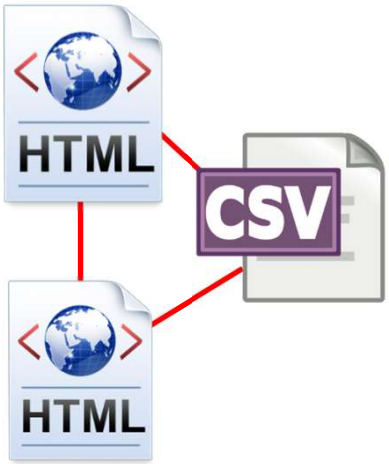
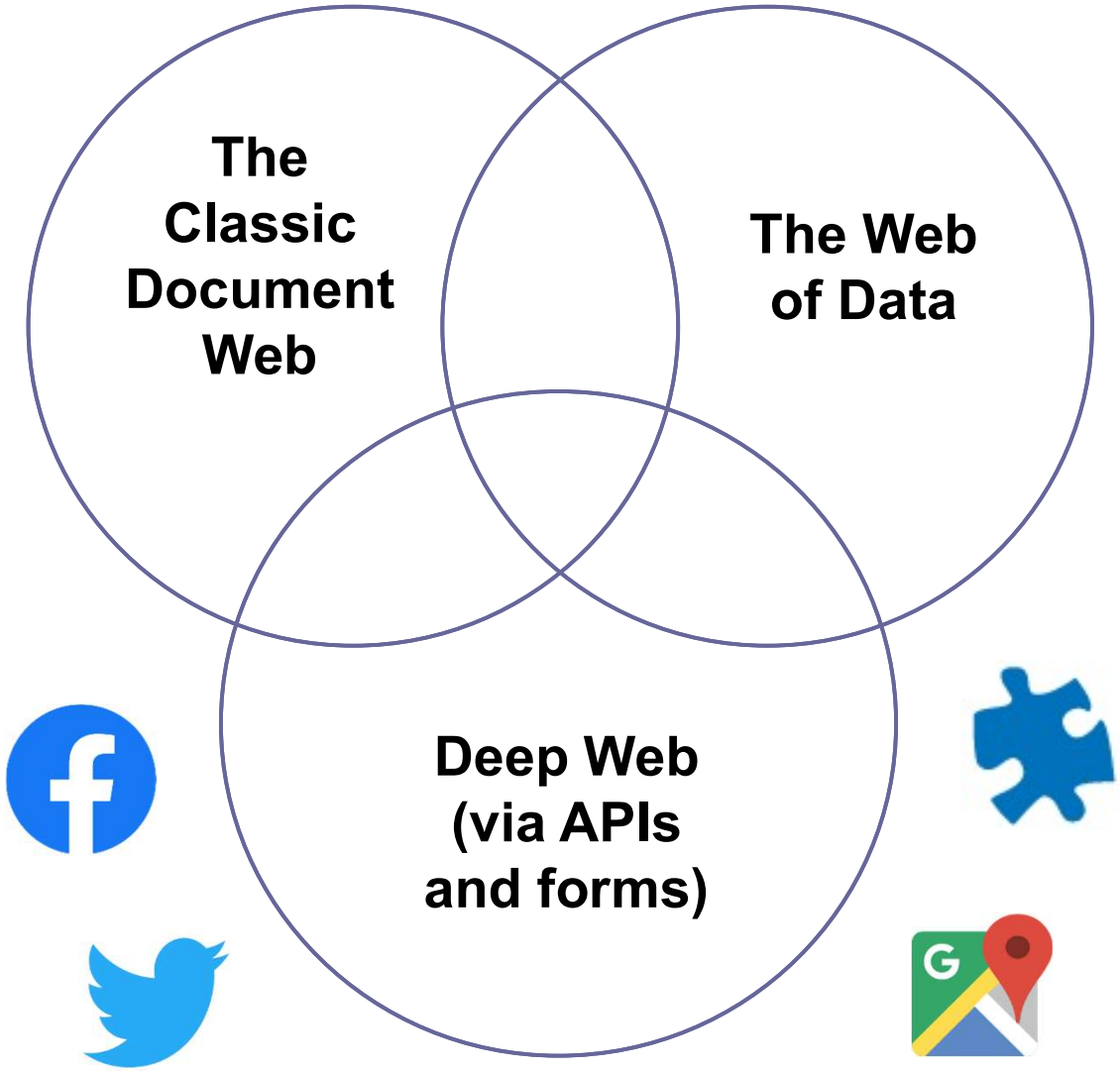
UNIVERSITY OF MANNHEIM

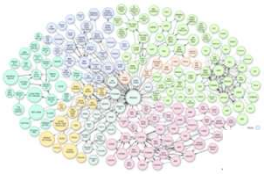
Web Data Integration

Types of Structured Data on the Web

Topology of the Web Today





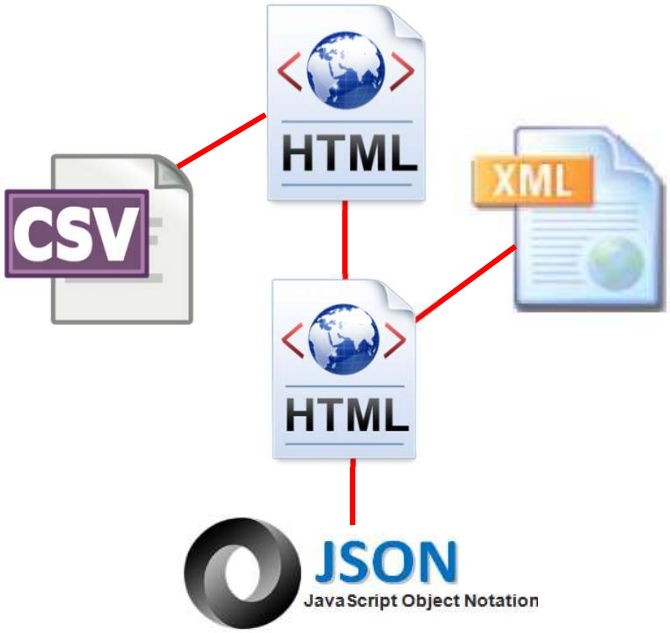


Outline

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

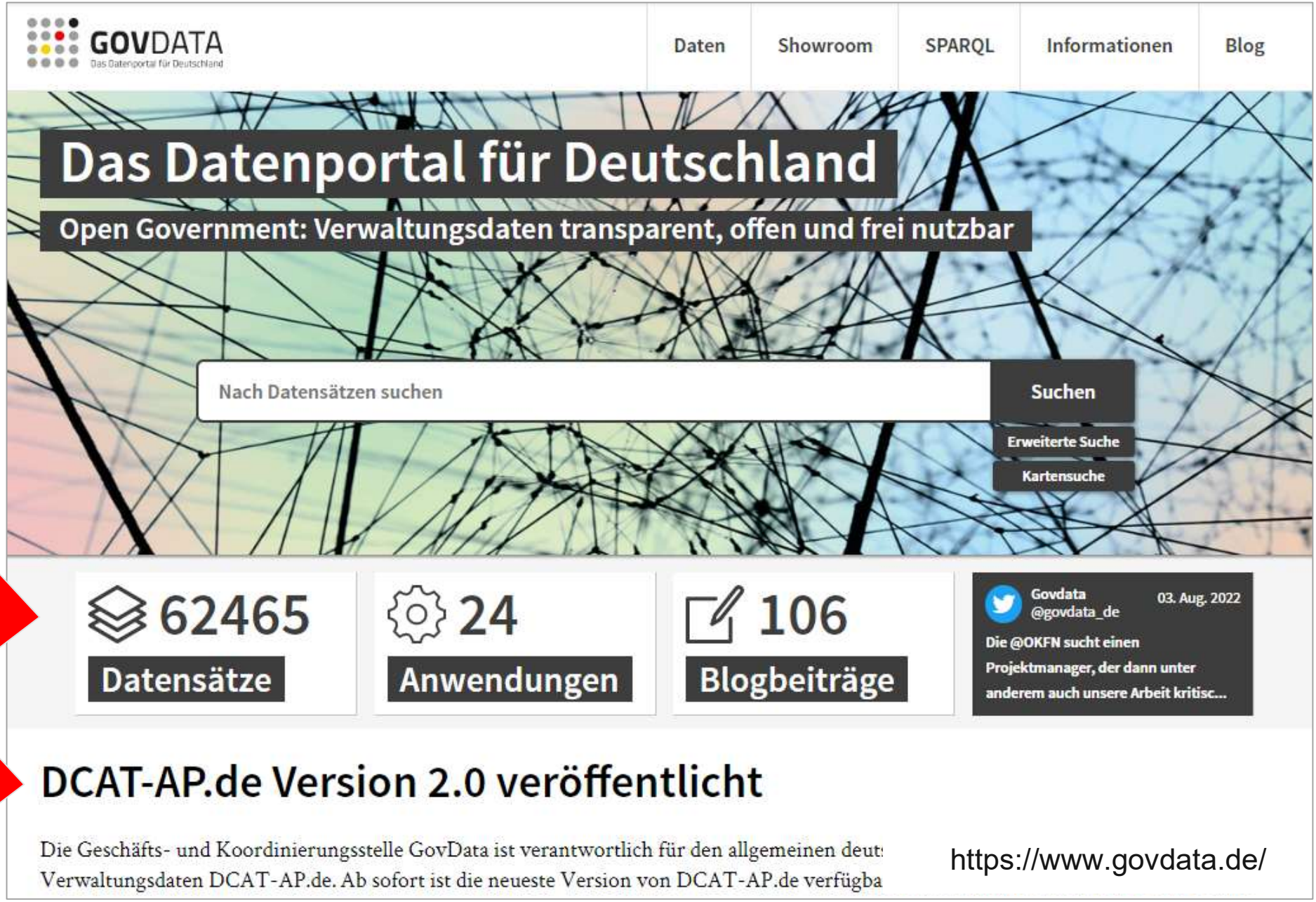


- Data Portals and Data Marketplaces
  - collect and host datasets
  - collect and generate metadata describing the datasets
  - provide for data search and exploration
  - provide free or payment-based access to data

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: datarade, AWS Data Products

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

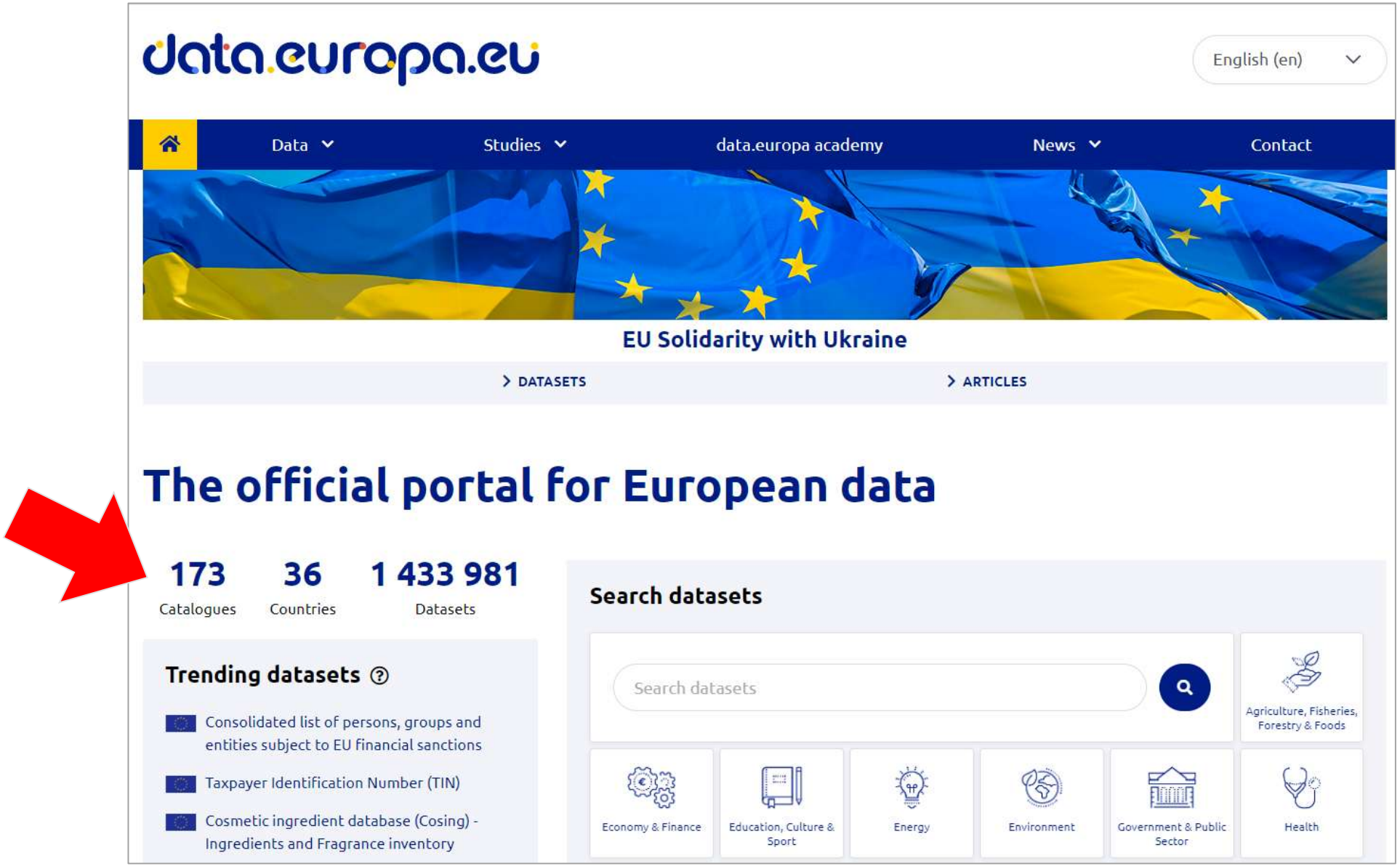
Dateiformate:
PDF

Letzte Änderung:
17.08.2022

Zeitraum:
-

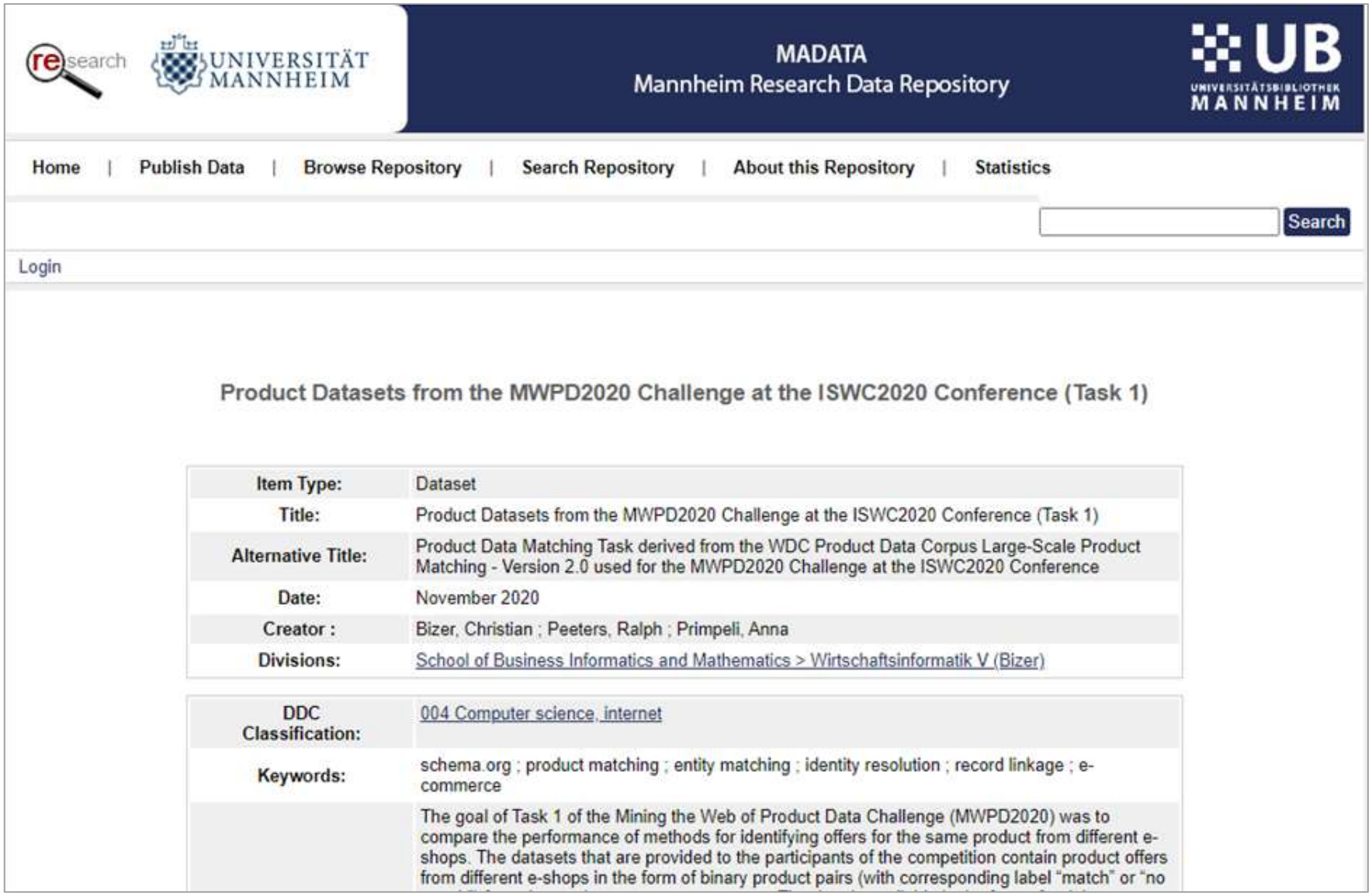
zum Datensatz

Example: Portal aggregating Metadata from other Portals



https://data.europa.eu/en

Example: Institutional Research Data Repository



MADATA
Mannheim Research Data Repository

Home | Publish Data | Browse Repository | Search Repository | About this Repository | Statistics

Search

Login

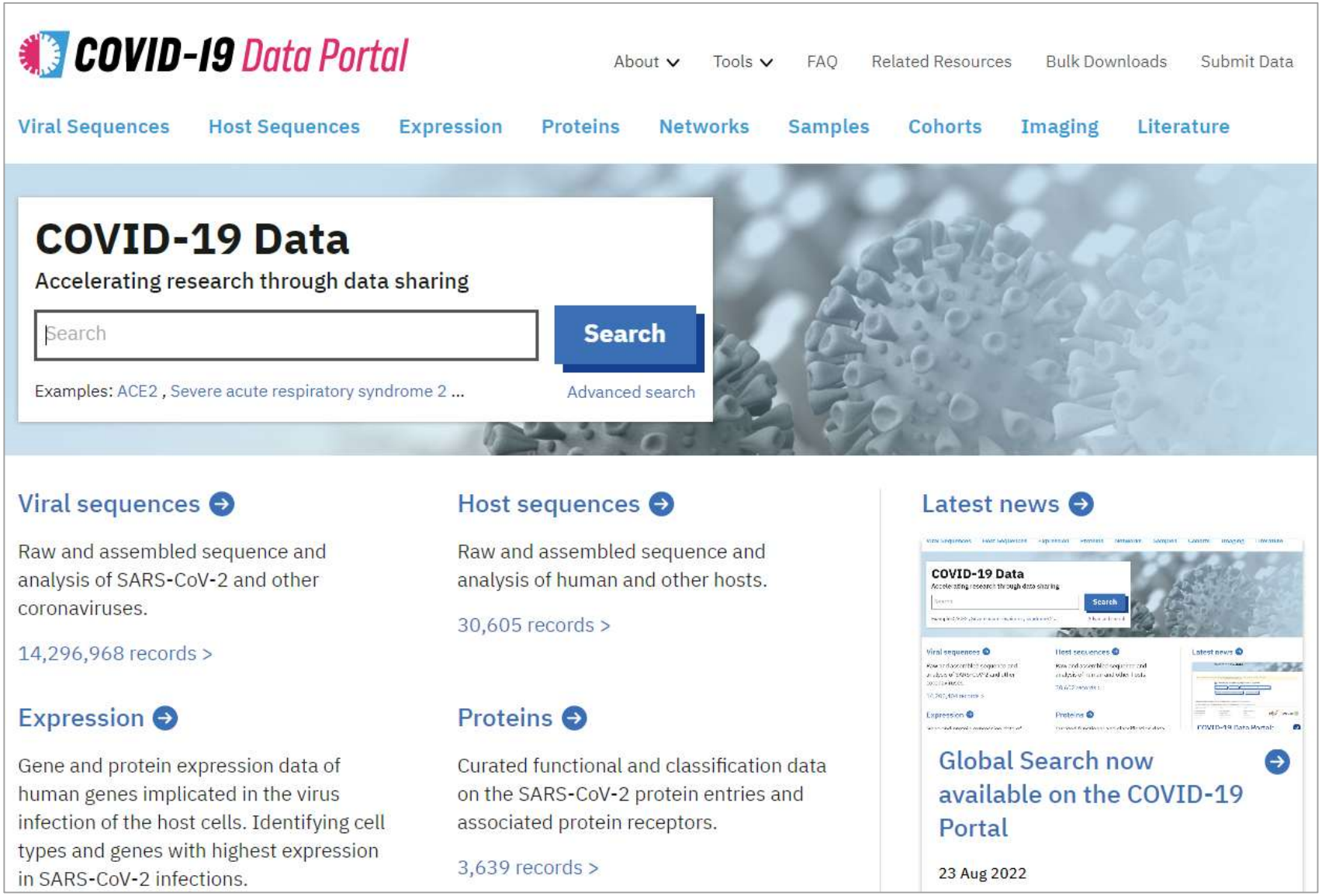
Product Datasets from the MWPD2020 Challenge at the ISWC2020 Conference (Task 1)

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



EMBL Heidelberg

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

Example: Cross-Topic Research Data Portal

zenodo

Search records...

Communities | My dashboard | Log in | Sign up

Published October 26, 2022 | Version 0.2.0

Dataset | Open

Data for replication of the publication: Probabilistic leak localization in water distribution networks using a hybrid data-driven and model-based approach

Ganjour Mazaev[1]; Michael Weyns[1]; Filip Vancoillie[2]; Guido Vaes[3]; Femke Ongenae[1]; Sofie Van Hoecke[1]

Show affiliations

20 to 30% of drinking water produced is lost due to leaks in water distribution pipes. In times of water scarcity, losing so much treated water comes at a significant cost, both environmentally and economically. In this paper, we propose a hybrid leak localization approach combining both model-based and data-driven modeling. Pressure heads of leak scenarios are simulated using a hydraulic model, and then used to train a machine-learning based leak localization model. A key element of our approach is that discrepancies between simulated and measured pressures are accounted for using a dynamically calculated bias correction, based on historical pressure measurements. Data of in-field leak experiments in operational water distribution networks were produced to evaluate our approach on realistic test data. Two problematic settings for leak localization were examined. In the first setting, an uncalibrated hydraulic model was used. In the second setting, an extended version of the water distribution network was considered, where large parts of the network were insensitive to leaks. Our results show that the leak localization model is able to reduce the leak search region in parts of the network where leaks induce detectable drops in pressure. When this is not the case, the model still localizes the leak but is able to indicate a higher level of uncertainty with respect to its leak predictions.

Files

data.zip

data.zip

603 VIEWS | 115 DOWNLOADS

Show more details

Versions

Version 0.2.0 — Oct 26, 2022
10.5281/zenodo.7255403

Version 0.1.0 — Jun 13, 2022
10.5281/zenodo.6637751

View all 2 versions

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.6637750. This DOI represents all versions, and will always resolve to the latest one. Read more.

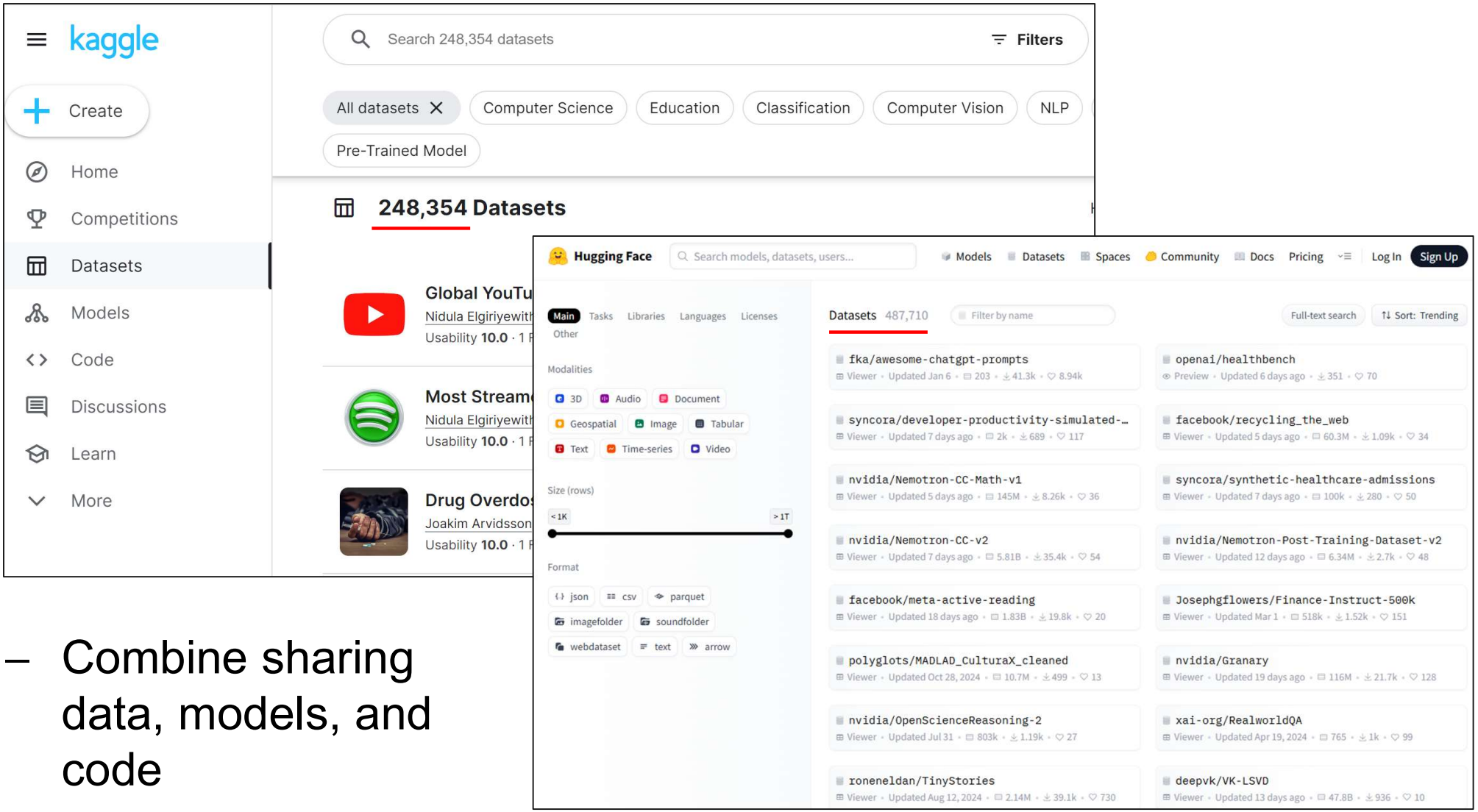
External resources

Indexed in

OpenAIRE

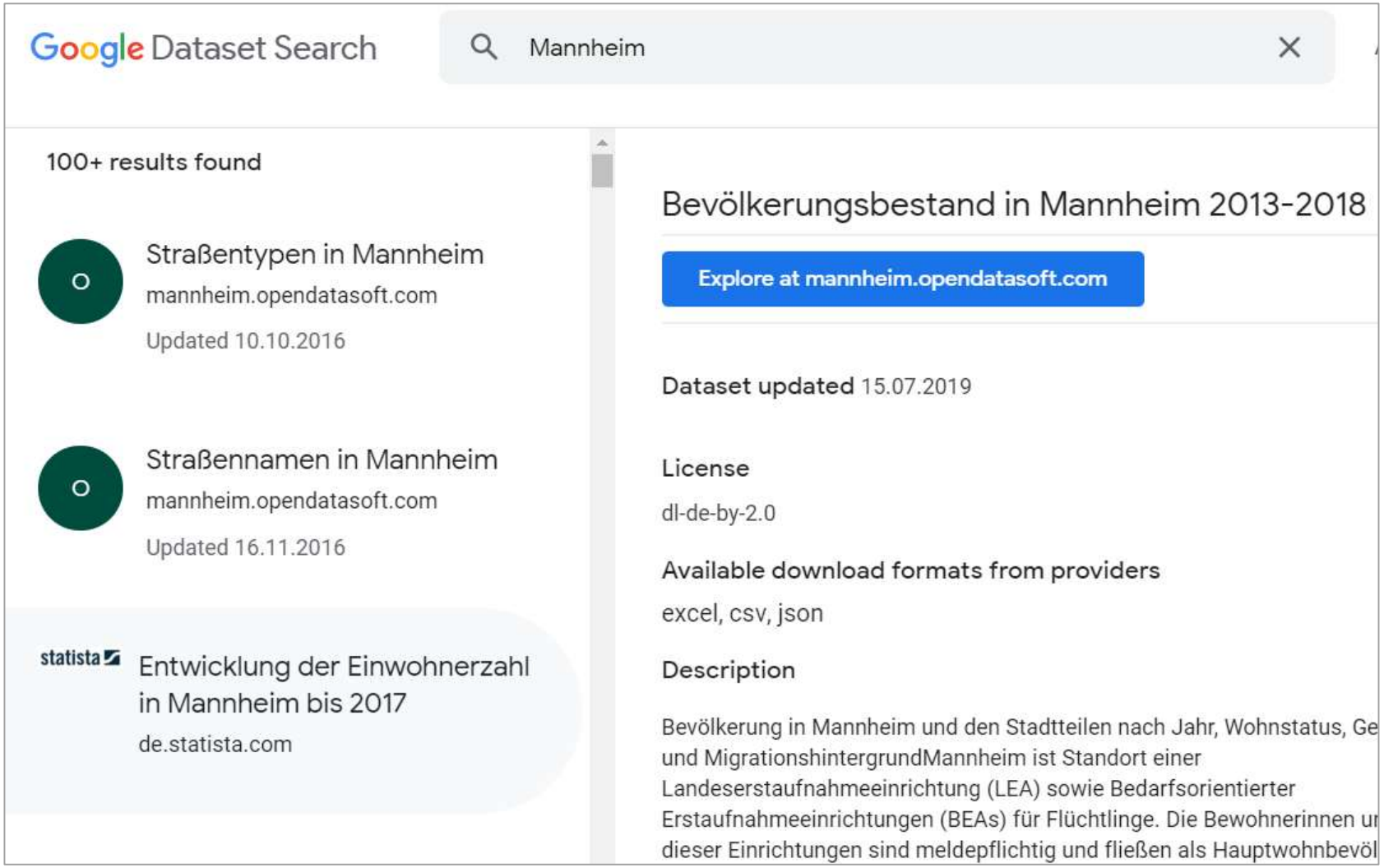
https://zenodo.org/

Example: Maschine Learning-oriented Data Portals



– Combine sharing data, models, and code

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

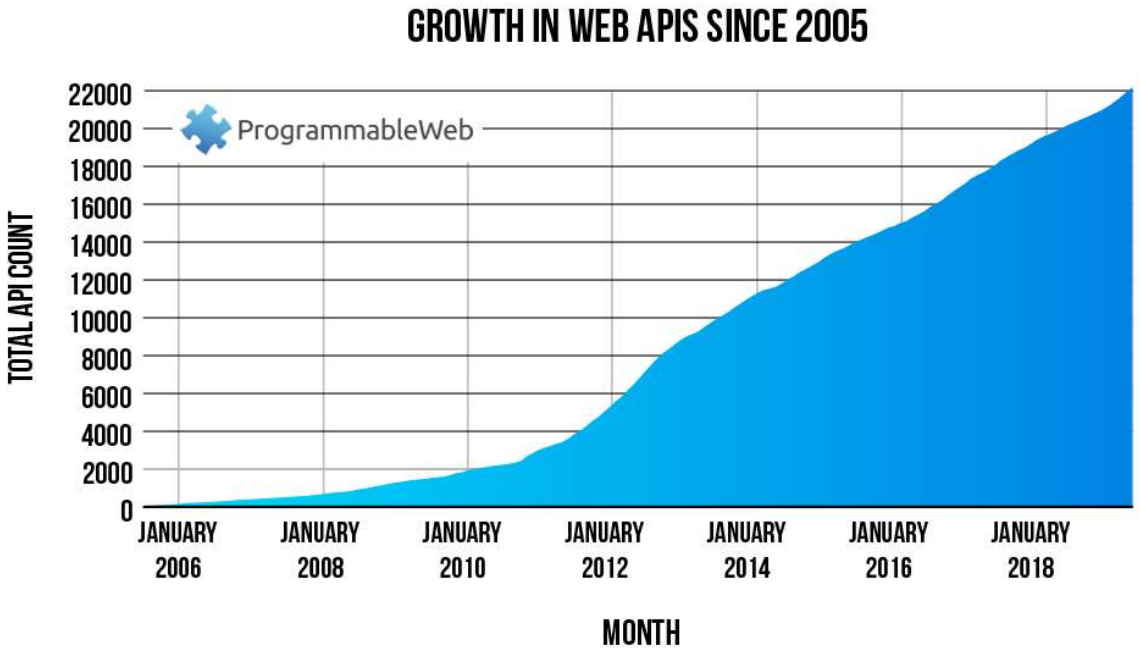
Example: Facebook Social Graph

- 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

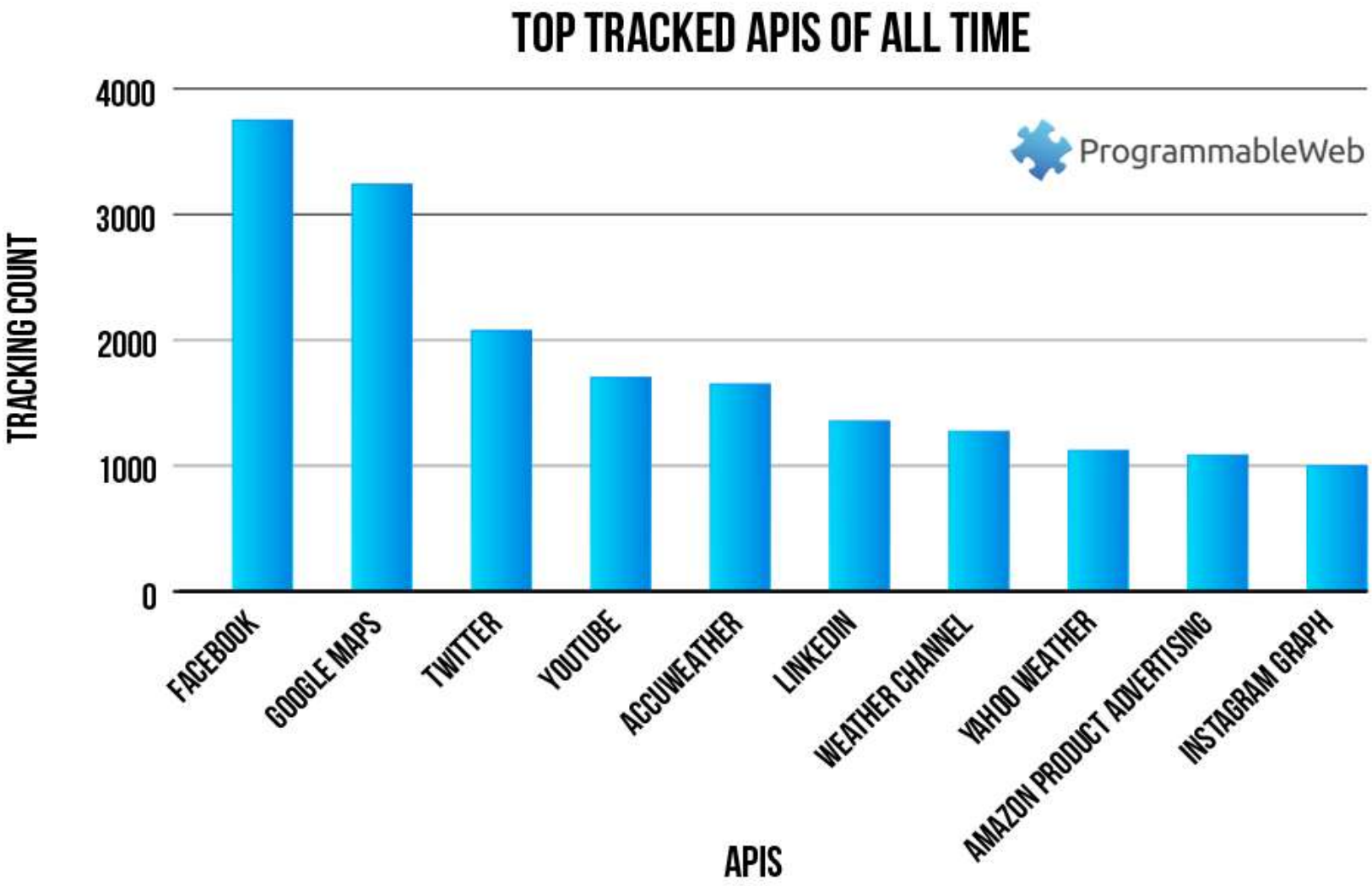
https://www.brandwatch.com/blog/facebook-statistics/
http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/

Web APIs

- Provide restricted access to the collected data
  - restricted to specific queries (canned queries)
  - restricted by number of queries / number of results
  - usually require registration
- ProgrammableWeb API Catalog
  - listed over 24,000 Web APIs (offline since 2022)
- Alternative Catalogs
  - https://publicapis.dev/
  - https://github.com/public-api-lists



Most Popular APIs



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

Resource-oriented vs. Function-oriented APIs

	Resource-oriented APIs	Function-oriented APIs
Types	**REST**	**JSON-RPC, SOAP**
Architectural paradigm	Retrieve and manipulate **resources** which are identified by HTTP URIs (e.g., api.example.com/users/123)	Call **remote procedures** (e.g., "method": "getUser")
Operations	HTTP verbs GET, POST, PUT, DELETE	Arbitrary functions having parameters
Response format	Typically JSON	JSON or XML

REST Request:

```
GET /users/123 HTTP/1.1
Host: api.example.com
Accept: application/json
```

JSON-RPC Request:

```
{ "jsonrpc": "2.0",
  "method": "getUser",
  "params": { "id": 123 }}
```

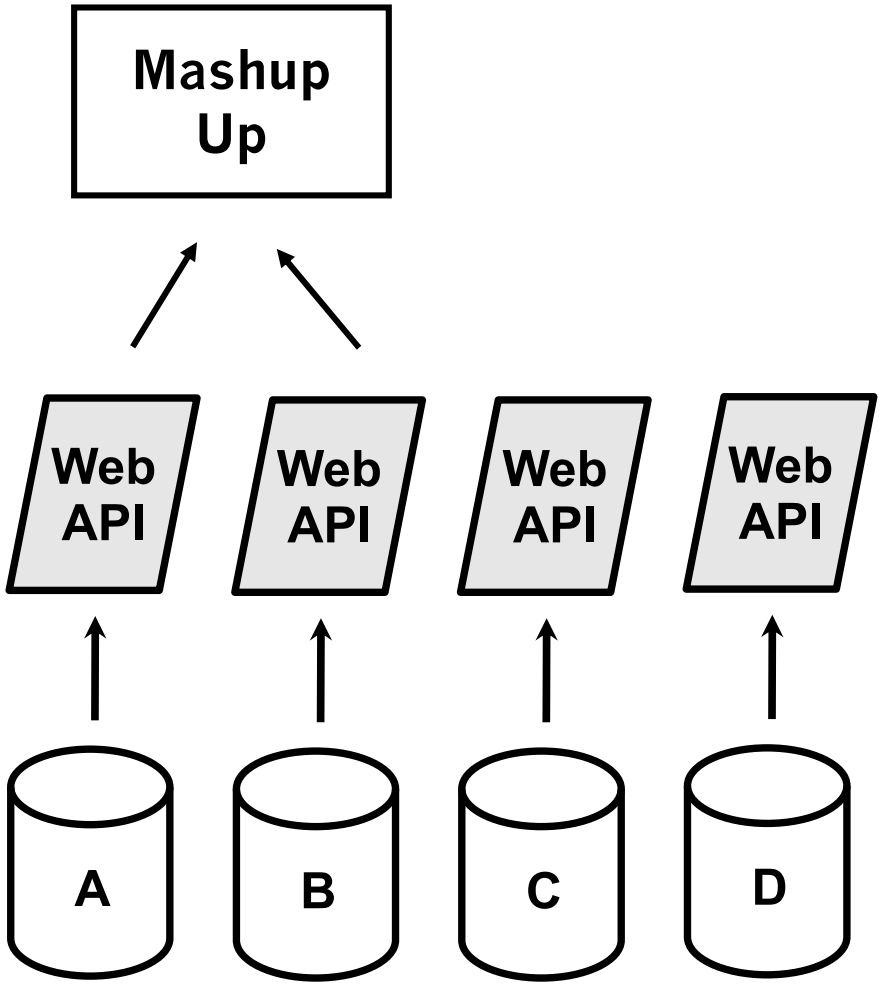
REST Response:

```
{ "id": 123,
  "name": "Alice",
  "email": "alice@ex.com" }
```

JSON-RPC Response:

```
{ "jsonrpc": "2.0",
  "result": {
    "id": 123,
    "name": "Alice",
    "email":"alice@ex.com" }}
```

Mashups are based on a fixed set of data sources



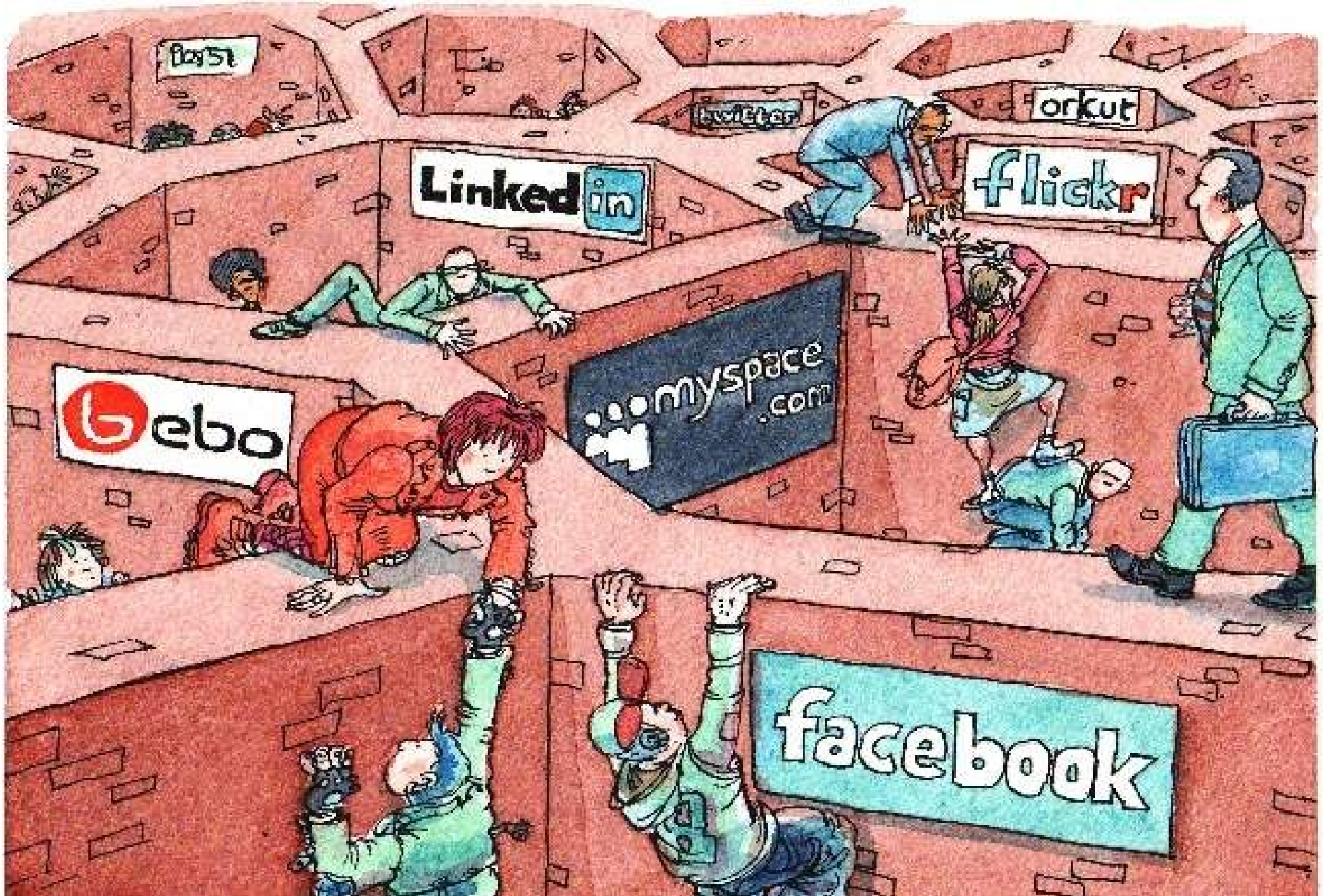
Web APIs expose proprietary interfaces and usually require registration

→ Not index-able by generic web crawlers

→ No automatic discovery of additional data sources

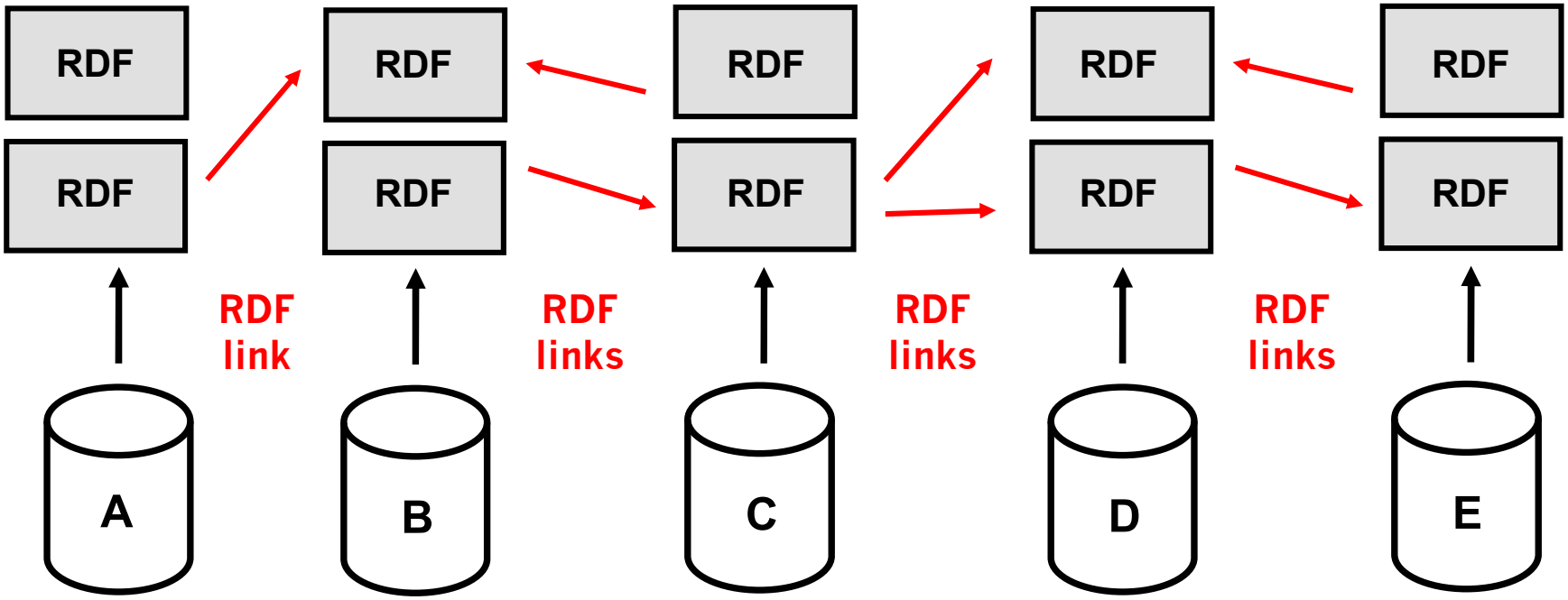
→ No single global data space

Web APIs slice the Web into Data Silos

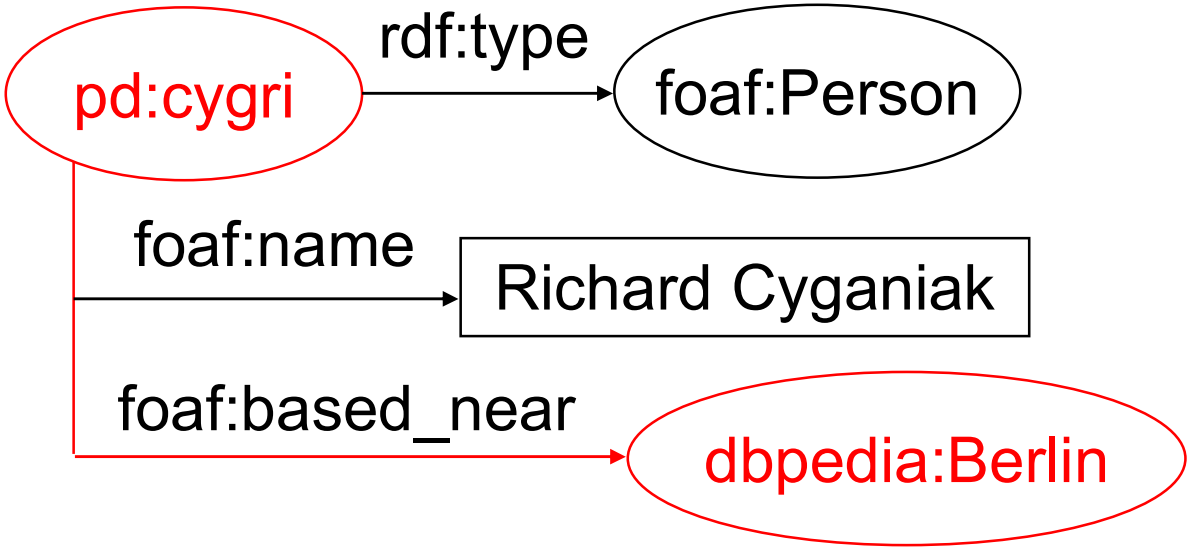


3. Alternative Approach: Linked Data

- 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

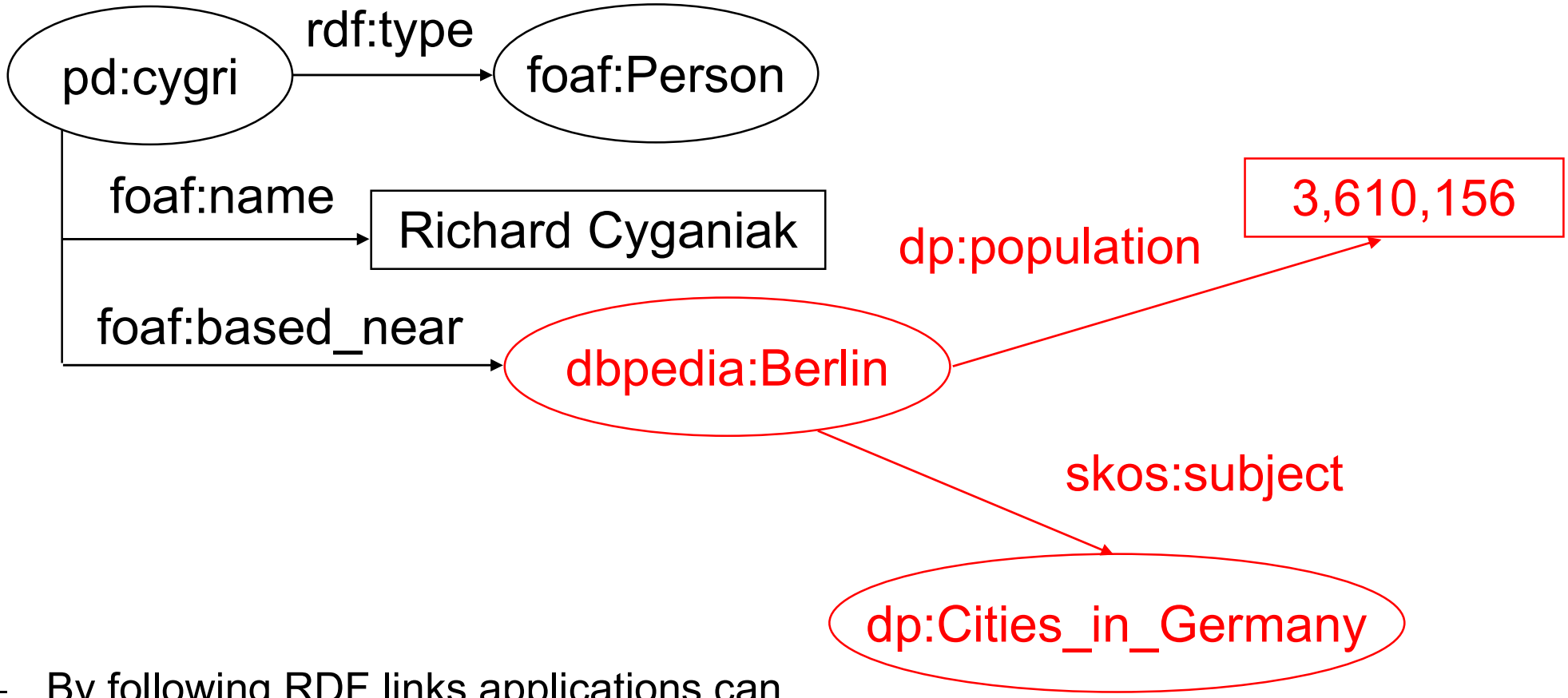


HTTP URIs take the role of global primary keys.

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
  - Principles Find1, Access1, Interop1, Interop2, Interop3

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

SIGMA 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

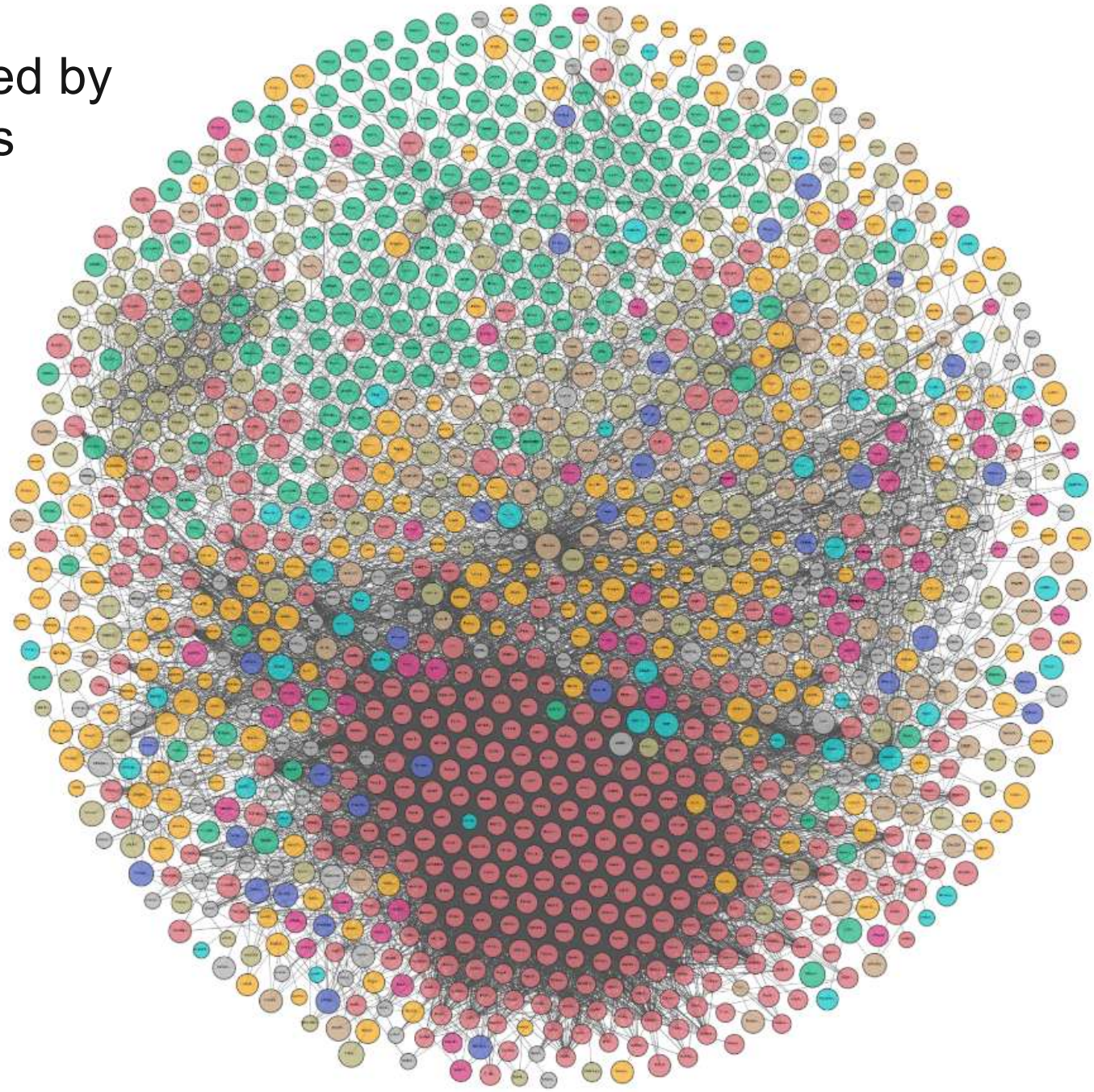
8 Flickr: Chris Bizer's Ph
BOSS http://flickr.com/pho

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

1,346 datasets connected by
16,174 sets of RDF links
(as of August 2024)



Legend

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

https://lod-cloud.net/

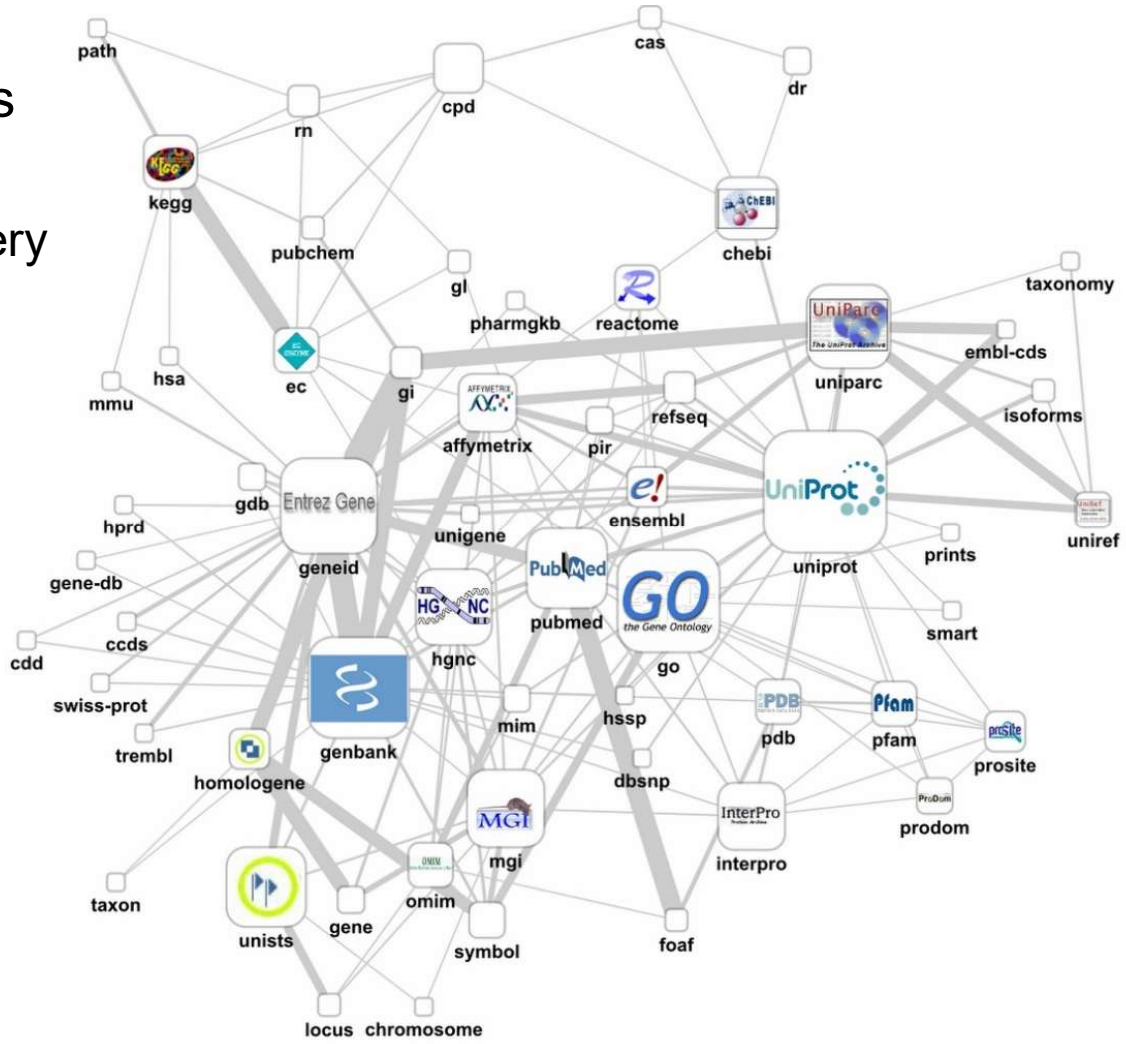
Uptake in the Libraries Community

- 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)
  - DBLP Computer Science Bibliography
  - 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.

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



BIO2RDF

Open PHACTS
Open Pharmacological Space

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

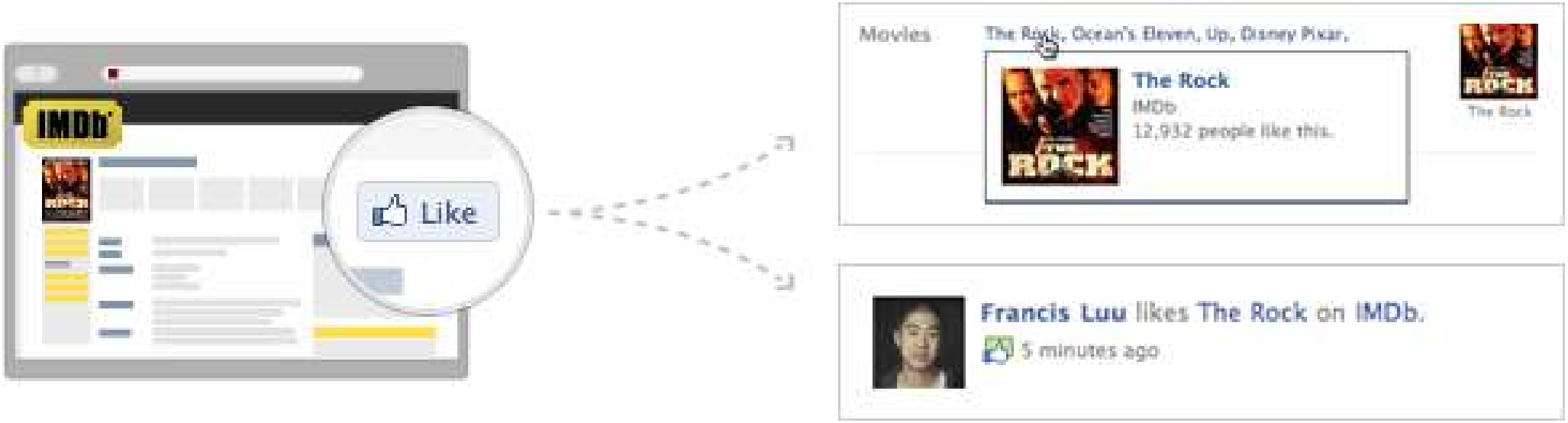
RDFa

- 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

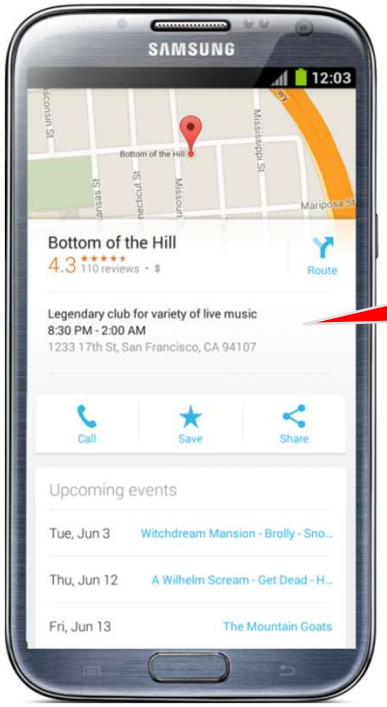
- ask site owners since 2011 to annotate data for enriching search results
- 675 Types: Event, Local Business, Product, Review, Job Offer
- Encoding: Microdata, RDFa, JSON-LD

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

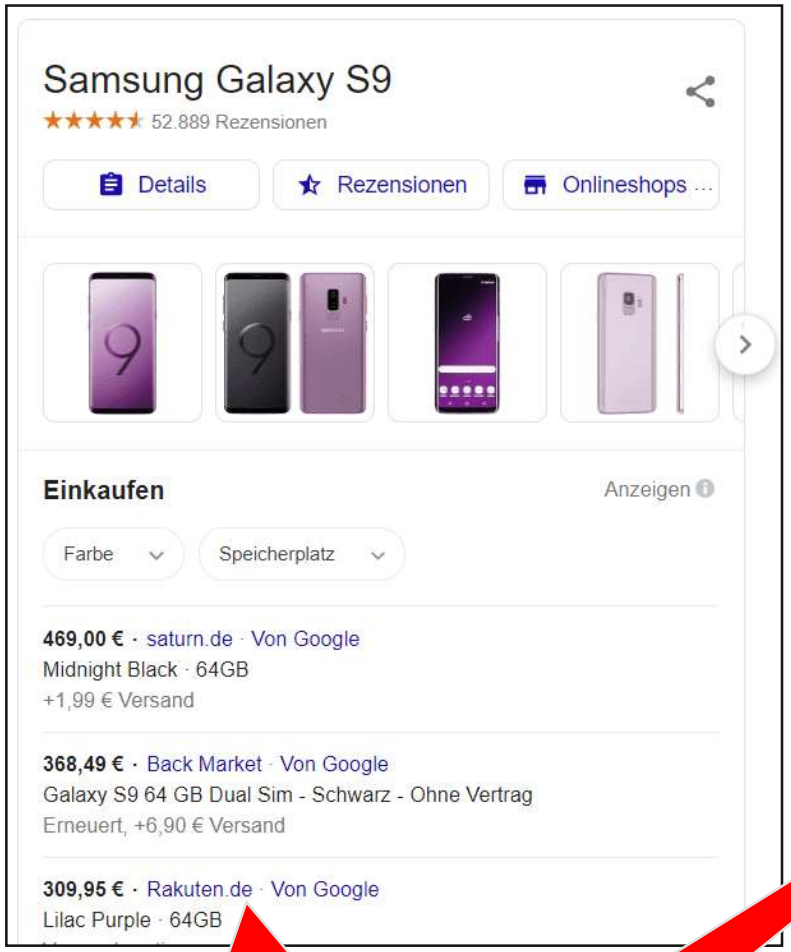
Record labels: Fat Possum Records, Nonesuch Records, V2 Records, Alive Naturalsound Records

Awards: Grammy Award for Best Rock Album, more

Upcoming events

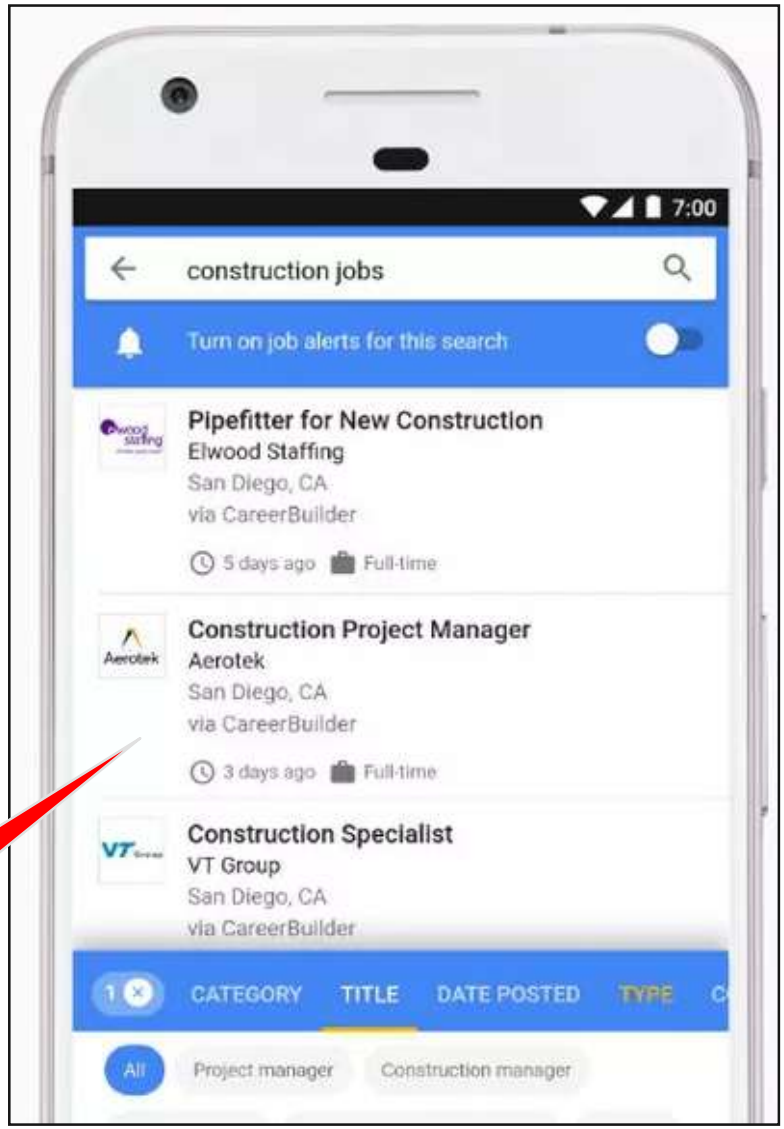
Jun 20 Fri	The Black Keys Neuhausen ob Eck (near you)
May 16 Fri	The Black Keys Gulf Shores, AL
Jun 22 Sun	The Black Keys Scheeßel

Usage of Schema.org Data @ Google



Product offers

Job offers



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

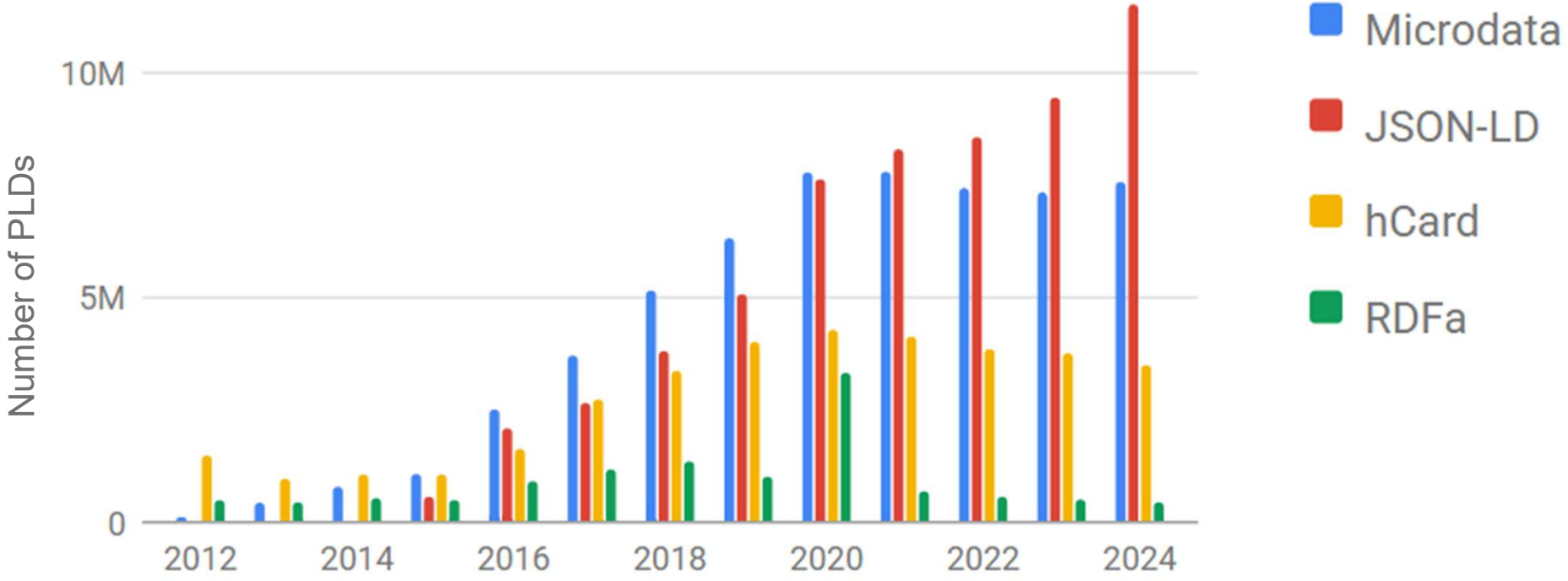
The Web Data Commons Project

- 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
  - 2024 CC Corpus: 2.4 billion HTML pages → 73.9 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 ➔ 1000 Euro
- http://webdatacommons.org/structureddata/

Overall Adoption 2024

1.3 billion HTML pages out of the 2.4 billion pages provide semantic annotations (51%).

16.5 million pay-level-domains (PLDs) out of the 37 million PLDs (websites) provide semantic annotations (44%).



http://webdatacommons.org/structureddata/2024-12/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 %
…	…

New Samsung Galaxy S4 GT-19505 16GB 5.0 inches Android Smartphone with 2-Year Sprint Contract - White Frost

99.00 US$

The Galaxy S4 is among the earliest phones to feature a 1080p Full HD display. The various connectivity options on the Samsung include …

000214632623

Samsung

GT-19505

000214632623

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						

Contestant	Age		
Kelly Louise Maguire	24	1.7	
Aquelle Plakaris	24	1.78 m (5 ft 10 in)	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

Most Requested Songs

Published December 28, 2011

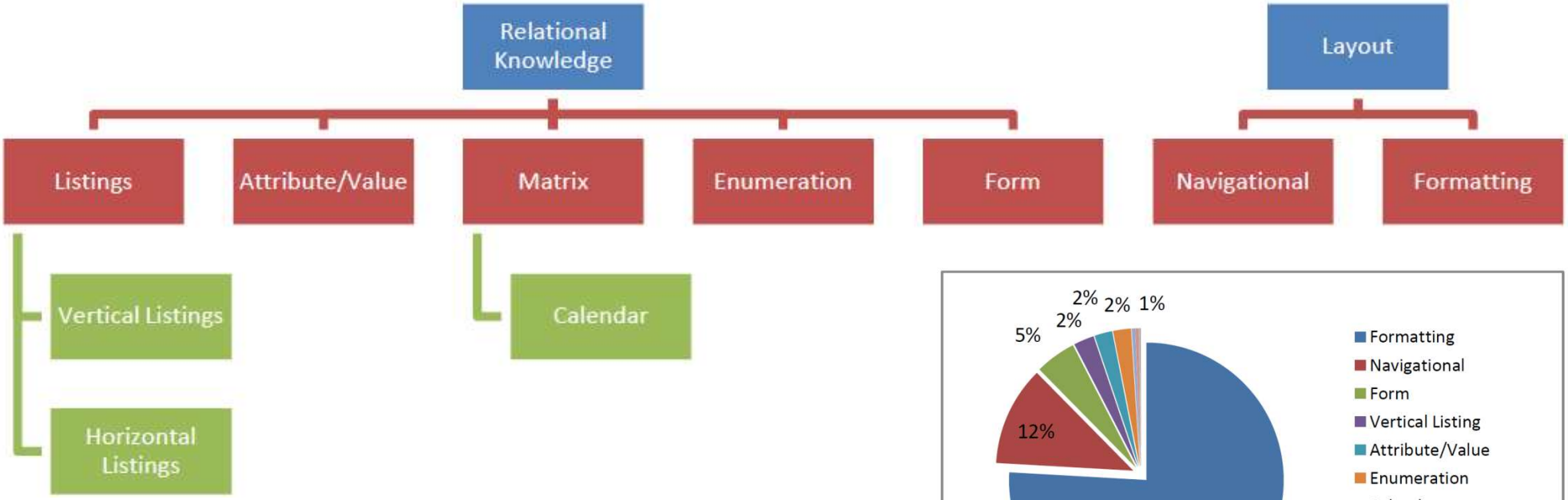
Here are the most requested songs of the past year:

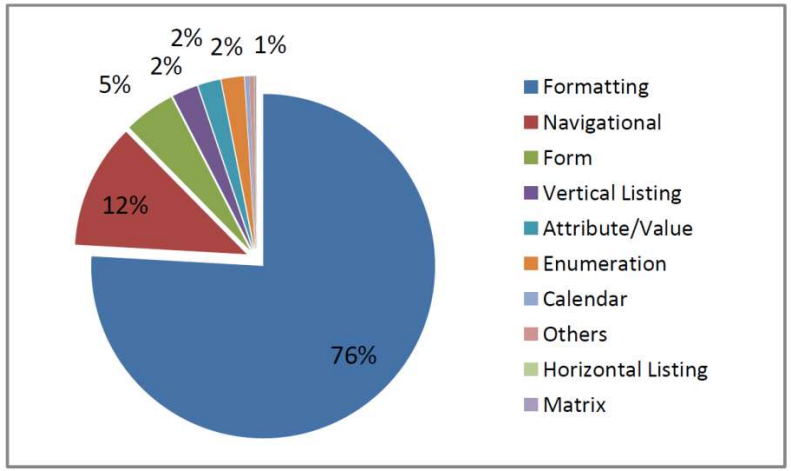
1	Black Eyed Peas	I Gotta Feeling
2	Journey	Don't Stop Believin'
3	Lady Gaga Feat. Colby O'donis	Just Dance
4	AC/DC	You Shook Me All Night Long
5	Cupid	Cupid Shuffle
6	Bon Jovi	Livin' On A Prayer
7	Beyonce	Single Ladies (Put A Ring On It)
8	Diamond, Neil	Sweet Caroline (Good Times Never Seemed So Good)
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

Types of Web Tables

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

Cafarella (2008)





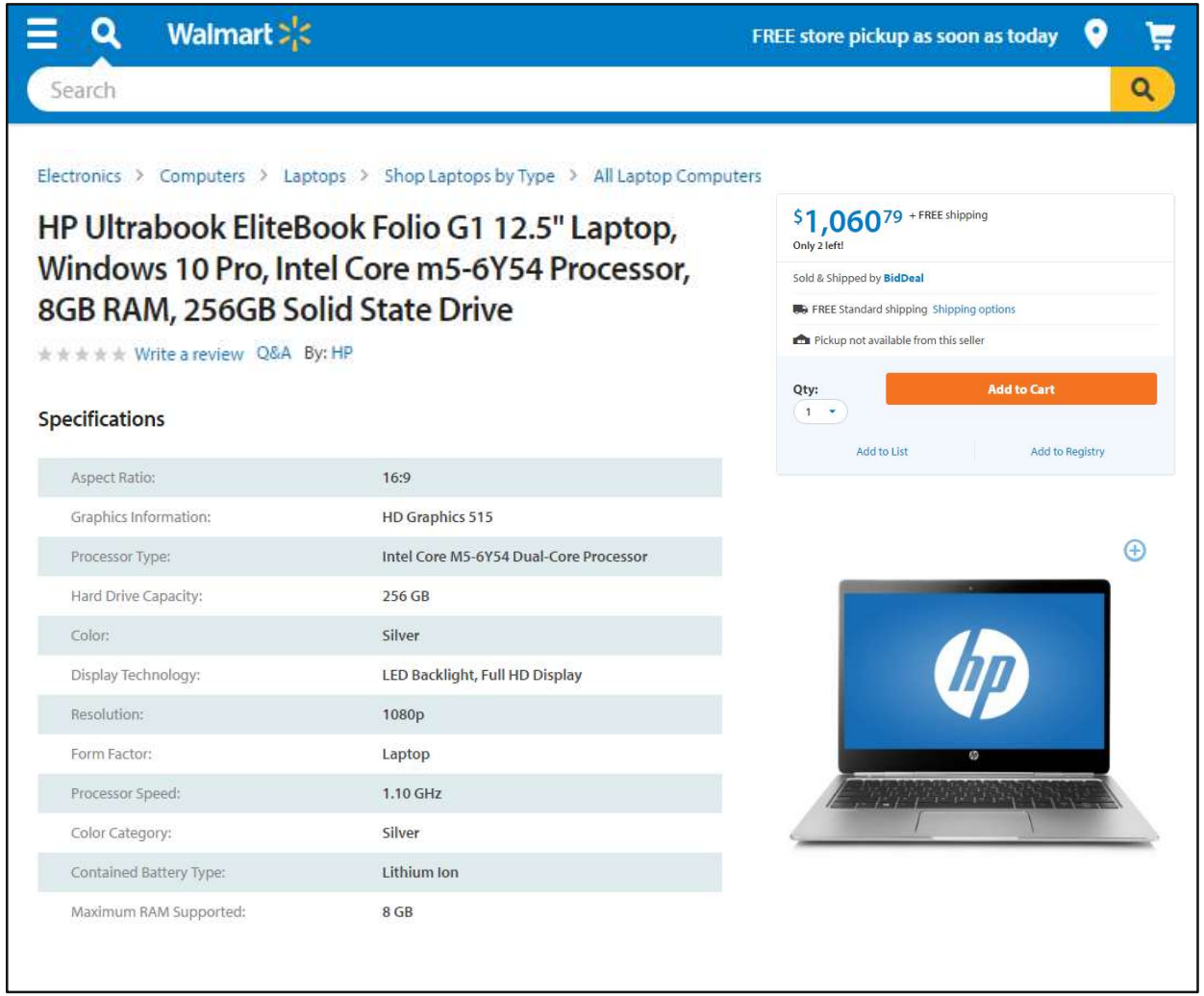
Cafarella, et al.: **WebTables: Exploring the Power of Tables on the Web.** VLDB 2008.
Crestan, Pantel: **Web-Scale Table Census and Classification.** WSDM 2011.

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

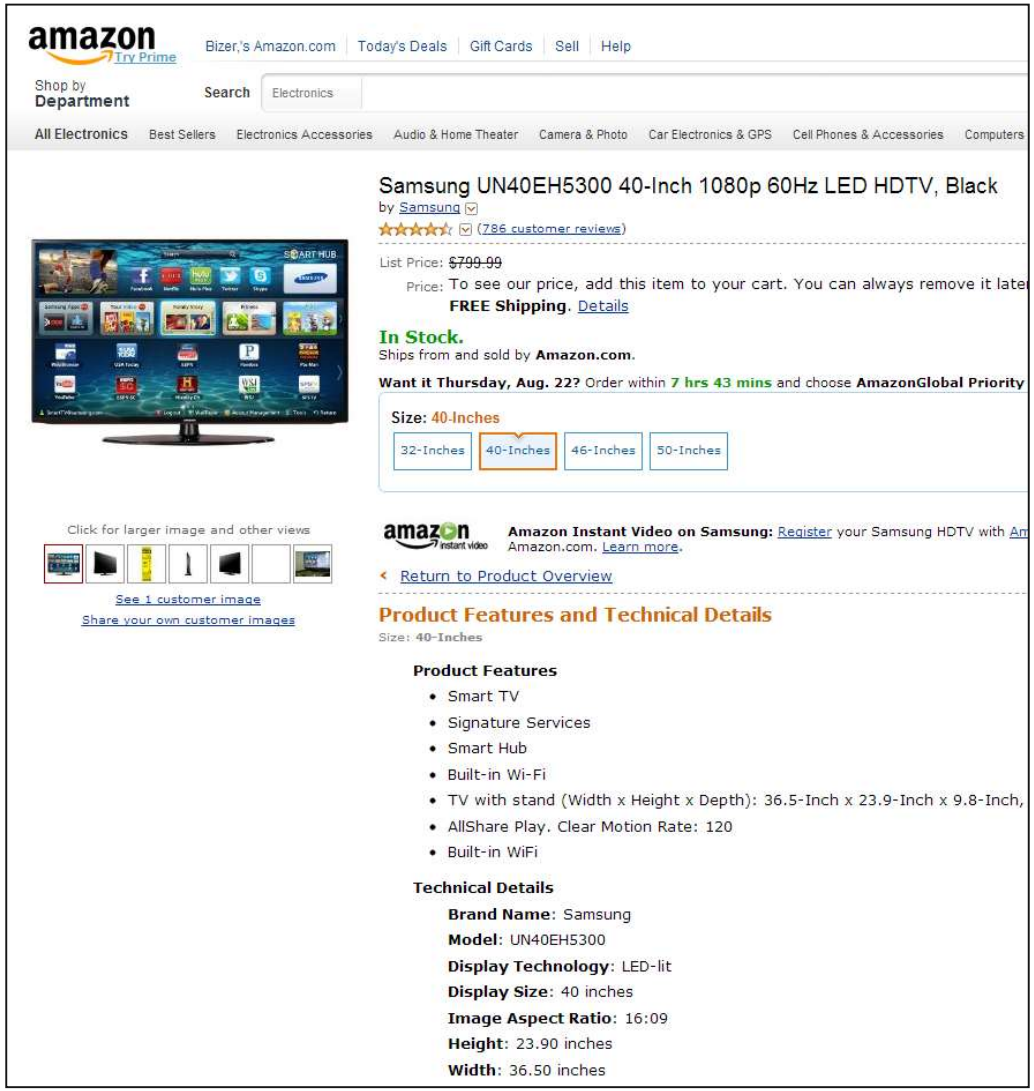
- Subject Attribute Values

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 ways out: 1. knowledge graphs as distant supervision, 2. LLM-based extraction (expensive)



4.3 Wikipedia as Data Source

Title

Description

Cross Language Links

Geo-Coordinates

Images

Infoboxes

Bristol - Wikipedia, the free encyclopedia - Mozilla Firefox

http://en.wikipedia.org/wiki/Bristol

article | discussion | edit this page | history

Log in / create account

Bristol

From Wikipedia, the free encyclopedia

Coordinates: 51°28'N 2°35'W

Bristol (pronunciation (help·info); IPA ˈbrɪstəl) 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 [hide]

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 [edit]

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.

Bristol

Bridge and the Avon Gorge

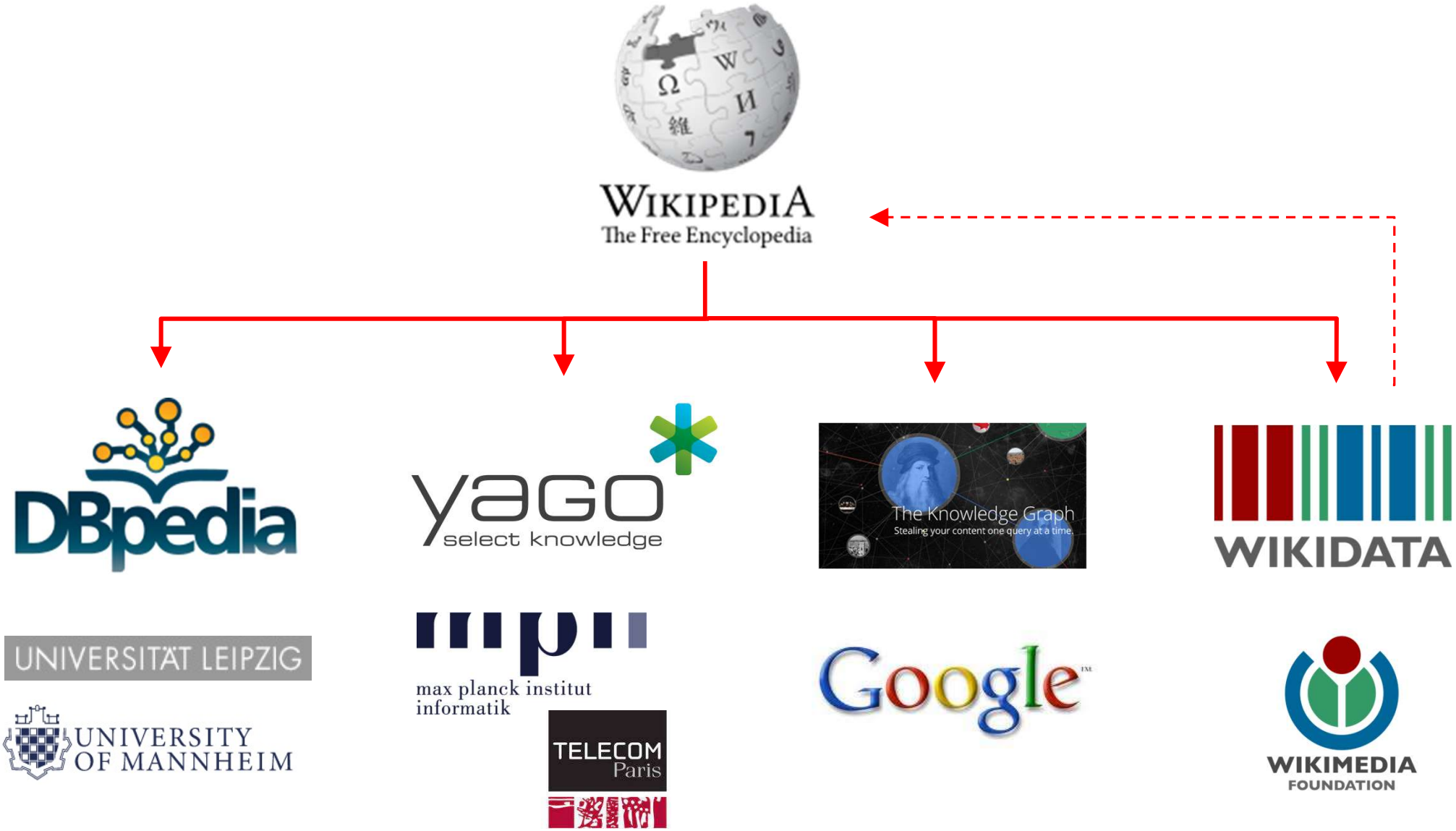
Coat of Arms of the City Council

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) Avon
Admin HQ	Bristol
Royal Charter	1155
County status	1373
Government	
- 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)

Anglo Saxon, العربية, Български, Català, Česky, Cymraeg, Dansk, Deutsch, Eesti, Español, Esperanto, Euskara, Français, हिन्दी, Ido, Bahasa Indonesia, Íslenska, Italiano, עברית, Latviešu, Lëtzebuergesch, Lietuvių, Magyar, Nederlands, 日本語, Norsk (bokmål)

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.

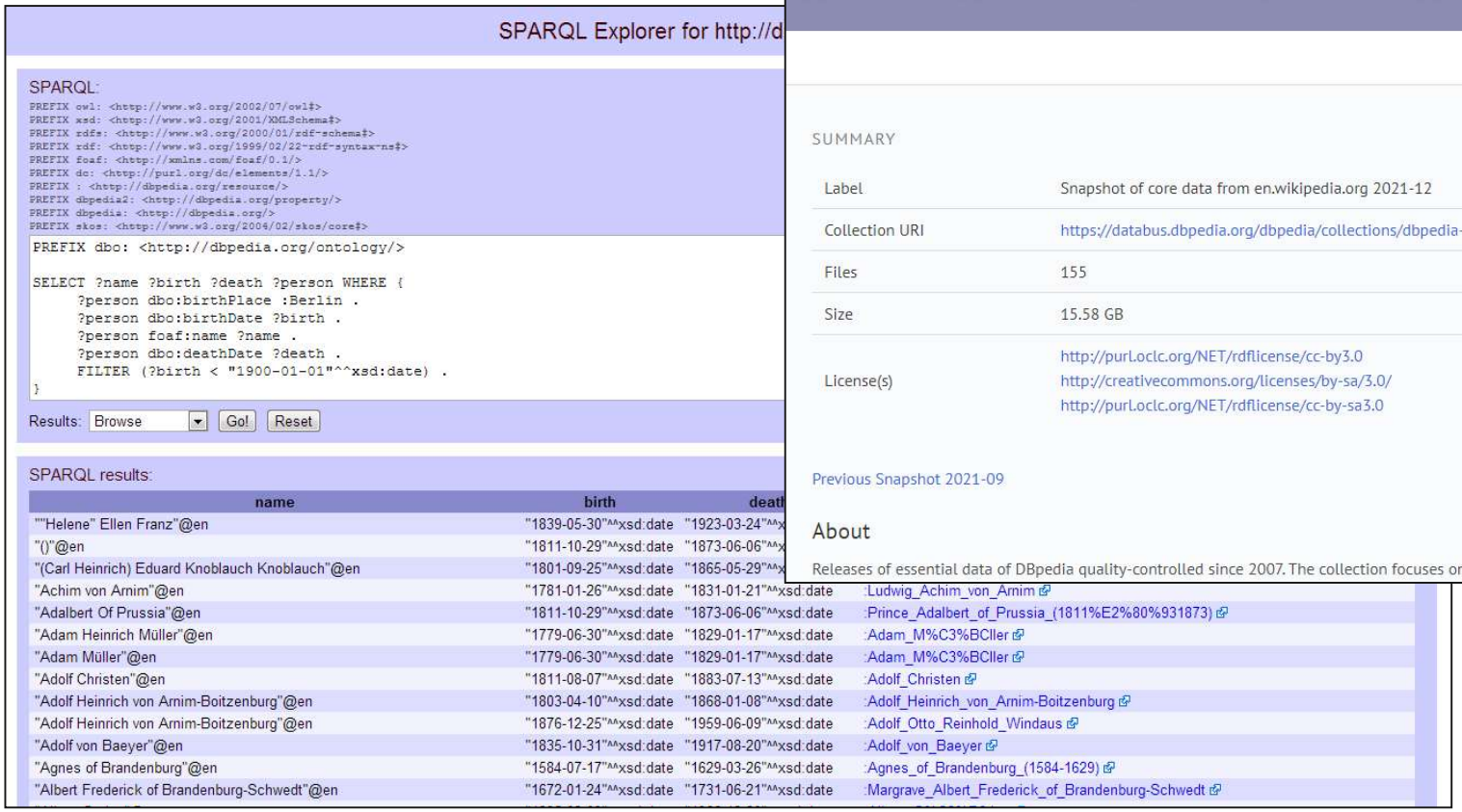
The Harbourside

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
  - 1 billion objects described by over 70 billion facts (2019)
  - 1500 classes, 35,000 properties
- Microsoft Knowledge Graph
  - revealed to the public in mid-2013
  - 2 billion objects described by 55 billion facts
- Knowledge Graphs employ RDF-style graph data models

See also: IE650 Knowledge Graphs

Noy et al.: Industry-scale Knowledge Graphs. Communications of the ACM, Vol. 62, 2019.

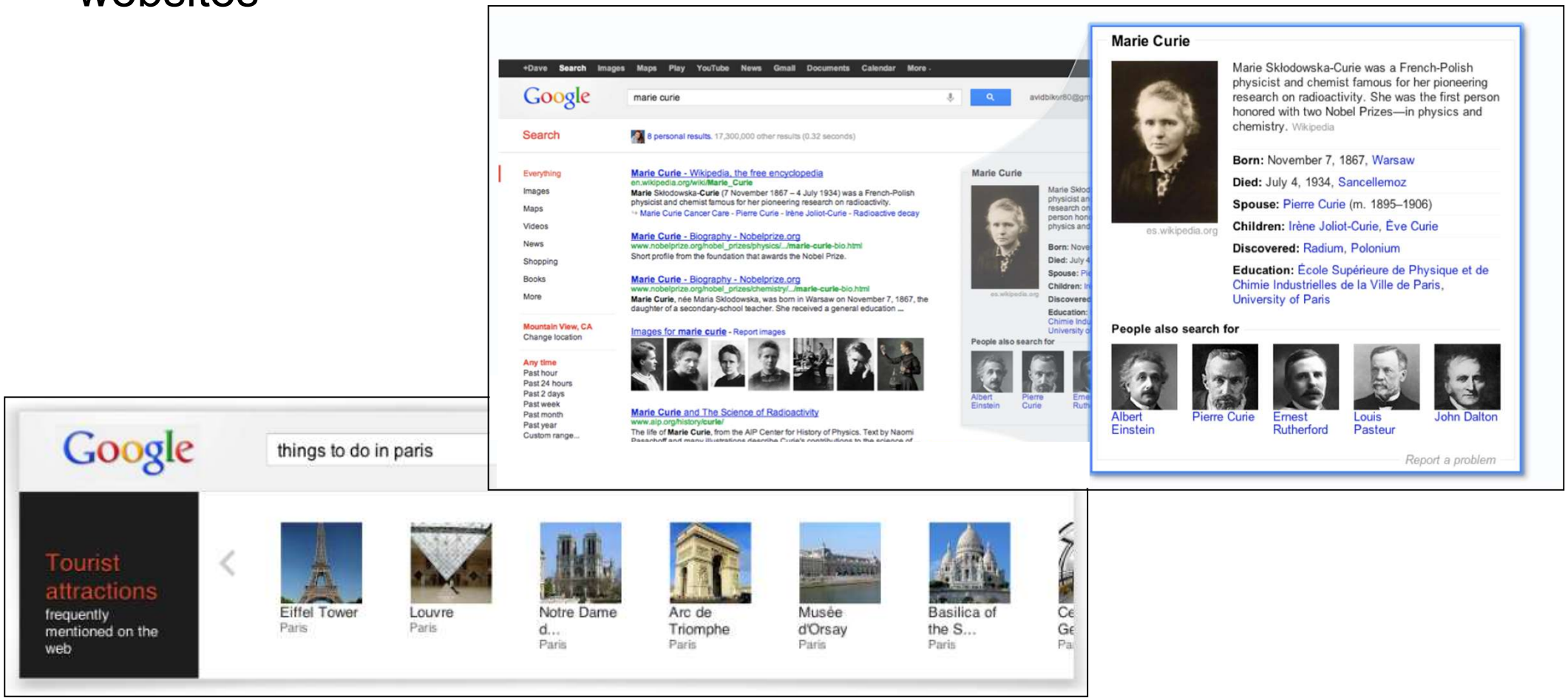
Data Sources used to Build Knowledge Graphs

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 curation / cleansing

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?

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

Summary

- Web data integration aims at integrating data from large numbers of data sources
  - integration projects in corporate data lakes usually involve less sources
- The topics of the published data partly correlate with the publication methods used:
  - Data Portals: public sector data, statistical data, research and ML data
  - Web APIs: user generated content, location-related data, weather data
  - Schema.org data: e-commerce, local business, event, job 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.
  - Alexander Brinkmann, et al.: The Web Data Commons schema.org Dataset Serie. Companion of the WWW Conference, 2024.
- 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.
- Knowledge Graphs
  - Jens Lehmann, et al: DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal, 2014.