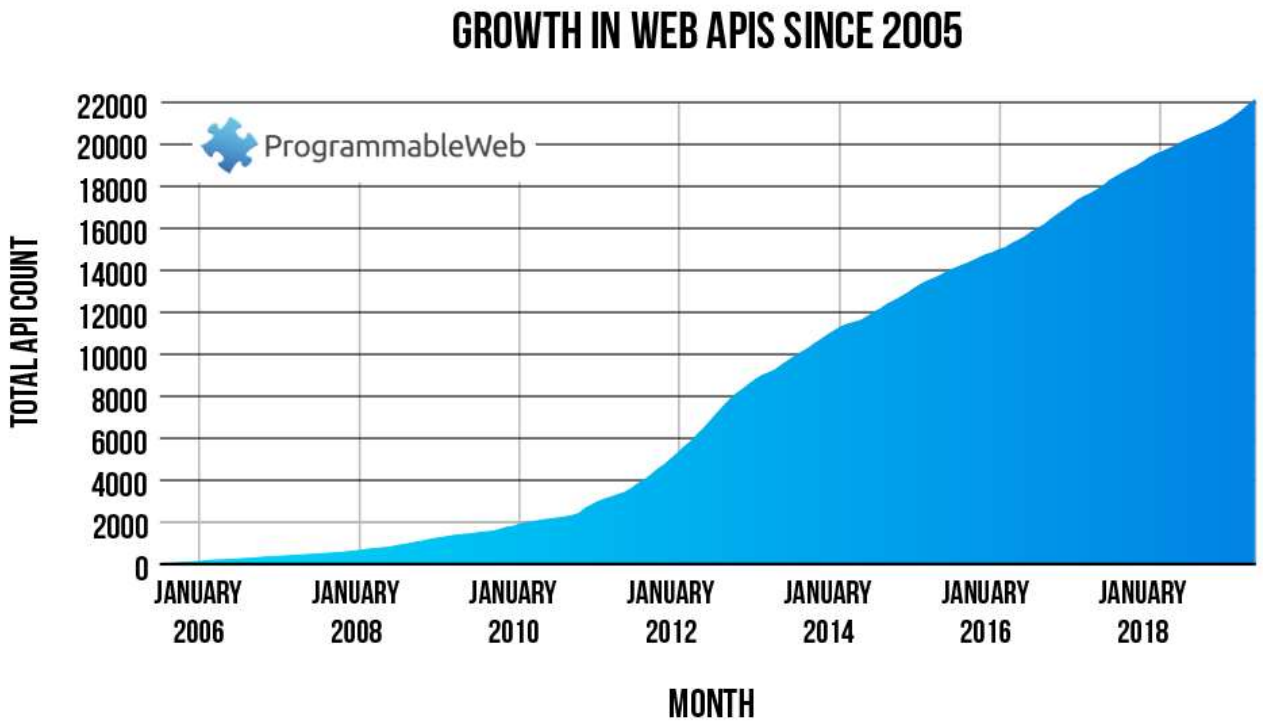
Example: Size of Facebook Social Graph

- 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

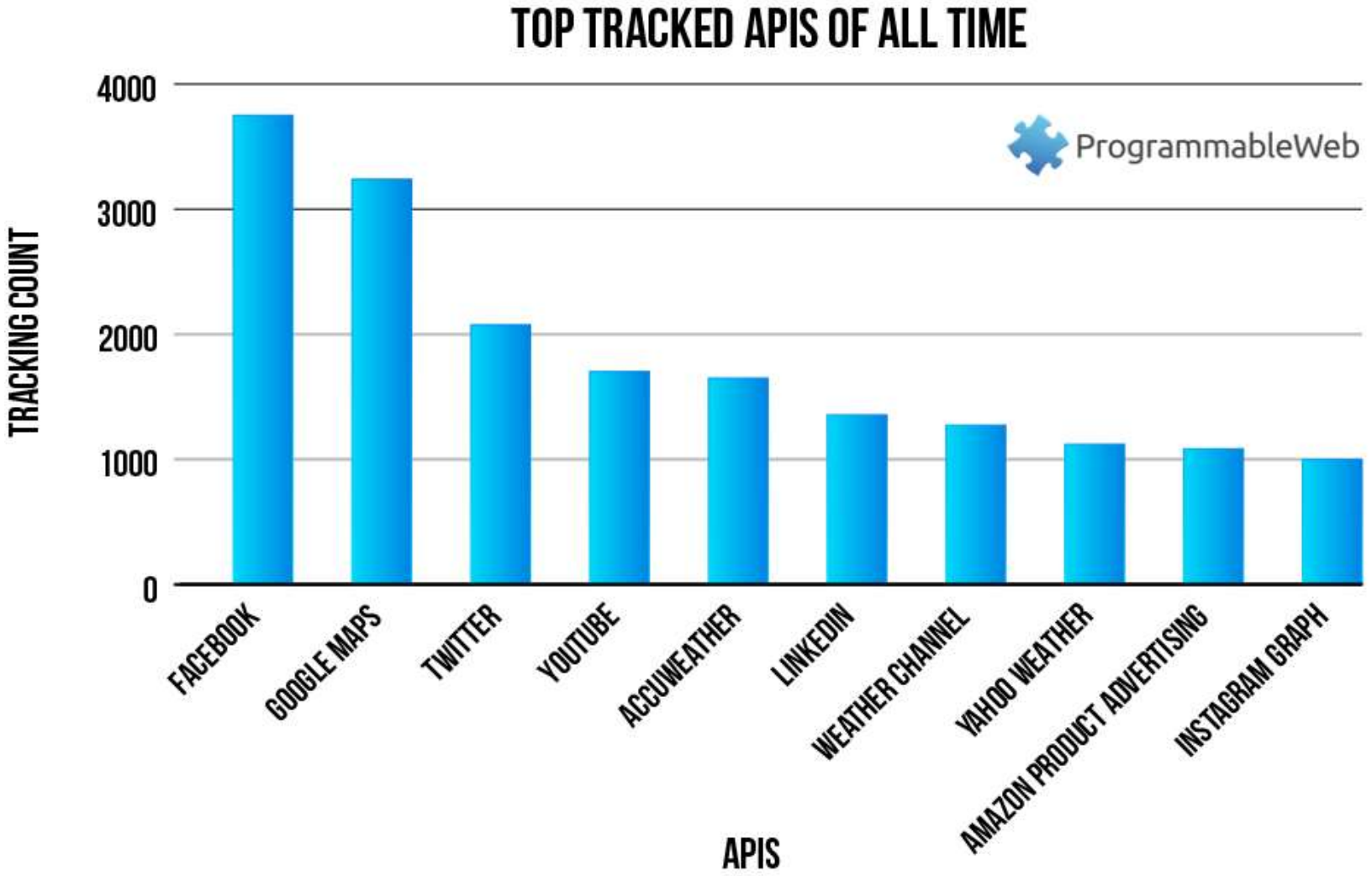
https://www.brandwatch.com/blog/facebook-statistics/
http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/

Web APIs

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

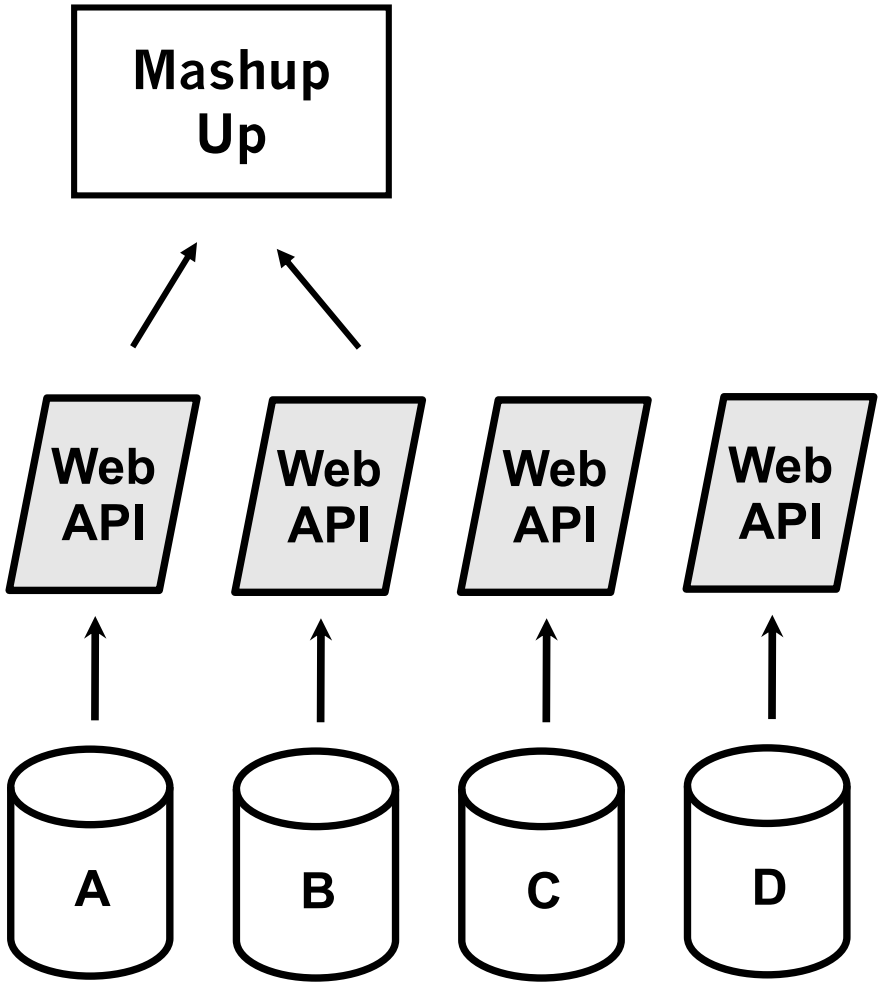


Most Popular APIs



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

Mashups are based on a fixed set of data sources



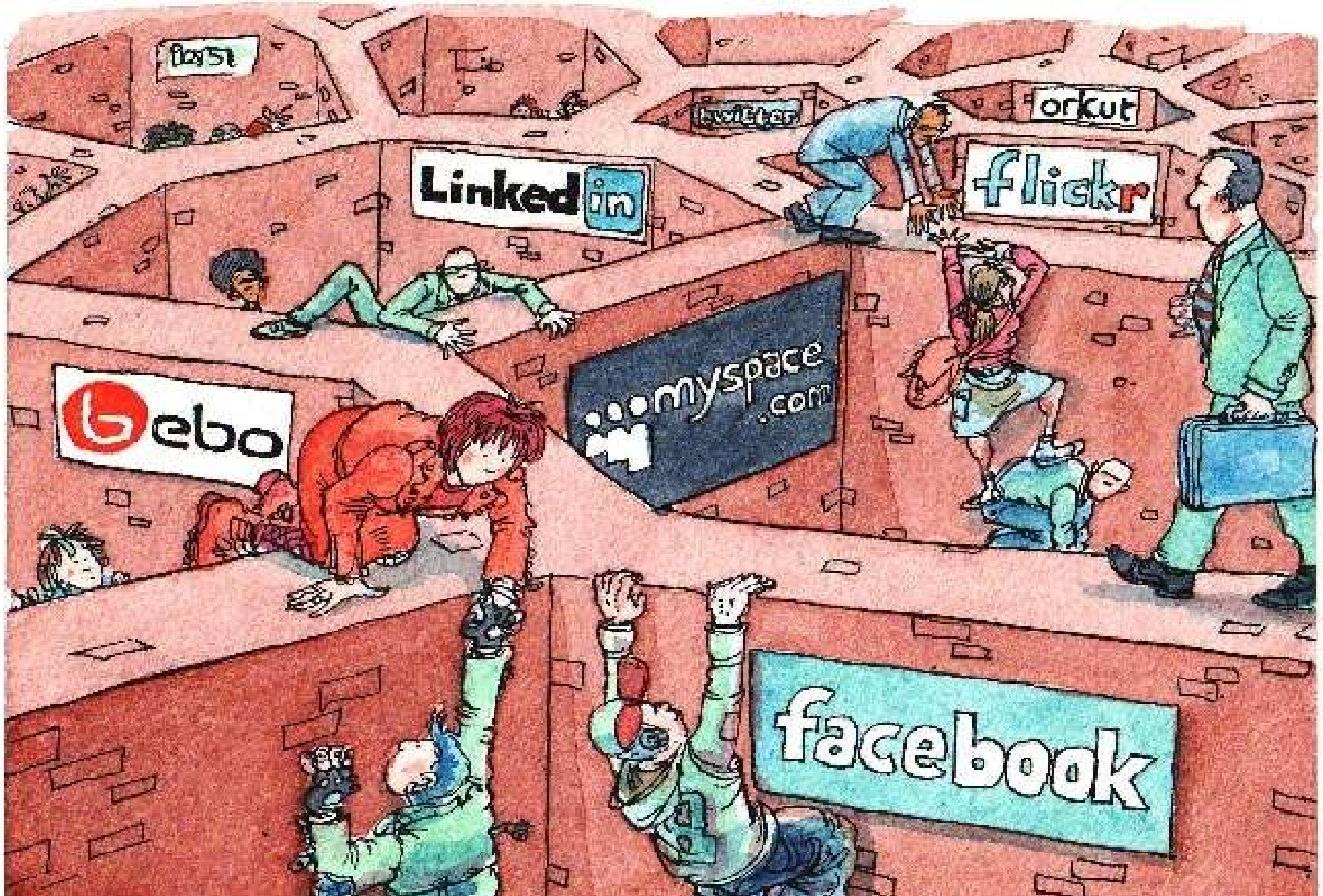
Web APIs expose proprietary interfaces

→ Not index-able by generic web crawlers

→ No automatic discovery of additional data sources

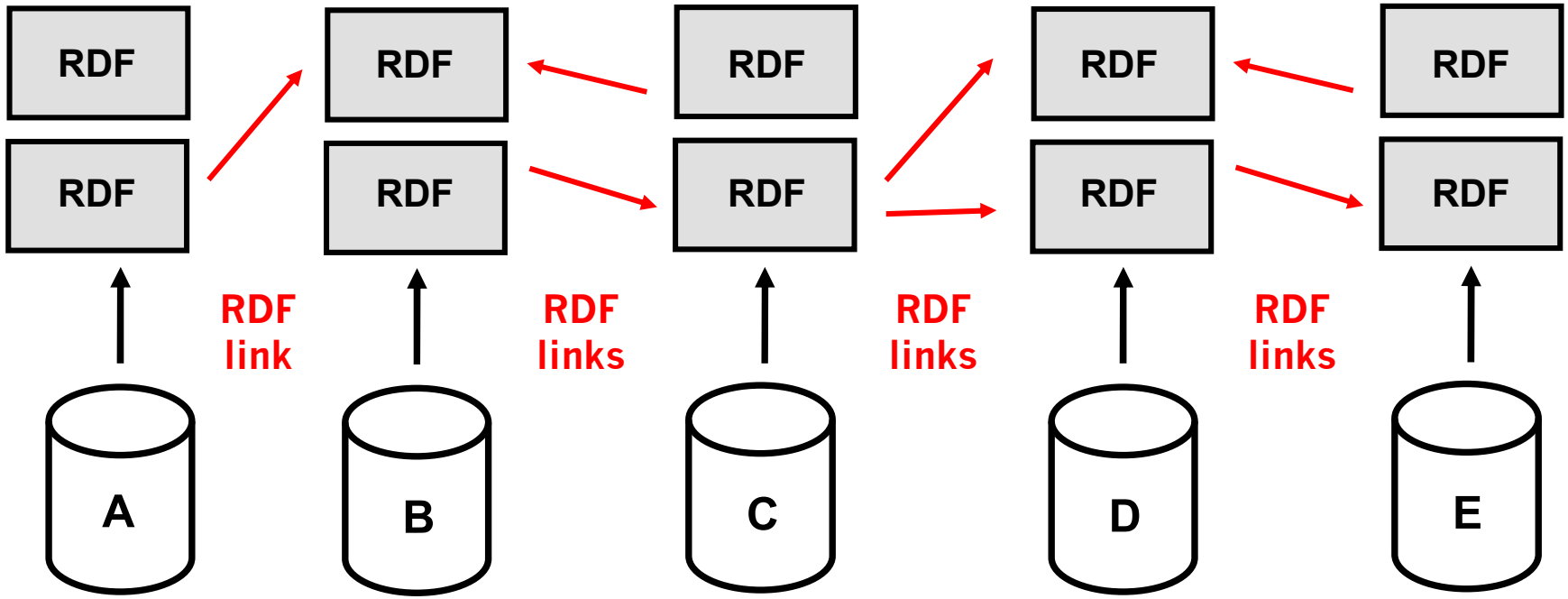
→ No single global data space

Web APIs slice the Web into Data Silos

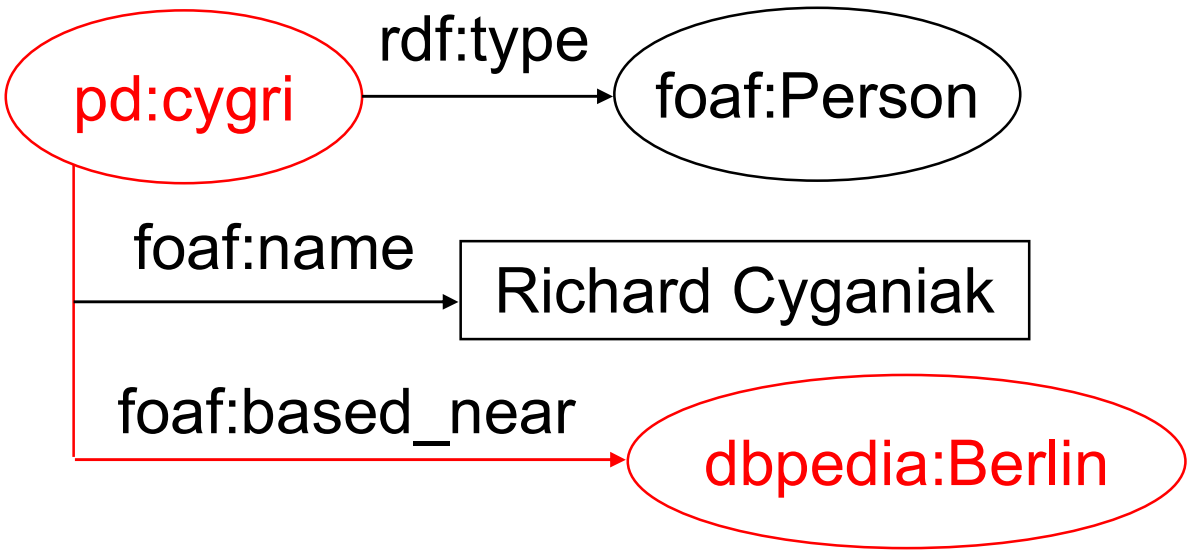


3. Alternative Approach: Linked Data

- 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

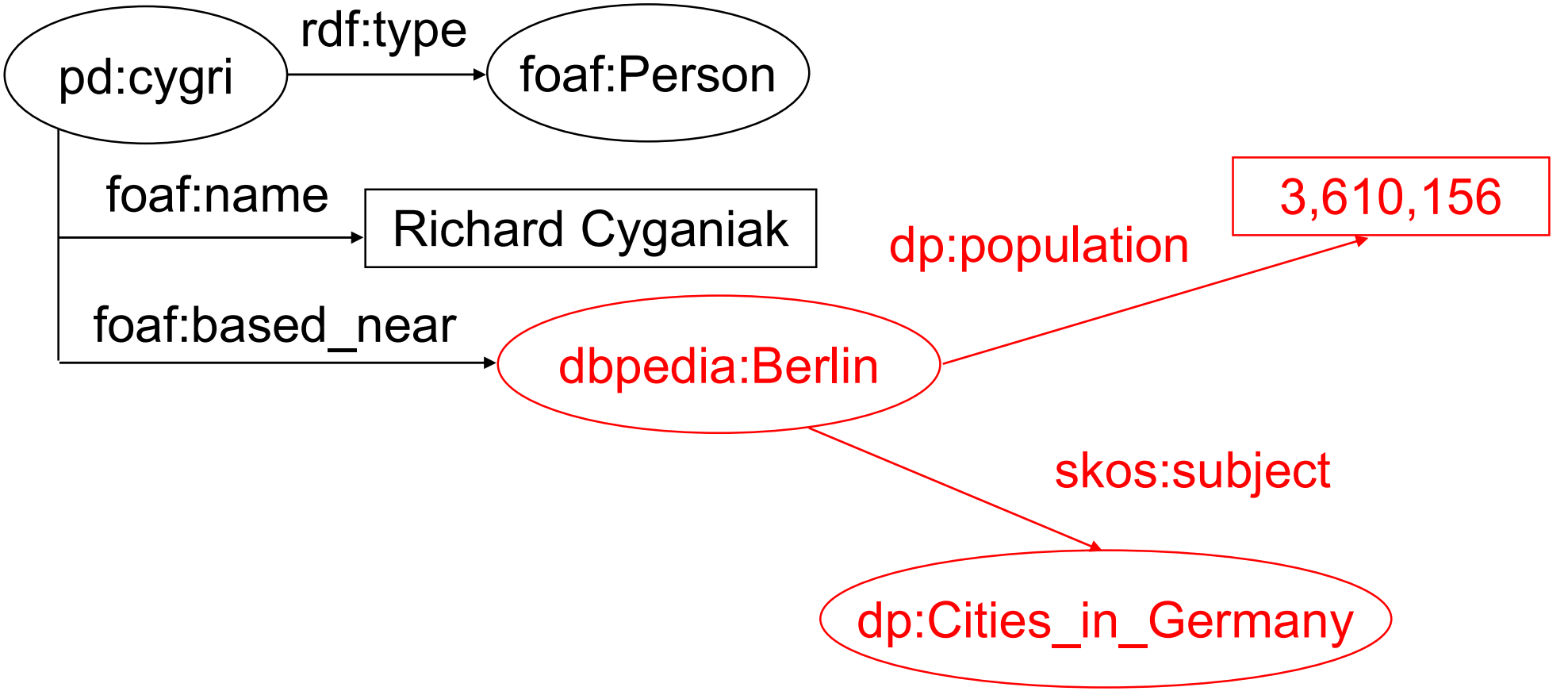


HTTP URIs take the role of global primary keys.

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

marbles

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.wiwiss.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	• 965c47c5a70db7407210cef6e4e6f5374a525c5c
workplace homepage	• http://www.w3.org/
nickname	• TimBL
nickname	• TimBL • timbl
personal mailbox	• mailto:timbl@w3.org

The SigMa Linked Data Search Engine

SIG.MA
SEMANTIC INFORI
MASHUP

Help | About | Forum

Chris Bizer | Add More Info | Start New | Options | Order | Permalink

Chris Bizer

picture:  [3]  [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 Releational Databases on the Semantic Web. [16]

Named Graphs, Provenance and Trust [16]

hide value | just this value | which sources | reject sources [6]

Sources (20) | Appro

1 Chris Bizer - Free Univ
http://videolectures.net/chris_

2 Chris Bizer - semantic
http://ontoworld.org/wiki/Chris

3 Untitled document 6 f
BOSS http://www.faceboo

4 Chris Bizer - semantic
http://semanticweb.org/wiki/Ch

5 Chris Bizer - LinkedIn
BOSS http://www.linkedin.

6 Chris Bizer 10 facts | 20
http://data.semanticweb.org/p

7 Chris Bizer - semantic
http://semanticweb.org/index.p

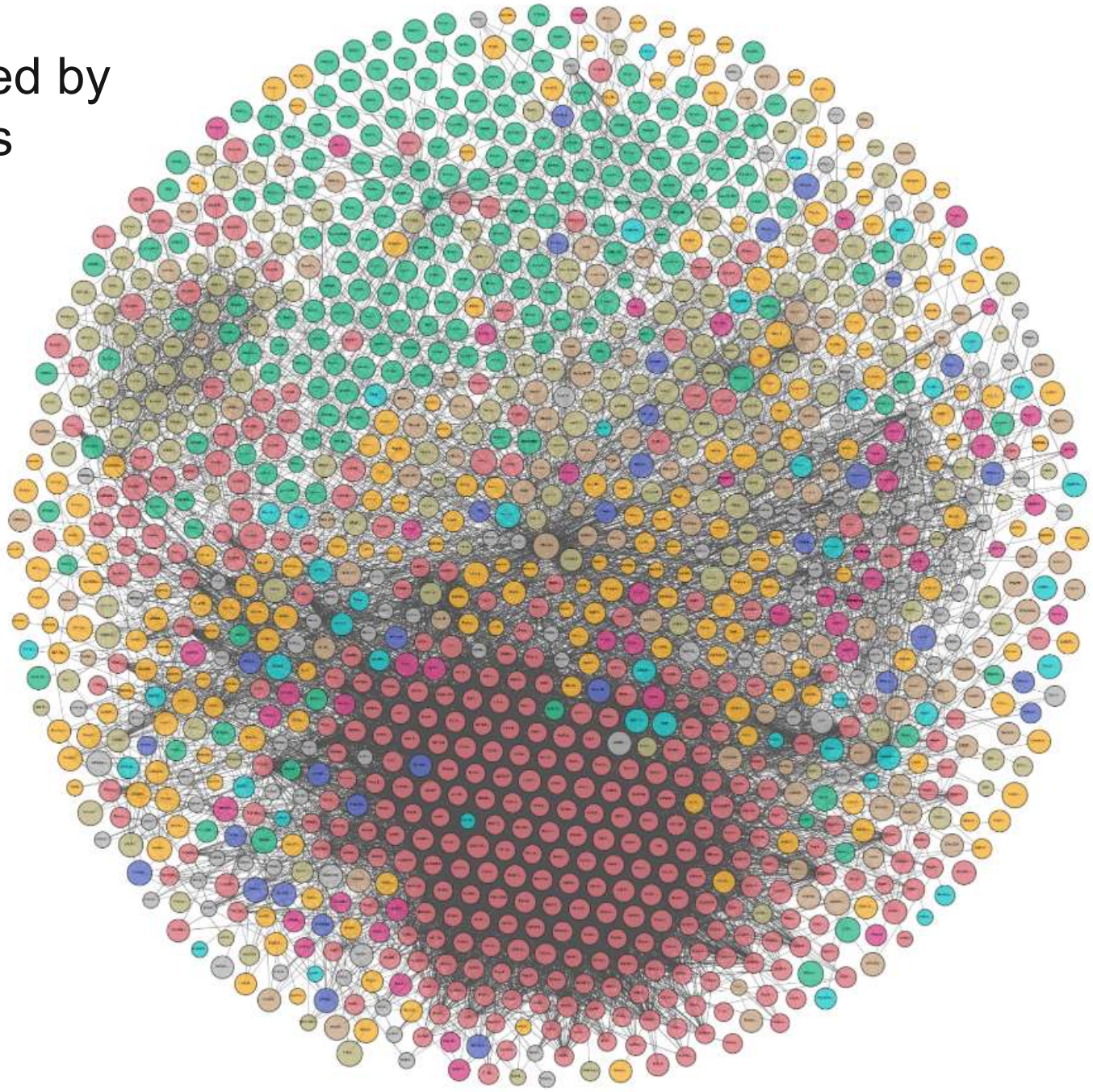
8 Flickr: Chris Bizer's Ph
BOSS http://flickr.com/pho

9 Untitled document 8 f
http://data.semanticweb.org/c

10 Chris Bizer 6 facts | 20
BOSS http://ebiquity.umbc

The Linked Open Data Cloud

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



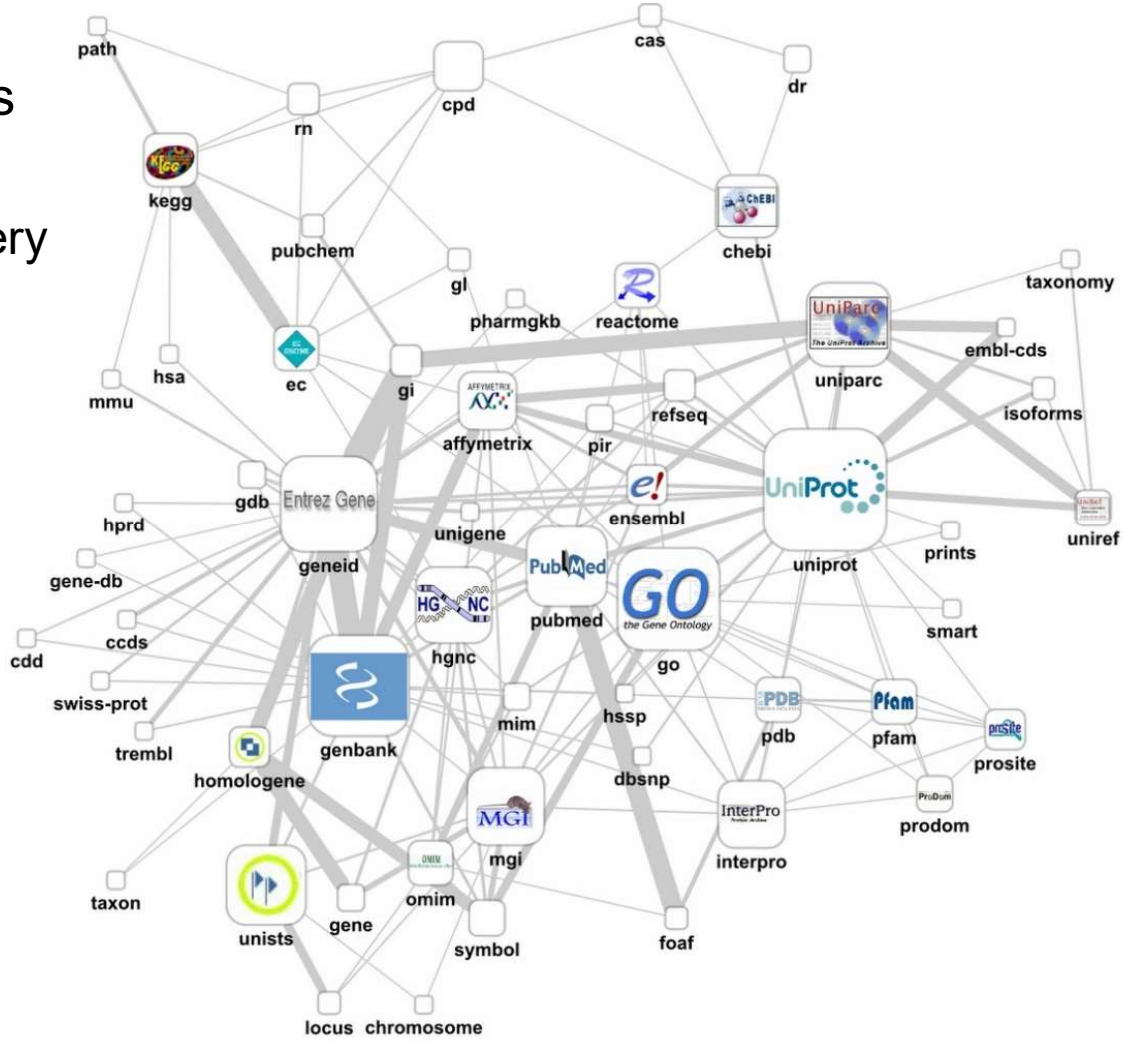
Legend

- Cross Domain
- Geography
- Government
- Life Sciences
- Linguistics
- Media
- Publications
- Social Networking
- User Generated

https://lod-cloud.net/

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



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)

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 

RDFa 

Microdata 

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

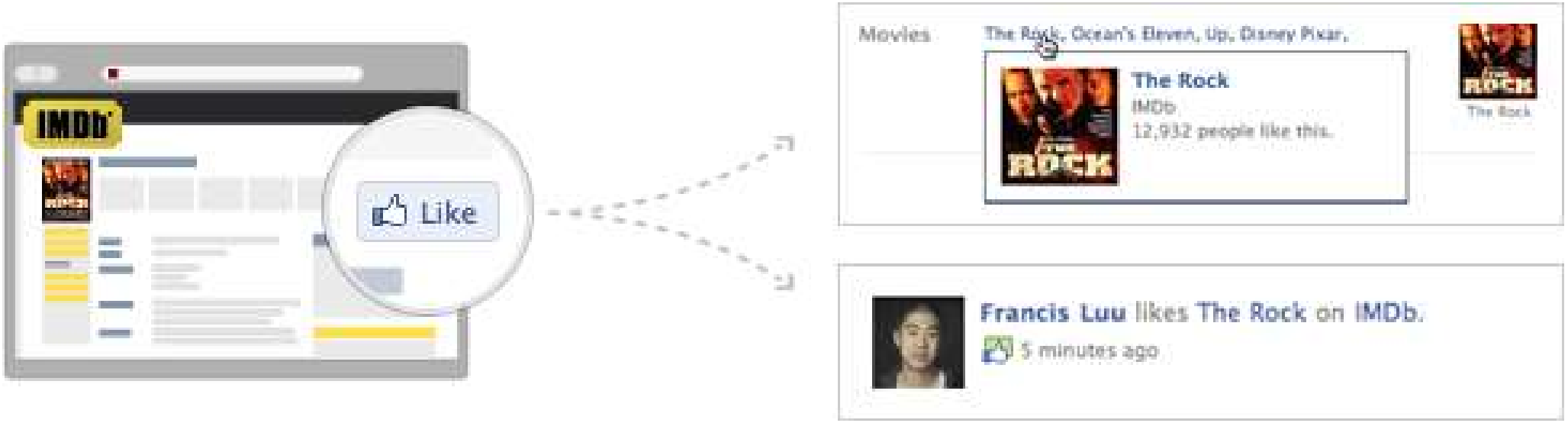
RDFa

- 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
            Jones</a>.
8     </div>
9 ...
```

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



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>
```

Schema.org

Google bing YAHOO! Yandex

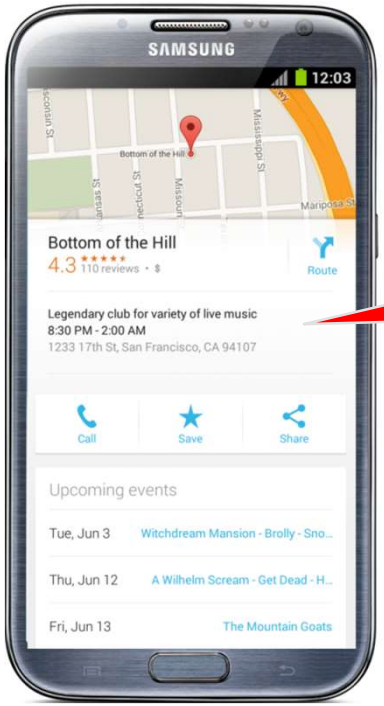
- ask site owners since 2011 to annotate data for enriching search results
- 675 Types: Event, Local Business, Product, Review, Job Offer
- Encoding: Microdata, RDFa, JSON-LD

HTML RDFa JSON-LD

Usage of Schema.org Data @ Google

Gramercy Tavern - Flatiron - New York, NY | Yelp
www.yelp.com › Restaurants › American (New)
★★★★½ 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
★★★★½ 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 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

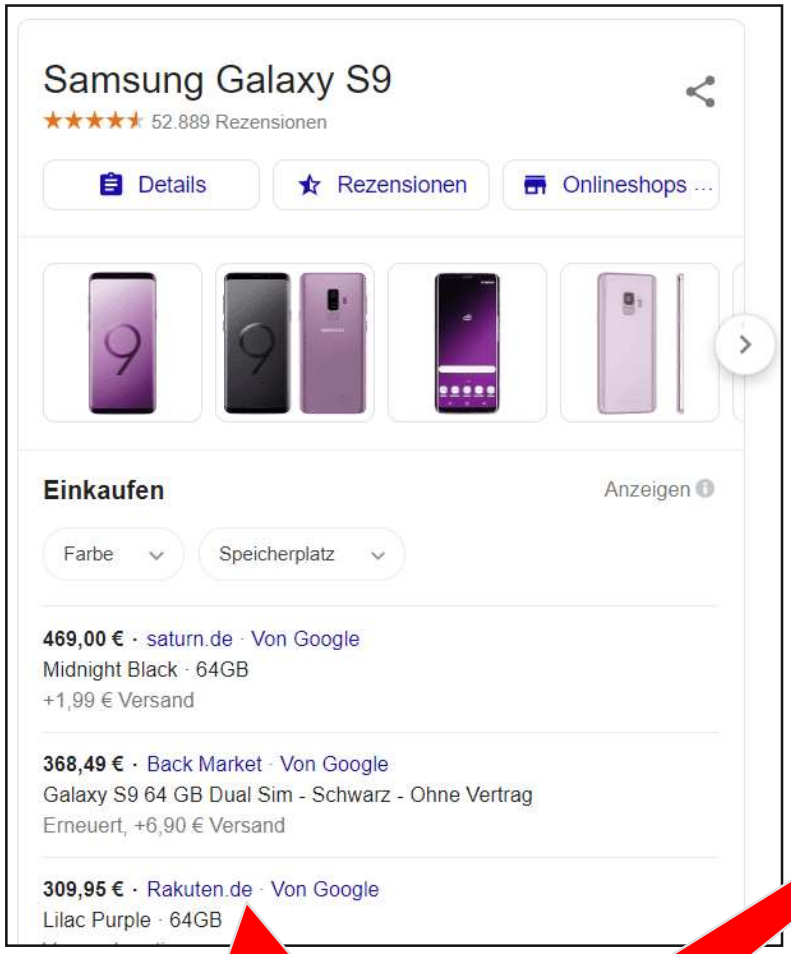
Record labels: Fat Possum Records, Nonesuch Records, V2 Records, Alive Naturalsound Records

Awards: Grammy Award for Best Rock Album, more

Upcoming events

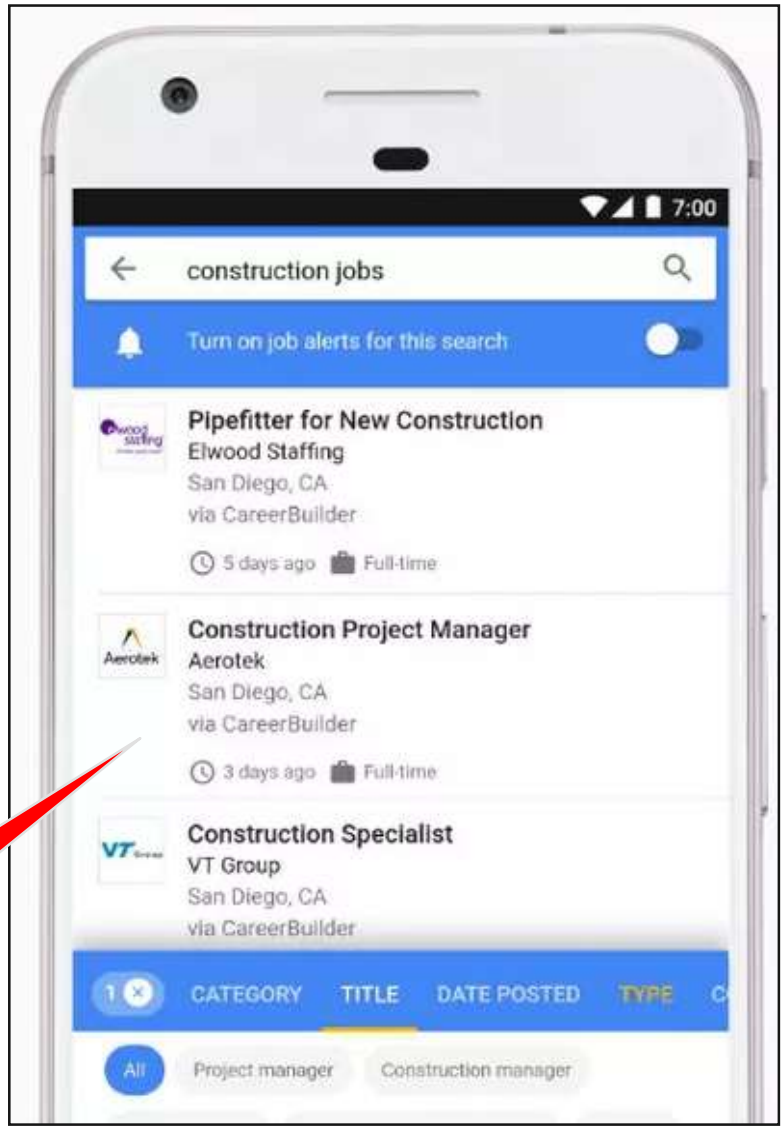
Jun 20 Fri	The Black Keys Neuhausen ob Eck (near you)
May 16 Fri	The Black Keys Gulf Shores, AL
Jun 22 Sun	The Black Keys Scheeßel

Usage of Schema.org Data @ Google



Product offers

Job offers



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

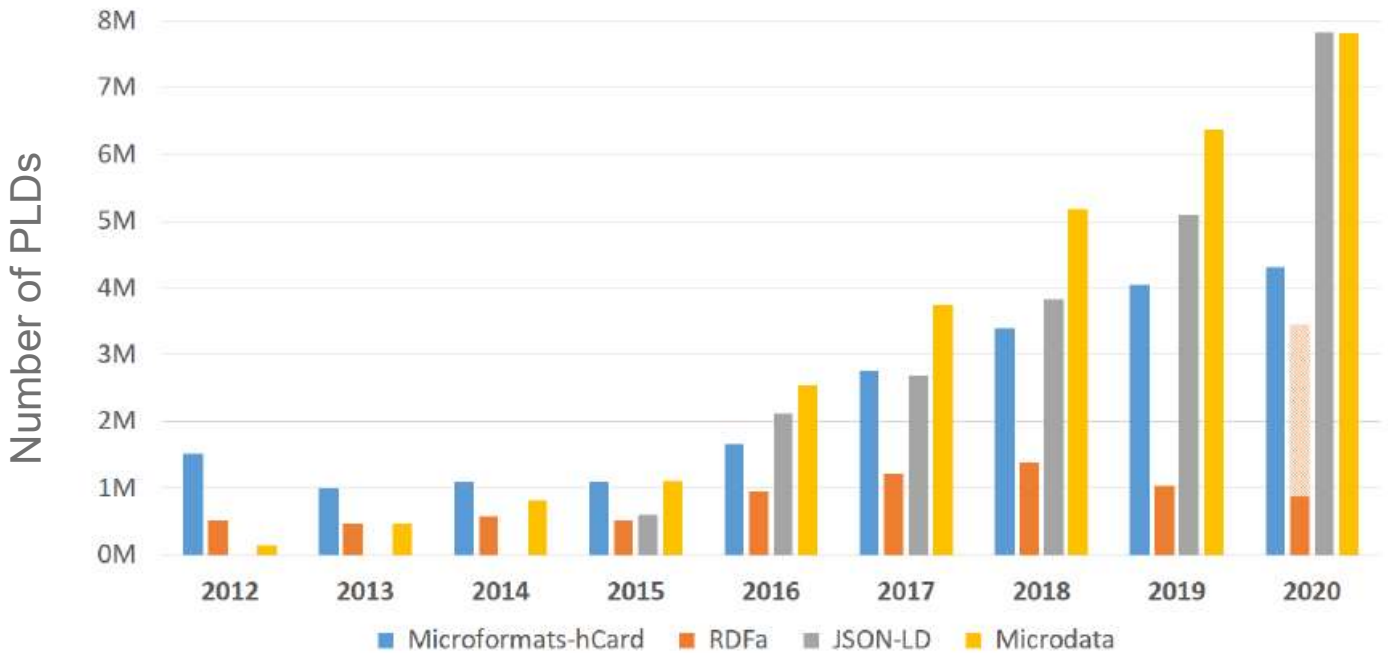
The Web Data Commons Project

- 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

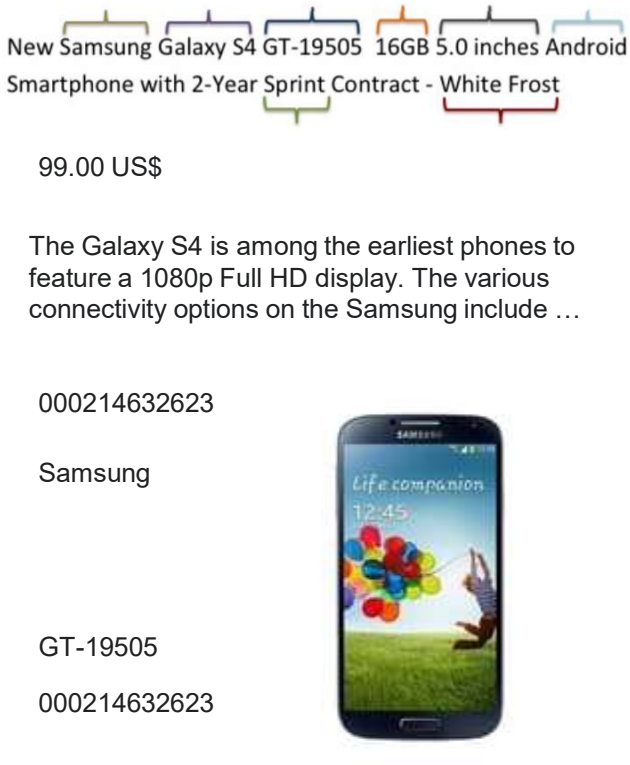
Adoption by Travel Websites

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

	Name	Population	Latitude/Longitude
1	Berlin, Berlin	3,426,354	52.524 / 13.411
2	Hamburg, Hamburg	1,739,117	53.575 / 10.015
3	Munich, Bavaria		
4	Cologne, North Rhine-		
5	Frankfurt am Main, He		
6	Essen, North Rhine-We		
7	Stuttgart, Baden-Würt		
8	Dortmund, North Rhin		
9	Düsseldorf, North Rhin		
10	Bremen, Bremen		

	WEBSITE	HUB	AIRLINE CODE
	www.aeroflot.ru/eng/	Moscow	SU
	www.airfrance.fr	Paris	AF
		Tokyo	NH
		Seoul	OZ
		Hong Kong	CX

150 INTERNATIONAL AFFAIRS		2016	2017	2018	2019	2020
8.	End Funding for the United Nations Development Program (UNDP)	81	81	82	83	85
9.	End Funding for the U.N. Intergovernmental Panel on Climate Change (IPCC)	10	10	10	10	11
10.	Eliminate the U.S. Trade and Development Agency (USTDA)					
11.	Reform Food Aid Programs					
12.	Eliminate Export-Import Bank					
13.	Eliminate the Overseas Private Investment Corporation (OPIC)					
14.	Eliminate Funding for the United Nations Population Fund (UNFPA)					
SUBTOTALS						

Most Requested Songs

Published December 28, 2011

Here are the most requested songs of the past year:

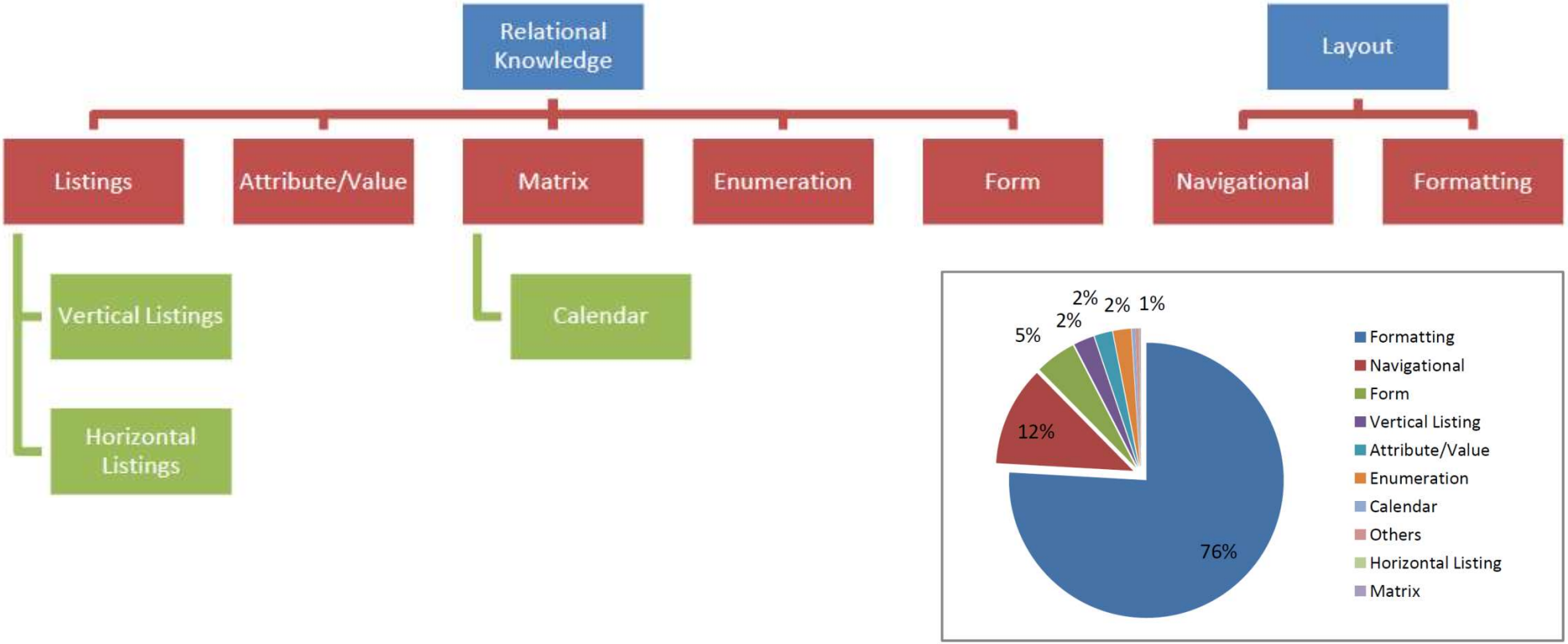
1	Black Eyed Peas	I Gotta Feeling
2	Journey	Don't Stop Believin'
3	Lady Gaga Feat. Colby O'donis	Just Dance
4	AC/DC	You Shook Me All Night Long
5	Cupid	Cupid Shuffle
6	Bon Jovi	Livin' On A Prayer
7	Beyonce	Single Ladies (Put A Ring On It)
8	Diamond, Neil	Sweet Caroline (Good Times Never Seemed So Good)
9	Morrison, Van	Brown Eyed Girl
10	Def Leppard	Pour Some Sugar On Me
11	B-52's	Love Shack
12	Lmfao Feat. Lauren Bennett And Goon Rock	Party Rock Anthem
13	Jackson, Michael	Billie Jean
14	DJ Casper	Cha Cha Slide
15	Usher Feat. Will I Am	Omg

Contestant	Age		
Kelly Louise Maguire	24	1.7	
Aquelle Plakaris	24	1.7	Nassau
Jessica van Moorlegh	18	1.70 m (5 ft 7 in)	Evergem
Yovana O'Brien	19	1.80 m (5 ft 11 in)	Santa Cruz

Types of Web Tables

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

Cafarella (2008)



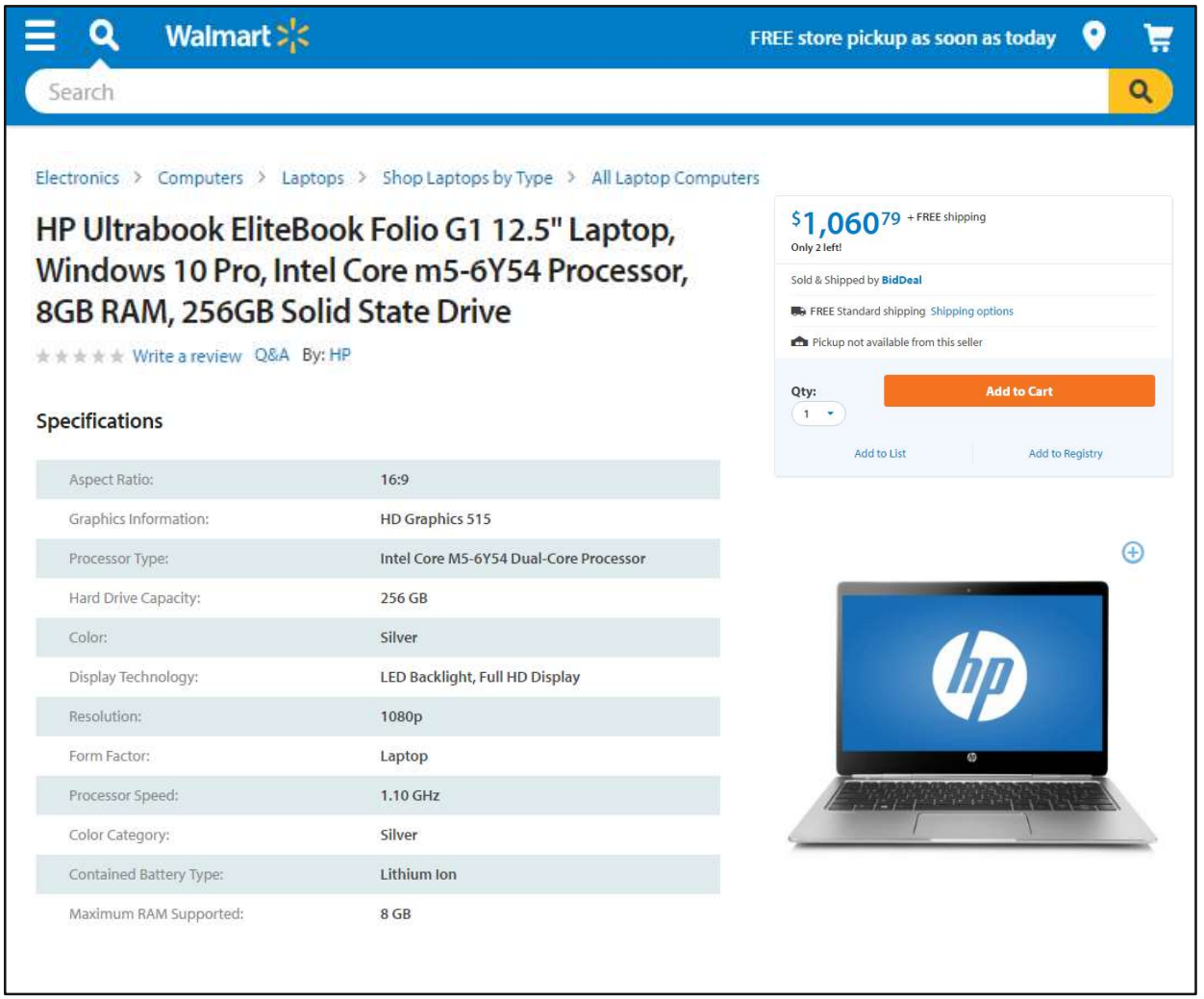
Cafarella, et al.: **WebTables: Exploring the Power of Tables on the Web.** VLDB 2008.
Crestan, Pantel: **Web-Scale Table Census and Classification.** WSDM 2011.

Attribute/Value Table together with schema.org Annotations

s:breadcrumb

s:name

Attribute/Value HTML Table



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/

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
counrty	340,000
isbn	99,000
area	95,000
population	86,000

28,000,000 different attribute labels

Subject Attribute Values

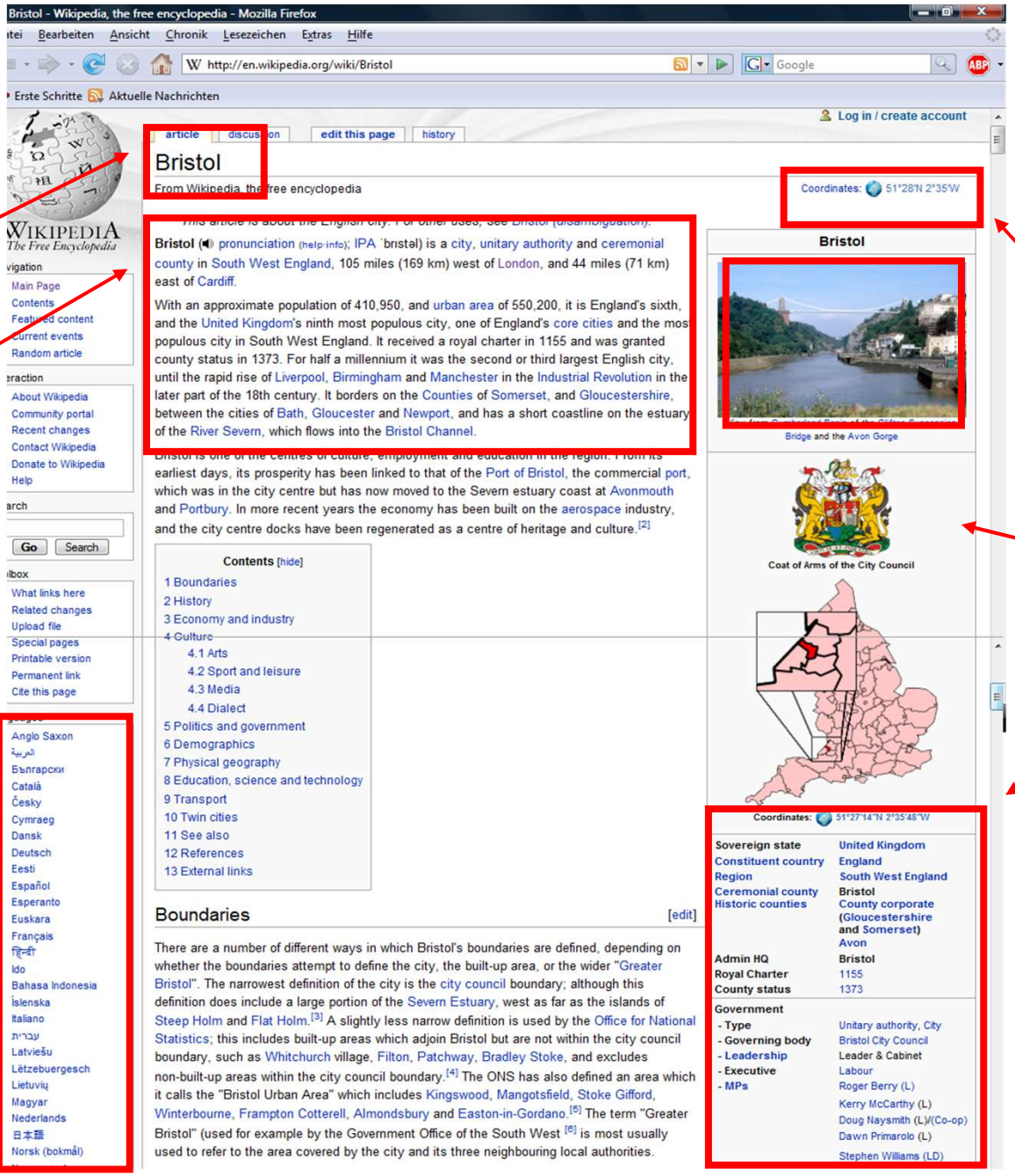
Value	#Rows
usa	135,000
germany	91,000
greece	42,000
new york	59,000
london	37,000
athens	11,000
david beckham	3,000
ronaldinho	1,200
oliver kahn	710
twist shout	2,000
yellow submarine	1,400

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



Title

Description

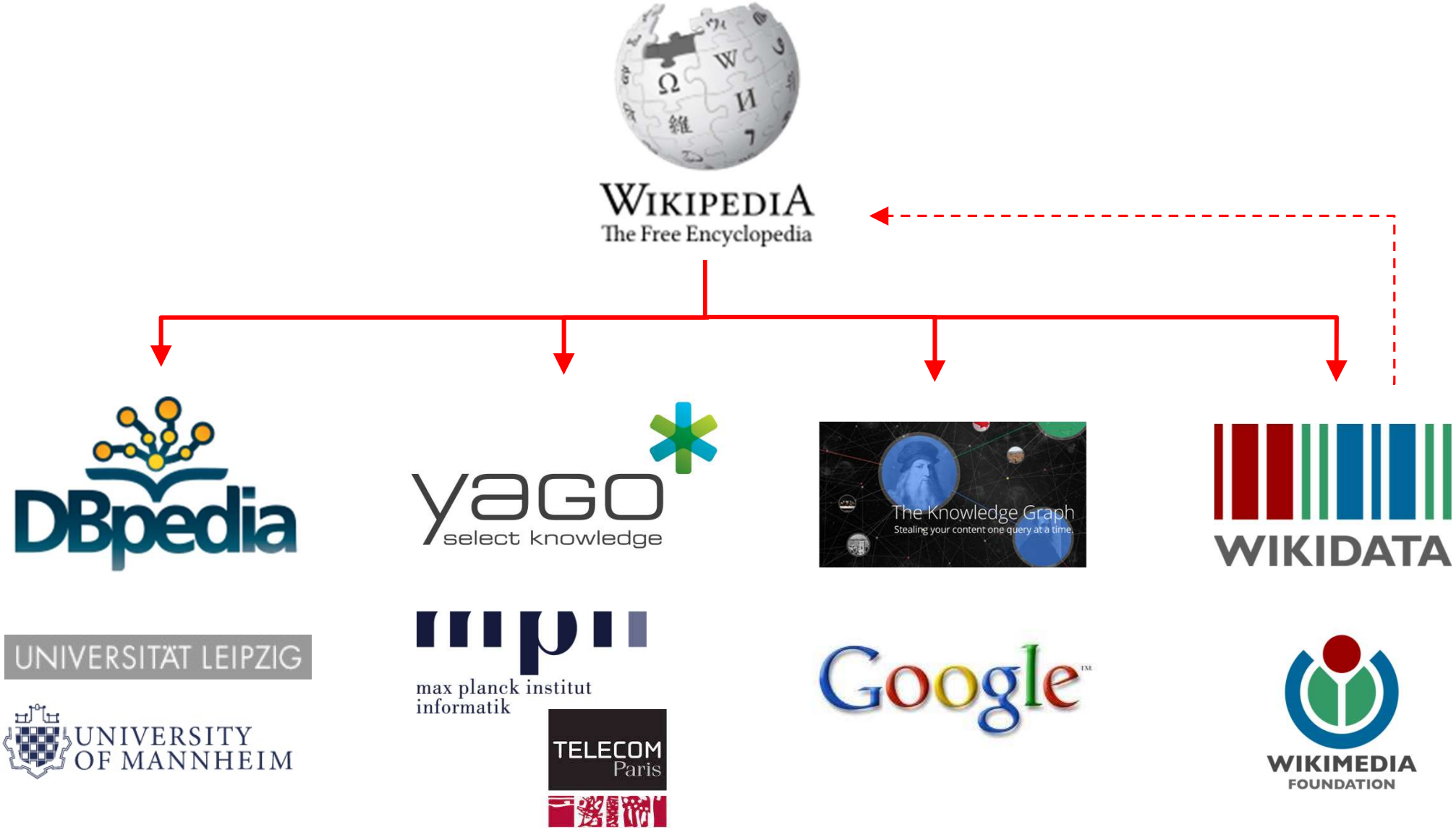
Cross Language Links

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

About neofonie | About DBpedia | Help

DBpedia
search powered by neofonie

enter search terms... Search

Deutschland Land der Ideen

First | Previous | Next | Last

item type
start typing...
Skyscraper (12)
Place (12)
Building (12)
more

location
start typing...
Hong Kong (12)
China (3)
Sham Tseng (1)
more

building started in year
start typing...
from... to... >
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.

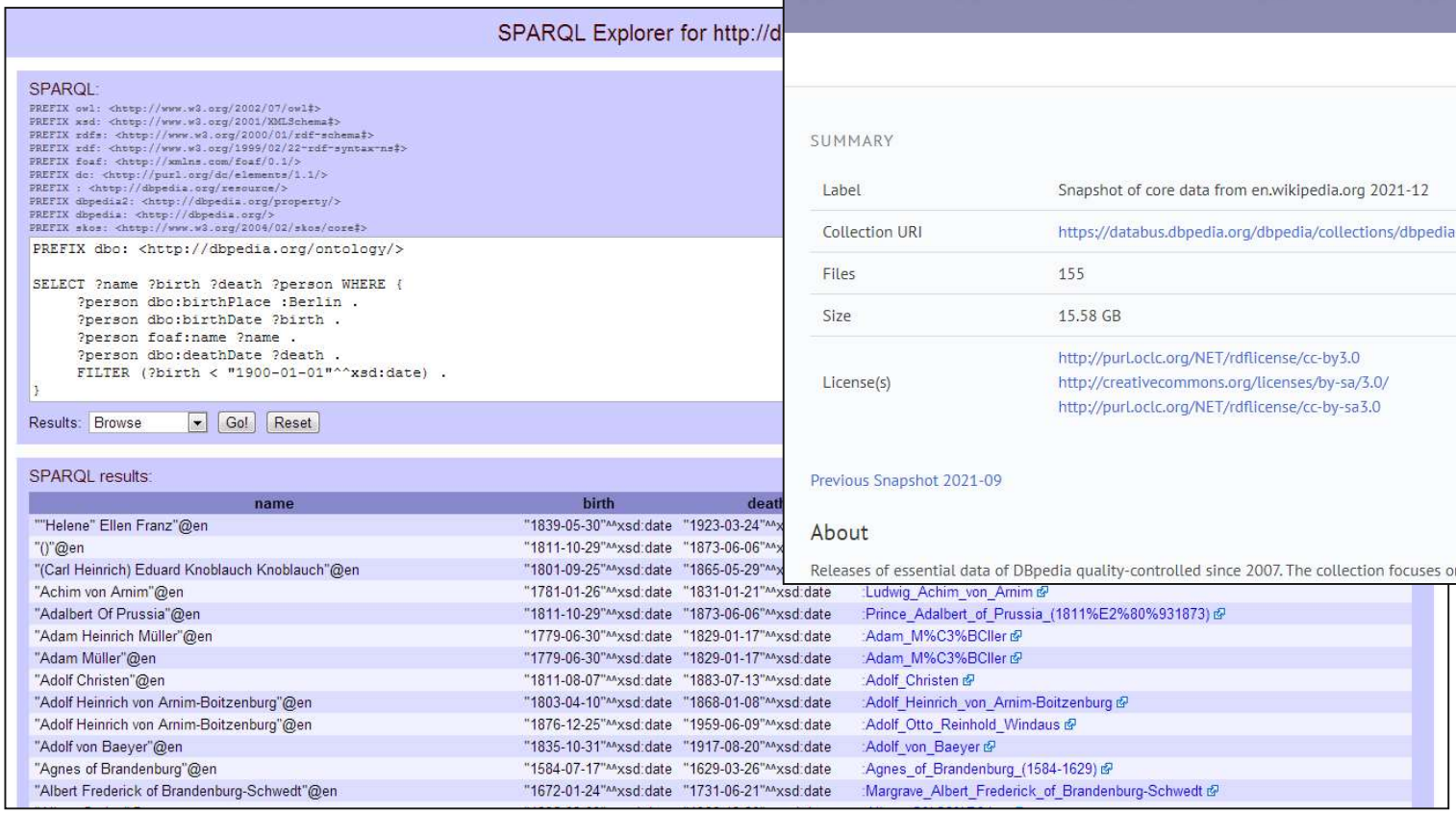
The Harbourside

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

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

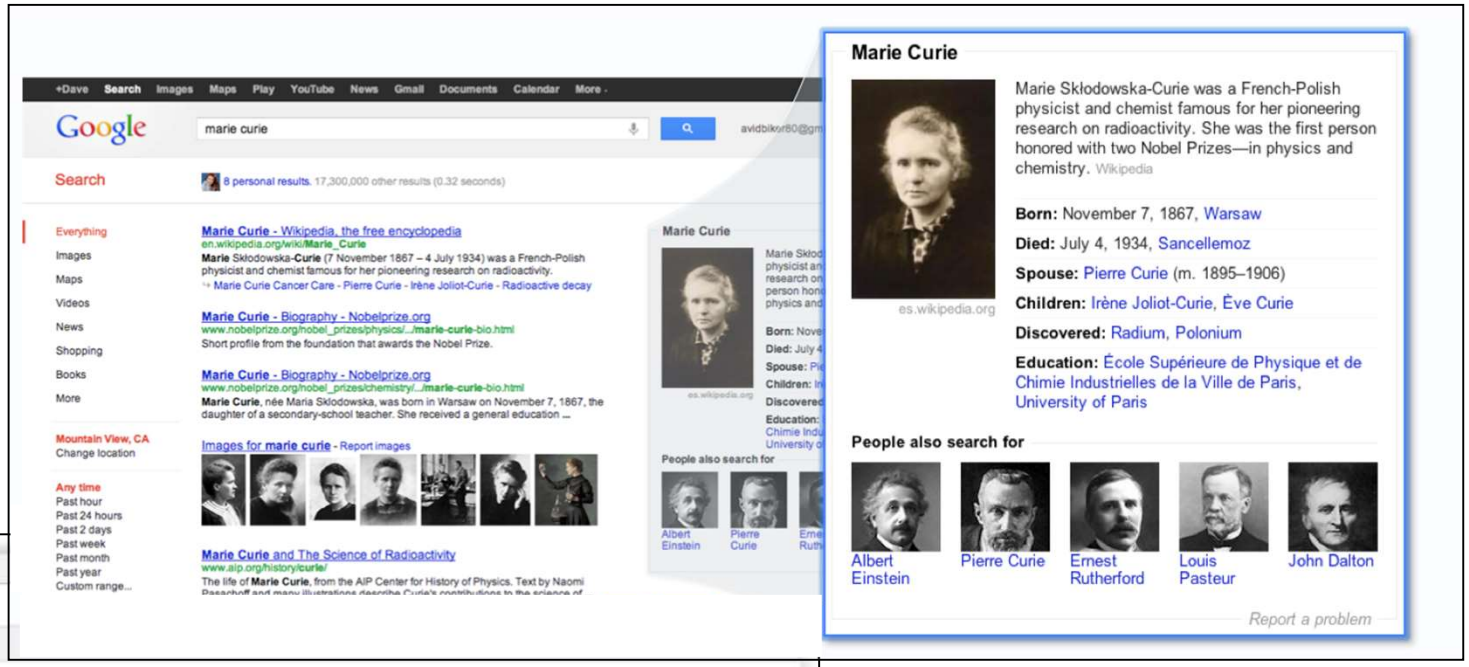
Data Sources used to Build Knowledge Graphs

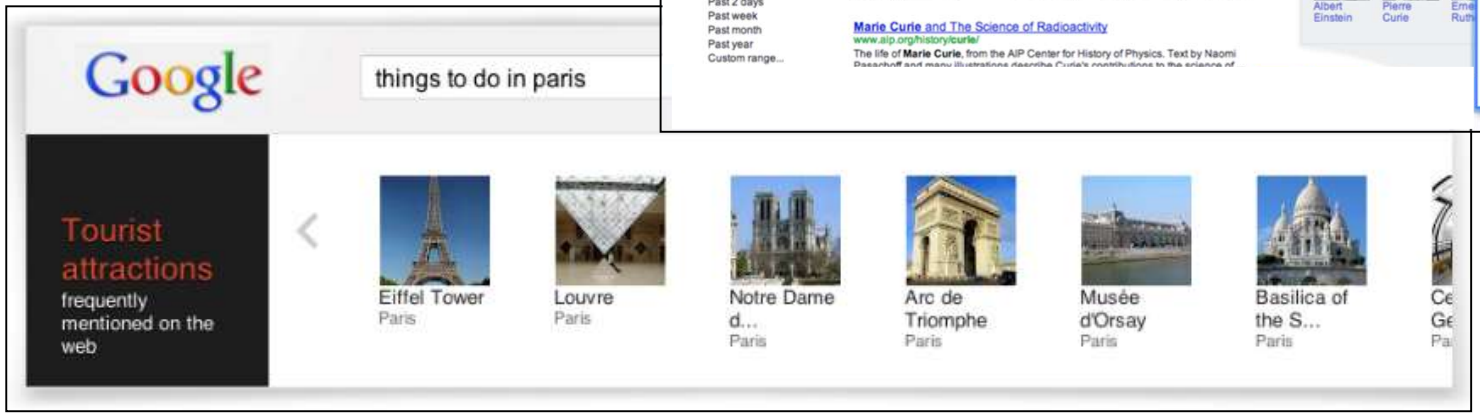
1. Wikipedia
   - infoboxes, category system, information extraction from text
2. Open license sources
   - e.g. CIA World Factbook, MusicBrainz, …
3. Commercial third-party data
   - e.g. IMDB, company listings, …
4. schema.org annotations in web pages
   - e.g. contact information for companies
   - e.g. logos of companies

Lots of effort is spend 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





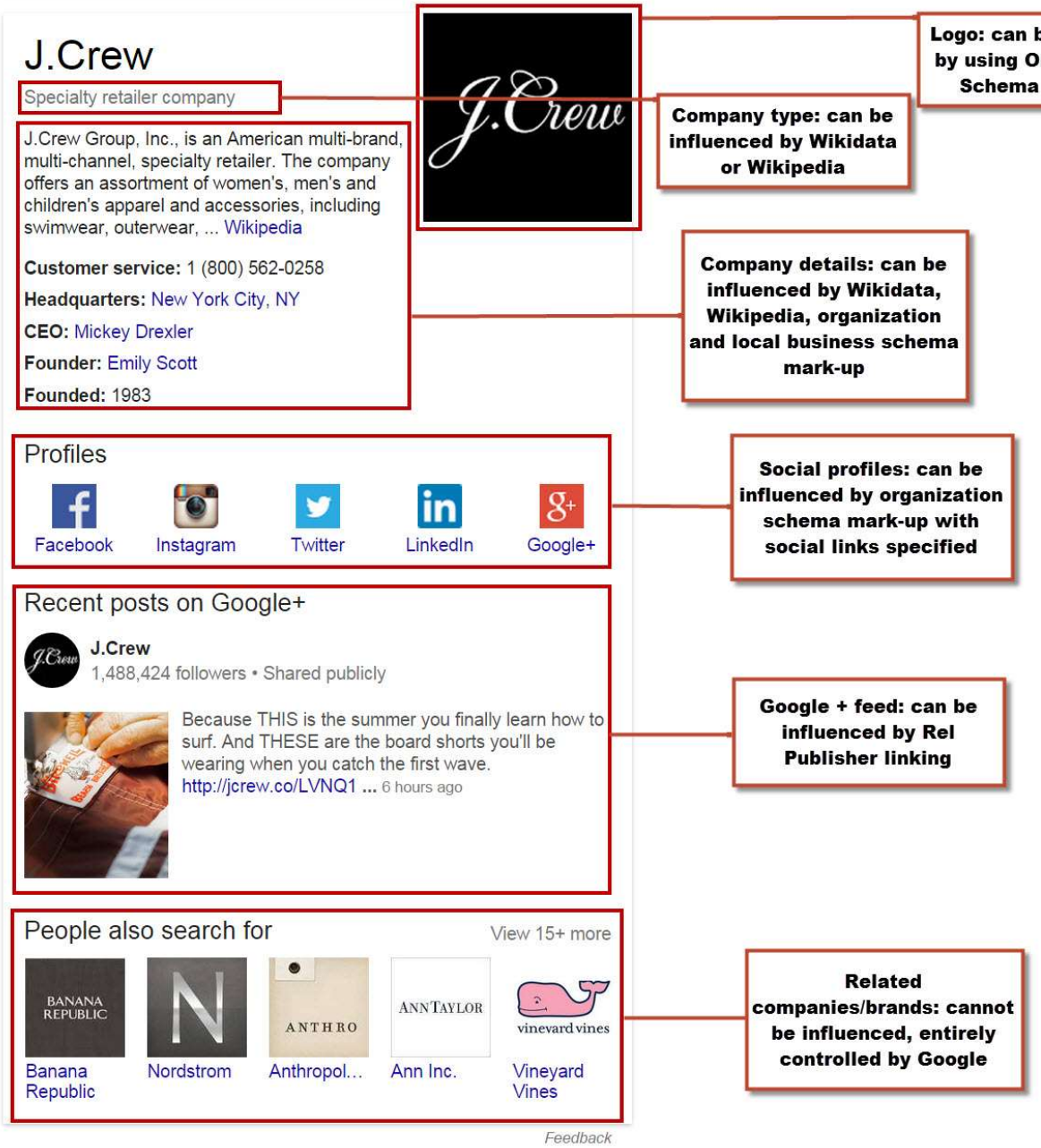
Behind-the-Scenes Applications of KGs

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

- Google
  - uses its knowledge graph to identity 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?



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

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

Google + feed: can be influenced by Rel Publisher linking

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

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

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.