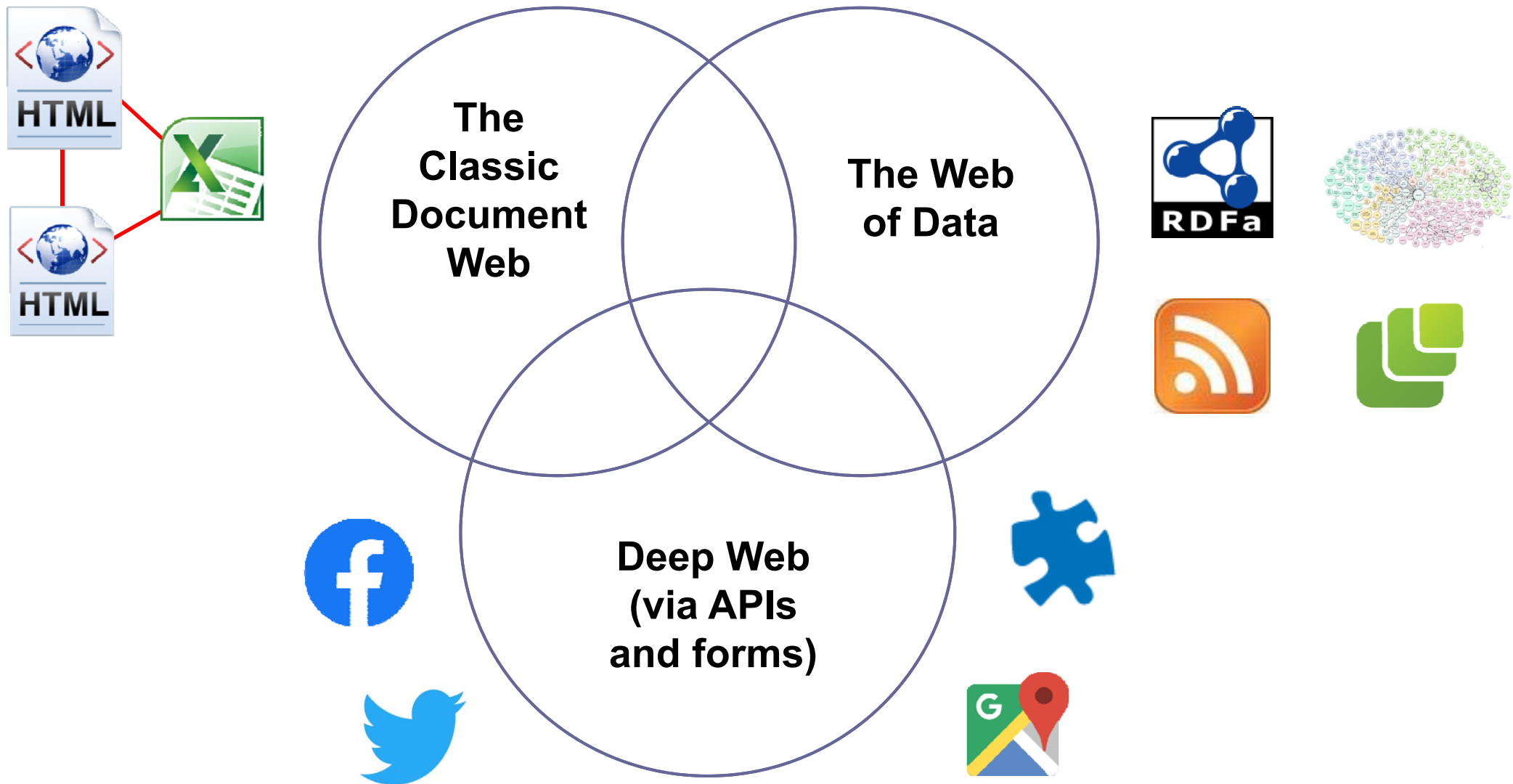


Web Data Integration

Types of Structured Data on the Web



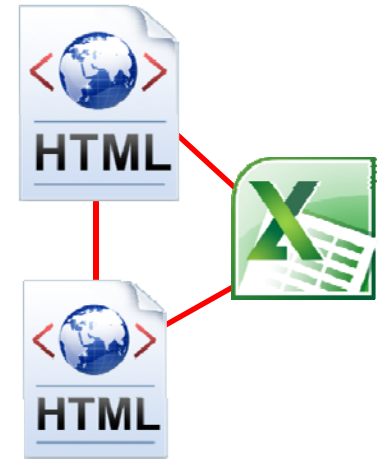
Topology of the Web Today



1. Data Catalogs and Marketplaces
2. Web APIs
3. Linked Data
4. HTML-embedded Data
 1. RDFa, Microdata, Microformats
 2. HTML Tables and Templates
 3. Wikipedia as Data Source
5. References

1. Data Catalogs and Marketplaces

- The Web traditionally contains structured data in various formats:
 - CSV files, Excel worksheets
 - XML documents, SQL dumps
- Data Catalogs and Data Marketplaces
 - collect and host data sets plus metadata
 - provide free or payment-based access to the data sets
- Examples
 - data.gov.uk, data.gov.us, publicdata.eu: host many public sector data sets
 - Factual, data.world, dmi.io: Commercial data providers / data markets
- List of Data Catalogs
 - <http://data.wu.ac.at/portalwatch/portalslist>





Data Search

753

Results

Tag: finance xSort by: Last Updated

Licence

[Open Government Licence \(742\)](#)[Non-Open Government Licence \(11\)](#)

Tags

finance x[transparency \(661\)](#)[disclosure \(642\)](#)[spend-transactions \(614\)](#)

Spend over £25,000 in Heywood, Middleton and Rochdale Clinical Commissioning Group

CSV

Publisher: [NHS Heywood, Middleton and Rochdale Primary Care Trust](#)

Updated: 20 Aug 2013

A monthly-updated list of all financial transactions spending over £25,000 made by Heywood, Middleton and Rochdale Clinical Commissioning Group, as part of the Government's commitment to...



Spend over £25,000 in NHS Bury Clinical

CSV

Nach Datensätzen suchen

Suchen

Erweiterte Suche

Kartensuche

31771 Treffer



Erweiterte Suche

Alles

Daten

Informationen

Blog-Beiträge

Sortieren nach

Filtermöglichkeiten

zurücksetzen

KATEGORIEN

☐ Bevölkerung und Gesellschaft

7444

☐ Bildung, Kultur und Sport

2501

☐ Energie

375

☐ Gesundheit

1160



Datensatz

Inhaltsstoffbestimmung im PM10-Feinstaub nach der 39. BImSchV 2017 (Datentabelle)

Nach der 39. BImSchV (Verordnung über Luftqualitätsstandards und Emissionshöchstmengen) ist die Luftqualität im Hinblick auf Staubinhaltsstoffe zu ermitteln. Diese sind die Schwermetalle Blei, Arse...

Veröffentlichende Stelle: -

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

100+ results found



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

Bevölkerungsbestand in Mannheim 2013-2018

[Explore at mannheim.opendatasoft.com](https://mannheim.opendatasoft.com)

Dataset updated 15.07.2019

License

dl-de-by-2.0

Available download formats from providers

excel, csv, json

Description

Bevölkerung in Mannheim und den Stadtteilen nach Jahr, Wohnstatus, Geschlecht und Migrationshintergrund Mannheim ist Standort einer Landeserstaufnahmeeinrichtung (LEA) sowie Bedarfsorientierter Erstaufnahmeeinrichtungen (BEAs) für Flüchtlinge. Die Bewohnerinnen und Bewohner dieser Einrichtungen sind meldepflichtig und fließen als Hauptwohnbevölkerung in die Statistik der Stadt Mannheim ein.

<https://toolbox.google.com/datasetsearch>

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** via the Web.
 - HTML interfaces
 - Web APIs



Example: Facebook

- 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/>

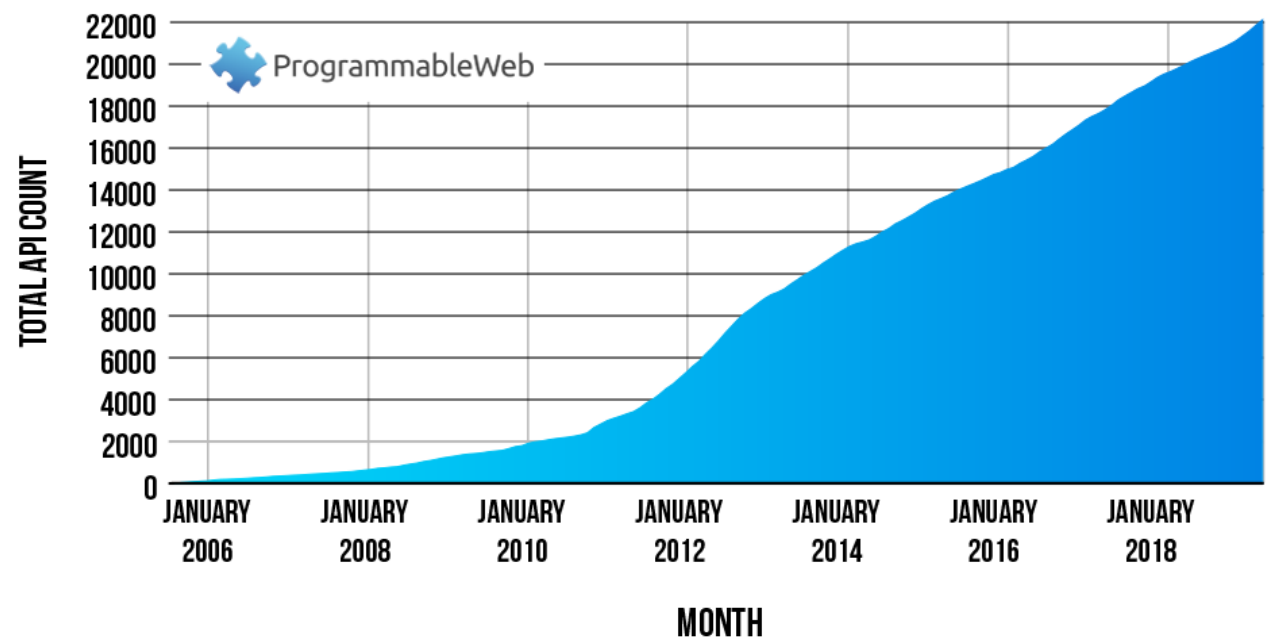
<http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/>

Web APIs

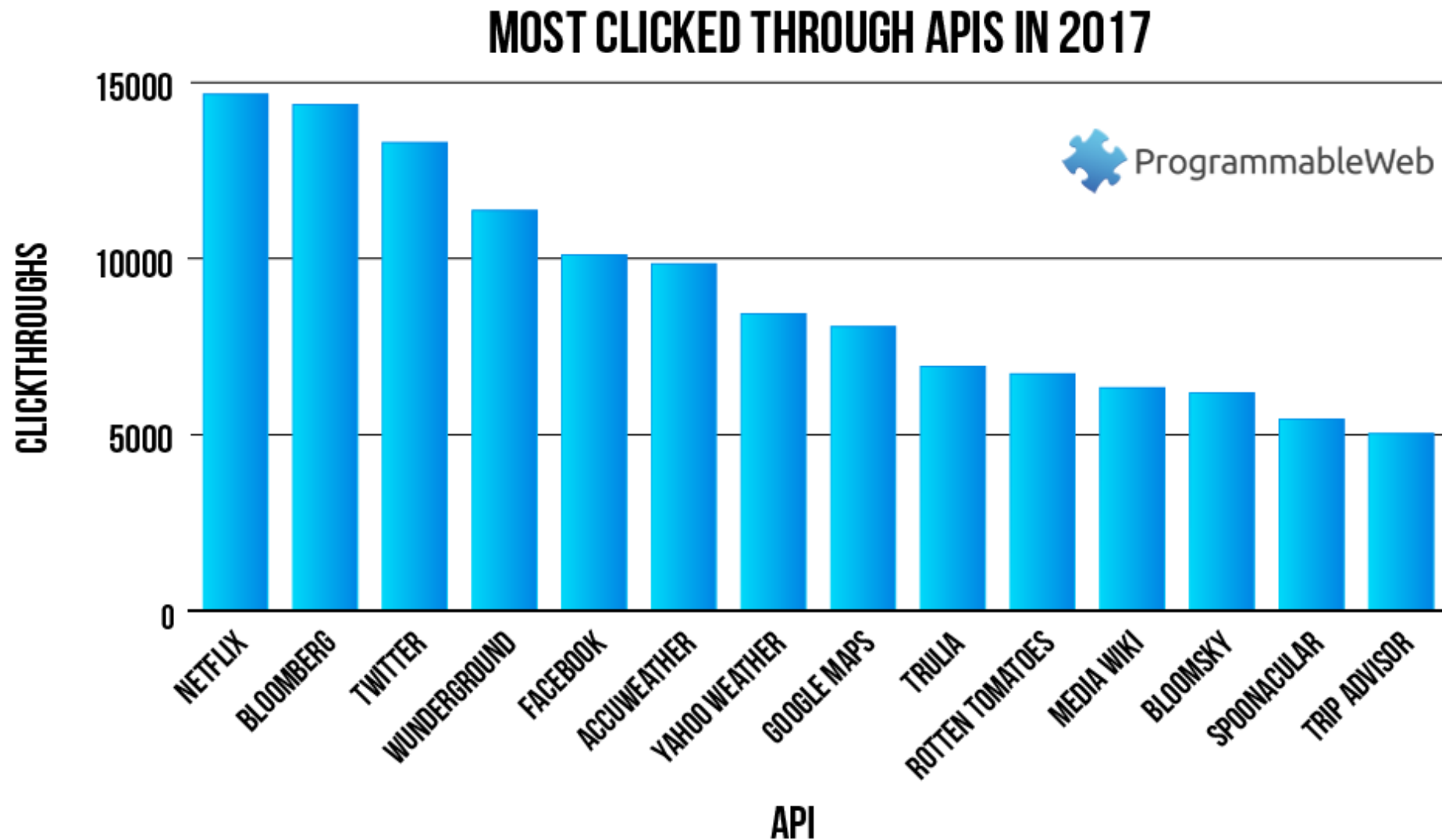
- Provide limited access to the collected data
 - restricted to specific queries (canned queries)
 - restricted by number of queries / number of results
- ProgrammableWeb API Catalog
 - lists over 20,000 Web APIs
 - lists over 6,800 mashups



GROWTH IN WEB APIS SINCE 2005

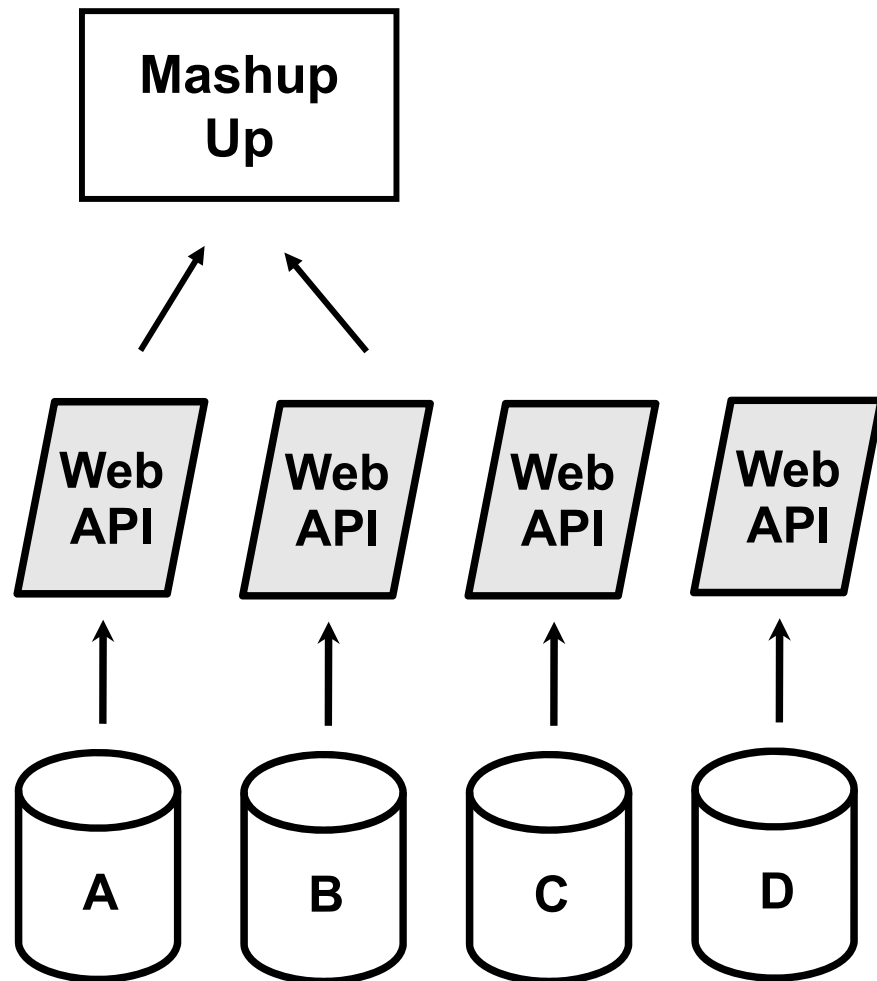


Most Popular APIs



<https://www.programmableweb.com/news/programmablewebs-most-popular-apis-2017/research/2018/02/21>

Mashups are based on a fixed set of data sources

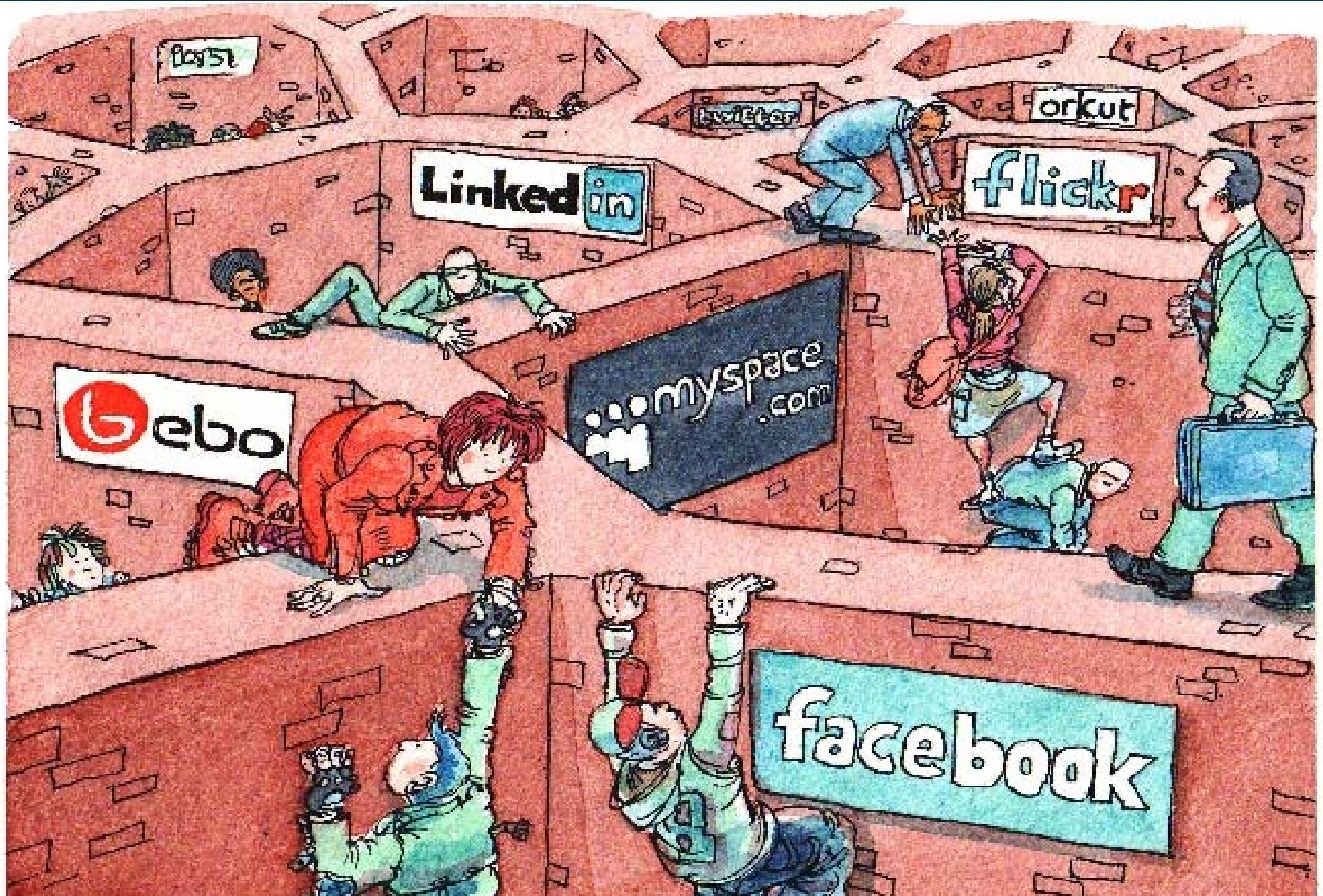


Web APIs expose proprietary interfaces.

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



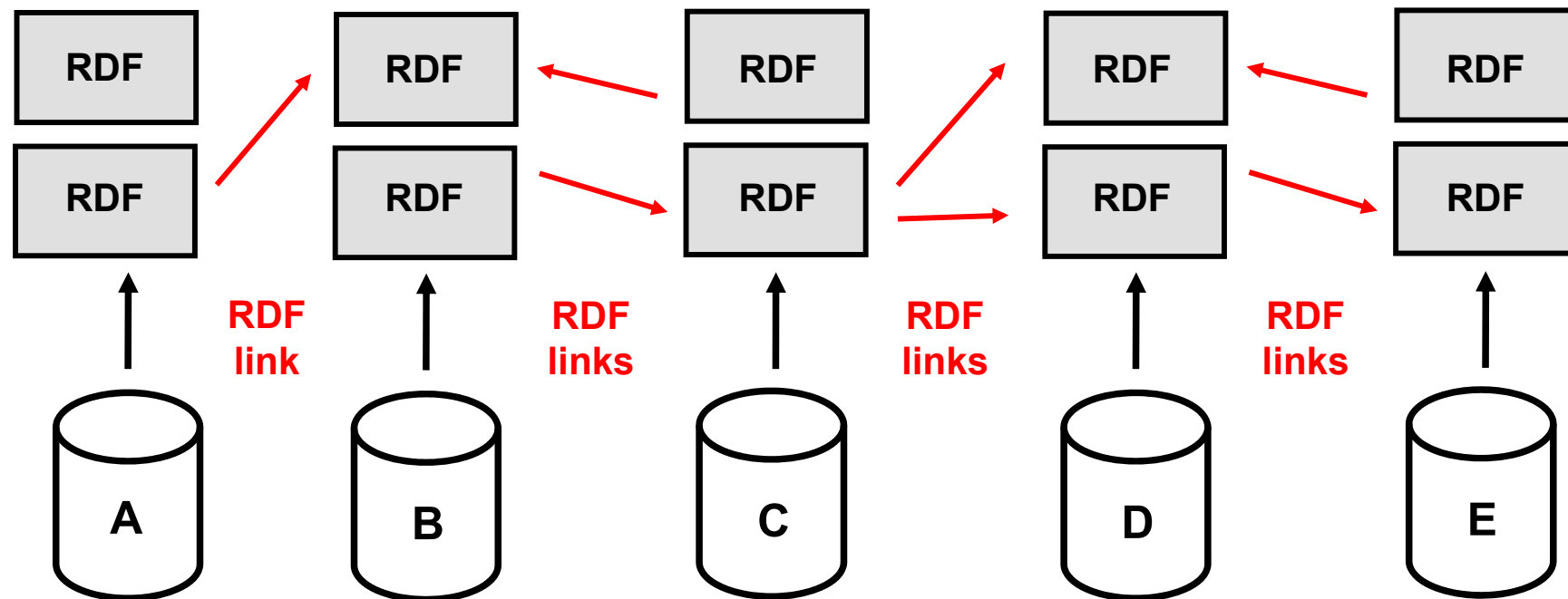
Web APIs slice the Web into Data Silos



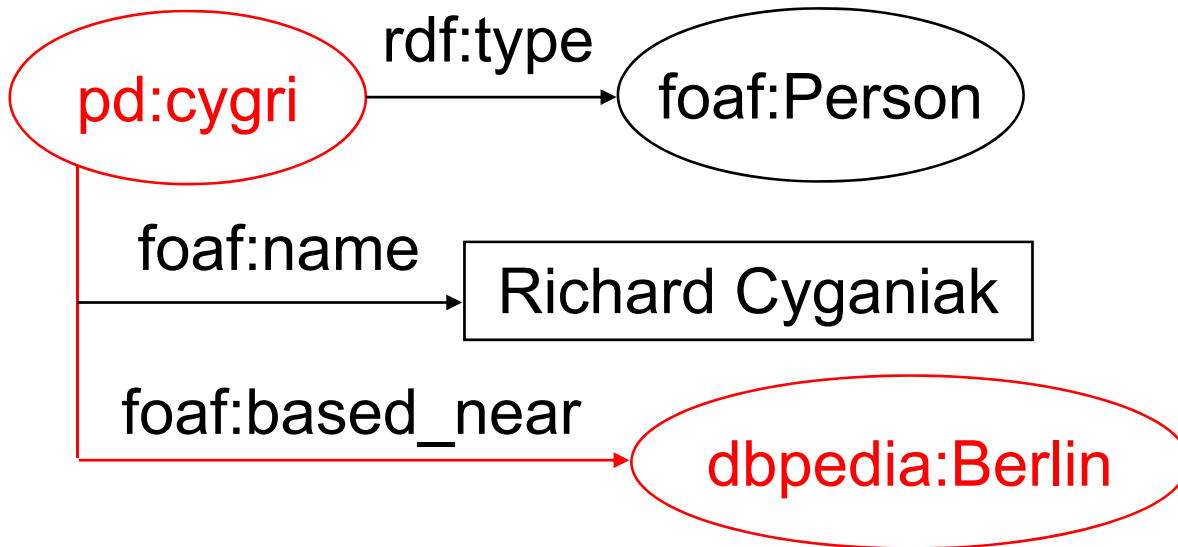
3. Alternative Approach: Linked Data



- 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

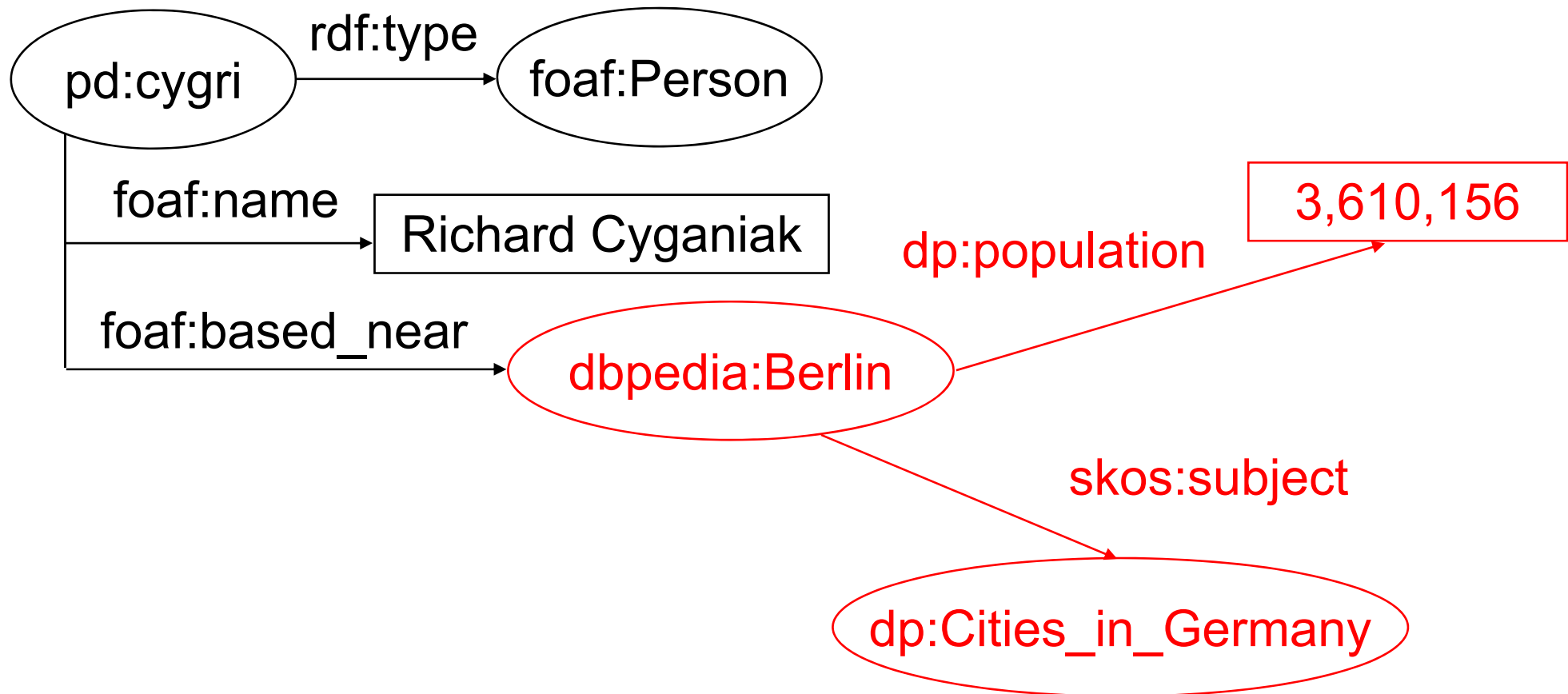


HTTP URIs take the role of global primary keys.

`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

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.wiwiw.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](#)

- [985c47c5a70db7407210cef6e4e8f5374a525c5c](#)   

[workplace homepage](#)

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

[nickname](#)

- [TimBL](#)    

[nickname](#)

- [TimBL](#)    

- [timbl](#)  

[personal mailbox](#)

- <mailto:timbl@w3.org>   

The SigMa Linked Data Search Engine

[Help](#)[About](#)[Forum](#)[Add More Info](#)[Start New](#)[Options](#)[Order](#)[Permalink](#)[Sources \(20\)](#)[Appro](#)

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]

1 [Chris Bizer - Free Uni](#)
http://videlectures.net/chris_

2 [Chris Bizer - semantic](#)
<http://ontoworld.org/wiki/Chris>

3 [Untitled document](#) 6 f
BOSS <http://www.facebook>

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

1,239 datasets connected by
16,147 sets of RDF links
(as of March 2019)

Legend

Cross Domain

Geography

Government

Life Sciences

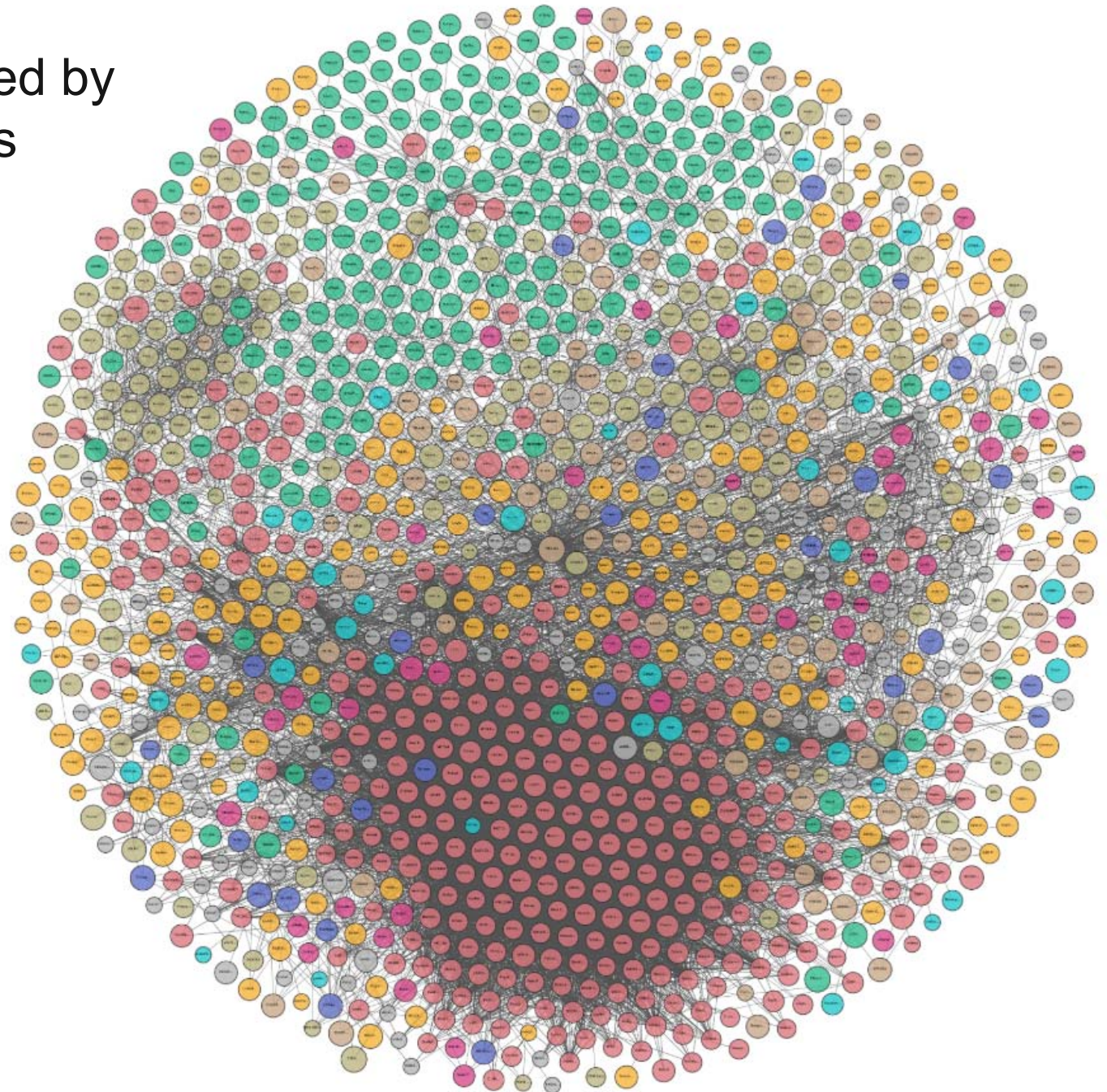
Linguistics

Media

Publications

Social Networking

User Generated



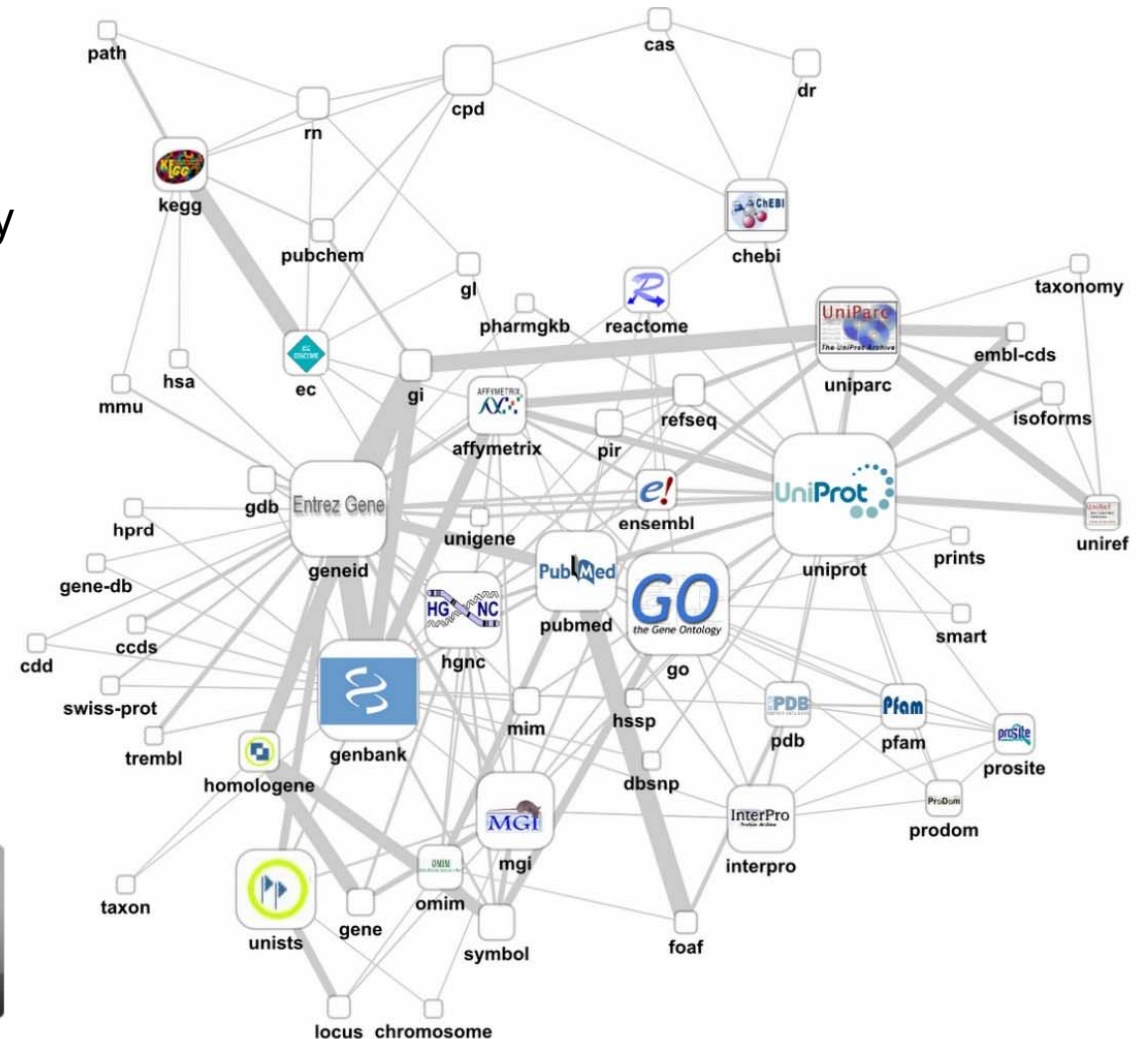
<https://lod-cloud.net/>

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

BIO↔RDF



Uptake in the Libraries Community

- Institutions publishing Linked Data
 - Library of Congress (subject headings and catalog)
 - German National Library (PND dataset and subject headings)
 - Swedish National Library (Libris catalog)
 - Hungarian National Library (OPAC and digital library)
 - Europeana Digital Library (catalog)
 - Springer Nature (publications, researchers, projects)
- Goals:
 1. Interconnect resources between repositories (by topic, by location, by historical period, by ...)
 2. Integrate library catalogs on global scale



LIBRARY OF
CONGRESS



National Library
of Sweden



Hands-on: How to get the Data?

- 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://sparqls.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



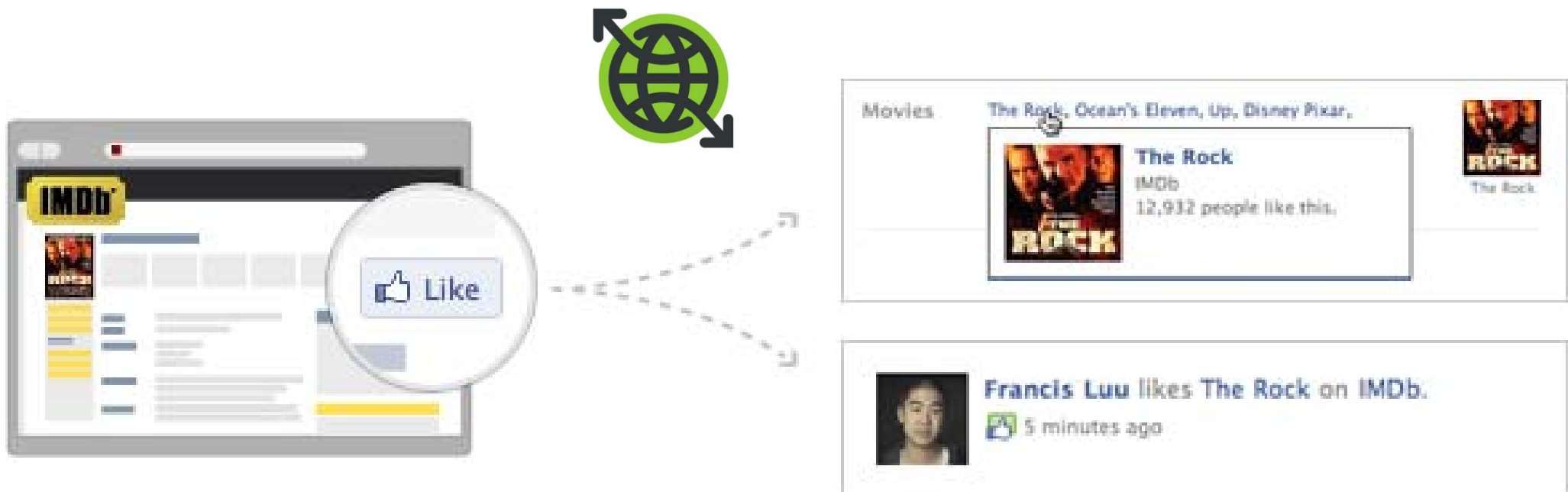
- Microformat effort dates back to 2003
- Small set of fixed formats
 - hcard : people, companies, organizations, and places
 - XFN : relationships between people
 - hCalendar : calendaring and events
 - hListing : small-ads; classifieds
 - hReview : reviews of products, businesses, events
- Shortcoming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



- 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 ...
```

- 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>
```

- used for embedding data into the HEAD of HTML pages
- putting data in 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.

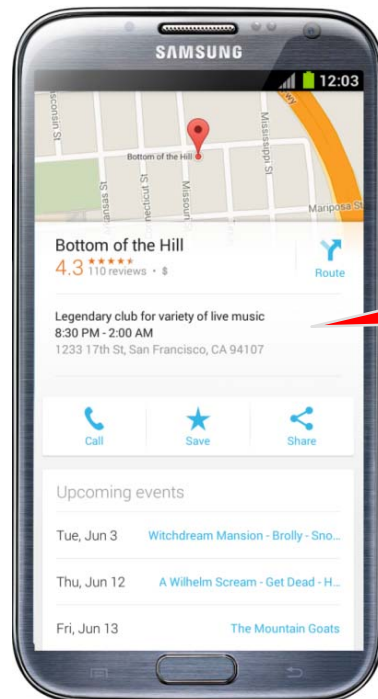
Property	Expected Type	Description
Properties from <u>Thing</u>		
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
url	URL	URL of the item.
Properties from <u>Place</u>		
address	<u>PostalAddress</u>	Physical address of the item.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

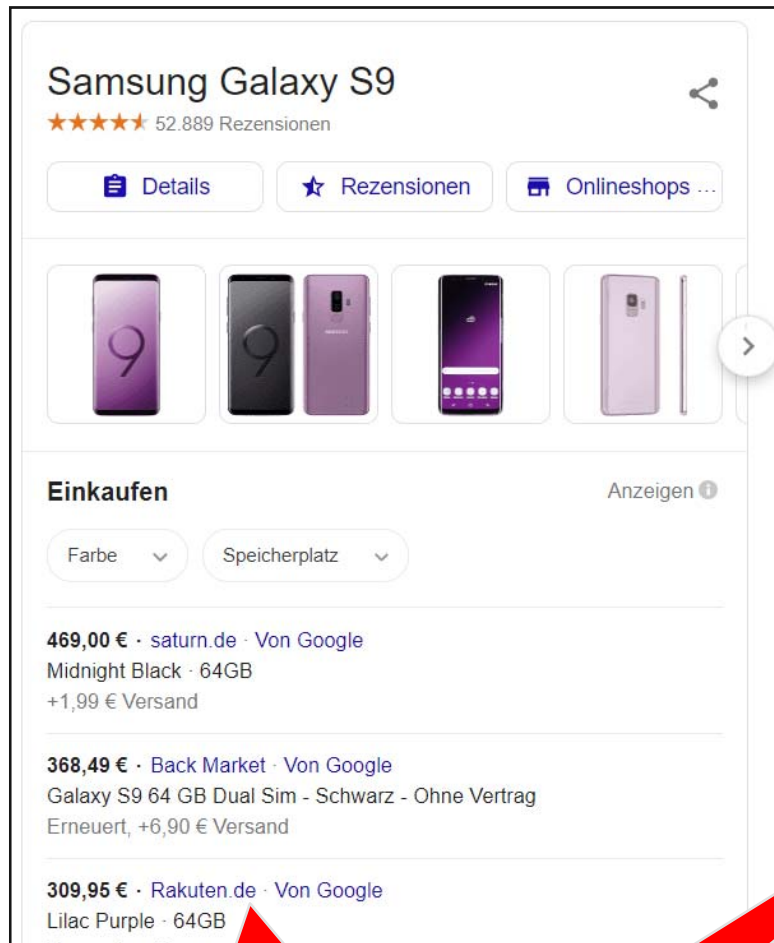
Record labels: Fat Possum Records, Nonesuch Records, V2 Records, Alive Natural Sound Records

Awards: Grammy Award for Best Rock Album, more

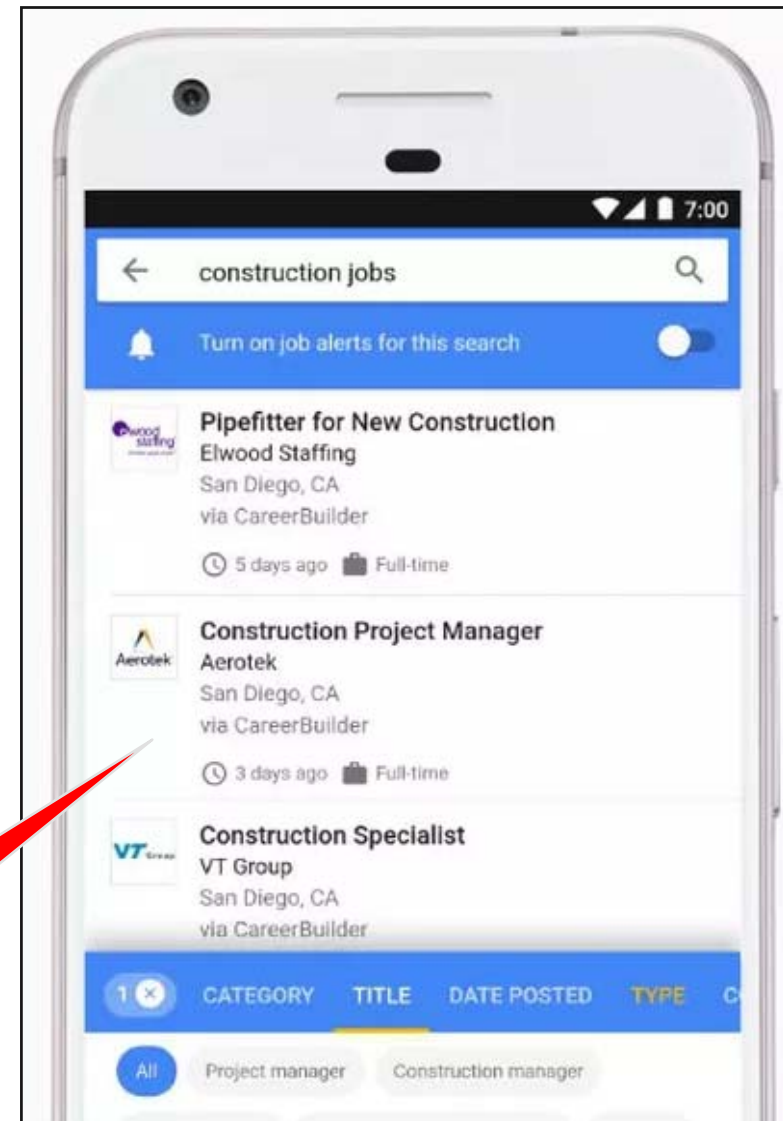
Upcoming events

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

Usage of Schema.org Data @ Google



Product
offers



Job
offers

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

The Web Data Commons Project

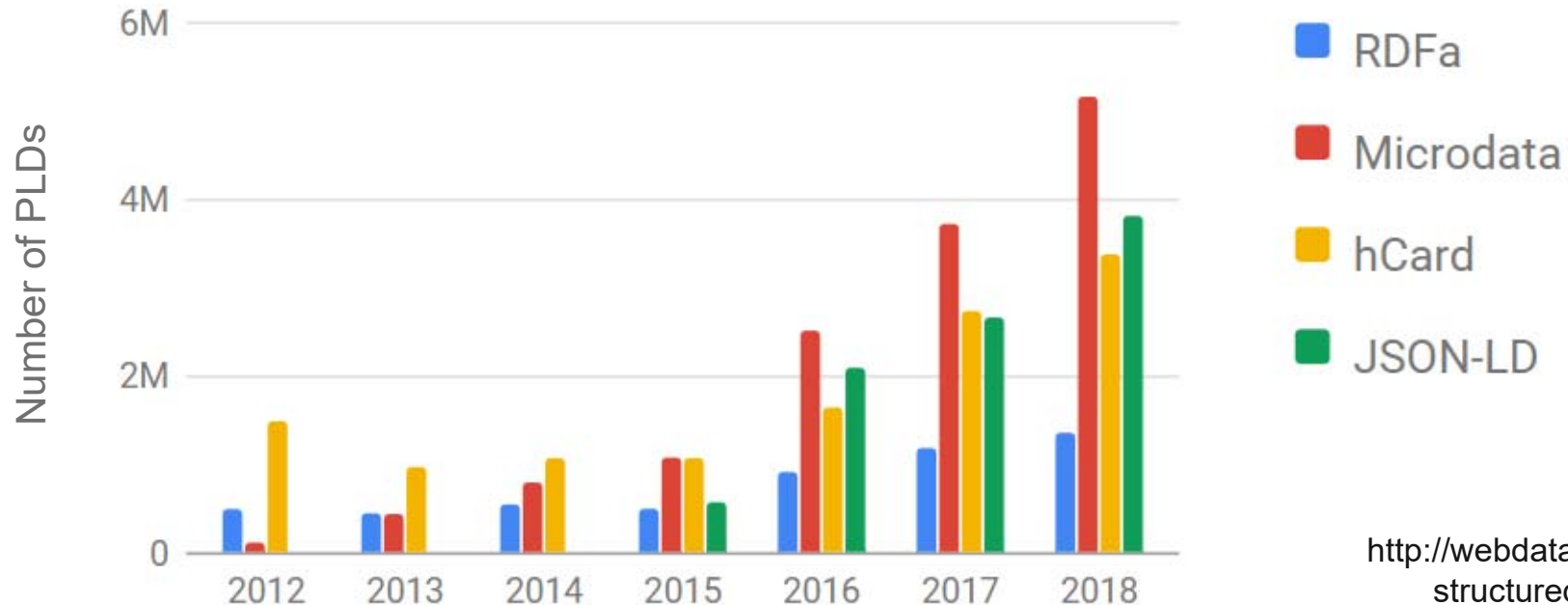


- extracts all Microformat, Microdata, RDFa, JSON-LD data from the **Common Crawl**
- analyzes and provides the extracted data for download
- statistics about some extraction runs
 - 2018 CC Corpus: **2.5 billion HTML pages** → **31.5 billion RDF triples**
 - 2017 CC Corpus: 3.1 billion HTML pages → 38.2 billion RDF triples
 - 2014 CC Corpus: 2.0 billion HTML pages → 20.4 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 → 250 Euro
- <http://webdatacommons.org/structureddata/>

Overall Adoption 2018

944 million HTML pages out of the 2.5 billion pages provide semantic annotations (37.1%).

9.6 million pay-level-domains (PLDs) out of the 32.8 million pay-level-domains covered by the crawl provide semantic annotations (29.3%).



Frequently used Schema.org Classes

Class	# Websites (PLDs)	
	Microdata	JSON-LD
schema:WebPage	1,124,583	121,393
schema:Product	812,205	40,169
schema:Offer	676,899	57,756
schema:BreadcrumbList	621,344	205,971
schema:Article	612,361	57,082
schema:Organization	510,069	1,349,775
schema:PostalAddress	502,615	176,500
schema:ImageObject	360,875	111,946
schema:Blog	337,843	12,174
schema:Person	324,349	335,784
schema:LocalBusiness	294,390	249,017
schema:AggregateRating	258,078	23,105
schema:Review	124,022	6,622
schema:Place	92,127	66,396
schema:Event	88,130	63,605

<http://webdatacommons.org/structureddata/2018-12/>

Properties used to Describe Products

Top Attributes	PLDs Microdata	
	#	%
schema:Product/name	754,812	92 %
schema:Product/offers	645,994	79 %
schema:Offer/price	639,598	78 %
schema:Offer/priceCurrency	606,990	74 %
schema:Product/image	573,614	70 %
schema:Product/description	520,307	64 %
schema:Offer/availability	477,170	58 %
schema:Product/url	364,889	44 %
schema:Product/sku	160,343	19 %
schema:Product/aggregateRating	141,194	17 %
schema:Product/brand	113,209	13 %
schema:Product/category	62,170	7 %
schema:Product/productID	47,088	5 %
...	...	...

<http://webdatacommons.org/structureddata/2018-12/stats/html-md.xlsx>

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 %

Hands-on: How to get the Data?

- <http://www.webdatacommons.org/structureddata/>

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-W				
7 Stuttgart  , Baden-Würt				
8 Dortmund  , North Rhine				
9 Düsseldorf  , North Rhin				
10 Bremen  , Bremen				

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	1.70 m (5 ft 7 in)	Nassau
Aquelle Plakaris	24	1.70 m (5 ft 7 in)	Evergem
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

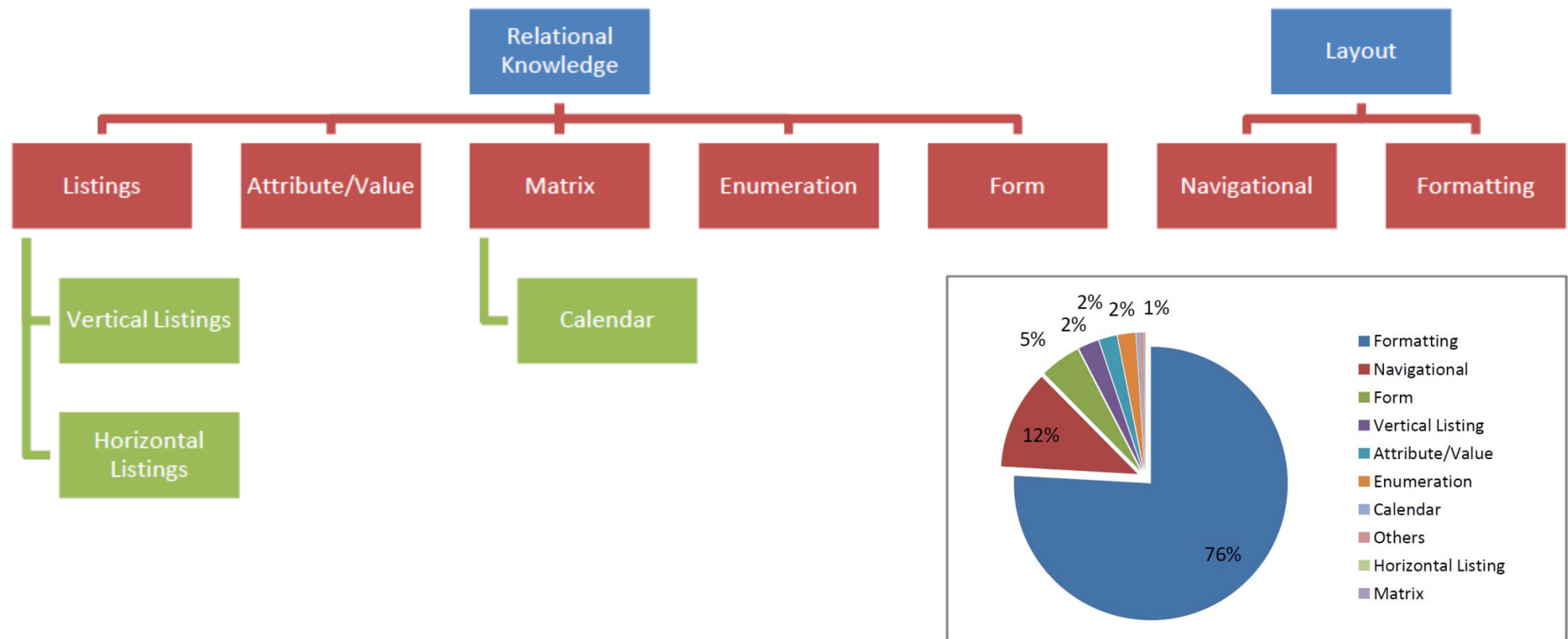
WEBSITE	HUB	AIRLINE CODE
www.aeroflot.ru/eng/	Moscow	SU
www.airfrance.fr	Paris	AF
	Tokyo	NH
	Seoul	OZ
	Hong Kong	CX

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	Omigod

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

Walmart

FREE store pickup as soon as today

Search

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

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

★★★★★ Write a review Q&A By: HP

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

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

Sold & Shipped by BidDeal

FREE Standard shipping Shipping options

Pickup not available from this seller

Qty: 1 Add to Cart

Add to List Add to Registry

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/>
- Alternative: Google Table Search
 - <https://research.google.com/tables>



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

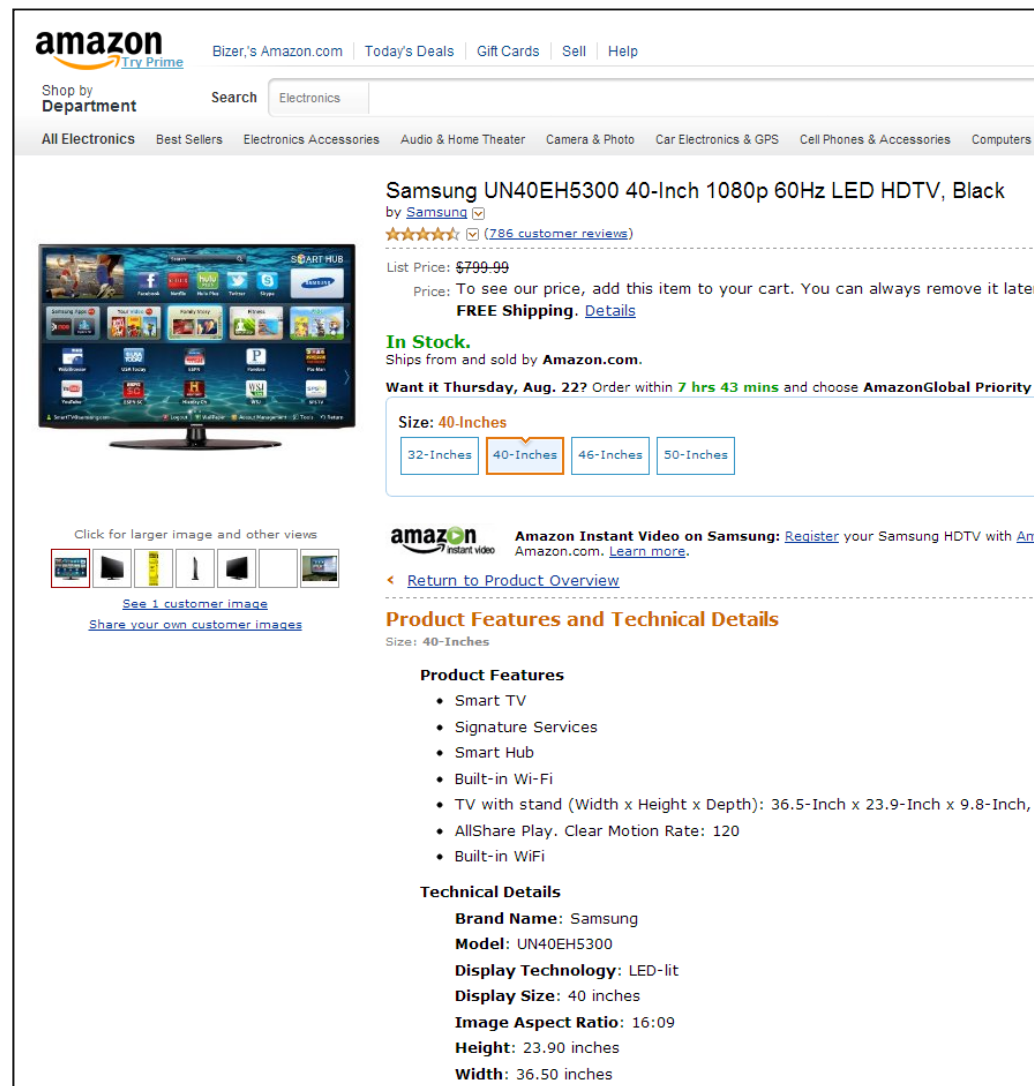
Value	#Rows
usa	135,000
germany	91,000
greece	42,000
new york	59,000
london	37,000
athens	11,000
david beckham	3,000
ronaldinho	1,200
oliver kahn	710
twist shout	2,000
yellow submarine	1,400

1.74 billion rows

253,000,000 different subject labels

Exploiting the Template-Structure of HTML Pages

- Many websites 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

Bristol - Wikipedia, the free encyclopedia - Mozilla Firefox

Wikipedia, the free encyclopedia

Bristol

From Wikipedia, the free encyclopedia

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

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 most populous city, 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]

- Boundaries
- History
- Economy and industry
- Culture
 - Arts
 - Sport and leisure
 - Media
 - Dialect
- Politics and government
- Demographics
- Physical geography
- Education, science and technology
- Transport
- Twin cities
- See also
- References
- 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 Stoop 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 / ﻿51.467°N 2.583°W﻿ / 51.467; -2.583

Bristol

Bridge and the Avon Gorge

Coat of Arms of the City Council

Coordinates: 51°27′14″N 2°35′48″W﻿ / ﻿51.454°N 2.597°W﻿ / 51.454; -2.597

sovereign state	 United Kingdom
constituent country	 England
region	 South West England
ceremonial county	 Bristol
historic counties	 County corporate (Gloucestershire and Somerset)
Avon	 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)

Title

Description

Cross
Language
Links

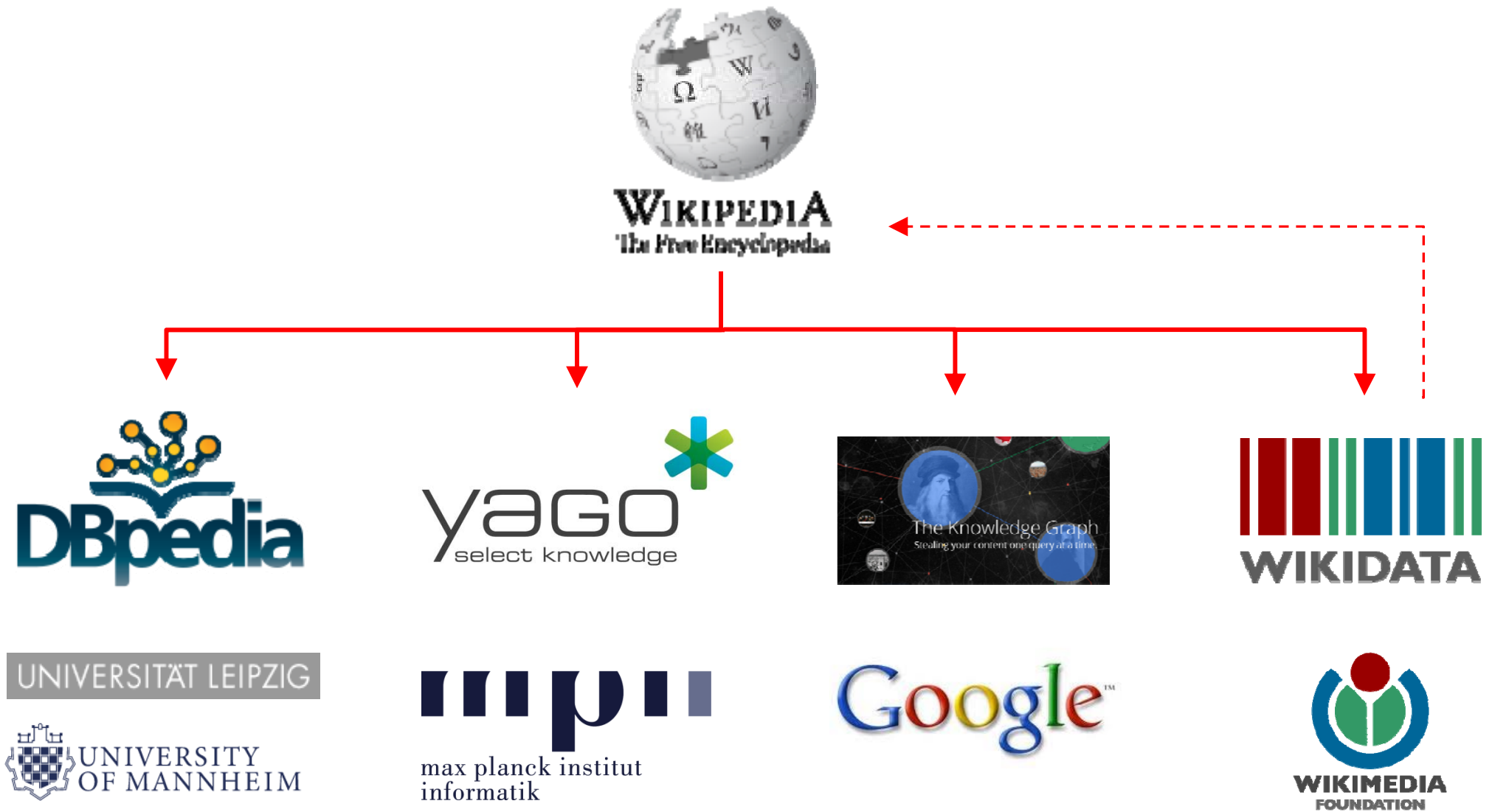
Geo-
Coordinates

Images

Infoboxes



Extracting Knowledge from Wikipedia



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

The DBpedia Knowledge Base - Release 2016



- Describes **6.6 million things**, out of which 5.5 million are classified in a consistent ontology using 760 classes and 2729 different properties
 - 1,500,000 persons
 - 840,000 places
 - 260,000 organizations
 - 139,000 music albums
- Altogether 13 billion pieces of information (RDF triples)
 - 1.7 million were extracted from the English edition of Wikipedia
 - 29,000,000 links to external web pages
 - 50,000,000 external links into 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

 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

SPARQL Explorer for <http://dbpedia.org/sparql>

SPARQL:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <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:

SPARQL results:

name	birth	death	person
""Helene"" Ellen Franz"@en	"1839-05-30"^^xsd:date	"1923-03-24"^^xsd:date	:Ellen_Franz ↗
"()@en	"1811-10-29"^^xsd:date	"1873-06-06"^^xsd:date	:Prince_Adalbert_of_Prussia_(1811%E2%80%931873) ↗
"(Carl Heinrich) Eduard Knoblauch Knoblauch"@en	"1801-09-25"^^xsd:date	"1865-05-29"^^xsd:date	:Eduard_Knoblauch ↗
"Achim von Arnim"@en	"1781-01-26"^^xsd:date	"1831-01-21"^^xsd:date	:Ludwig_Achim_von_Arnim ↗
"Adalbert Of Prussia"@en	"1811-10-29"^^xsd:date	"1873-06-06"^^xsd:date	:Prince_Adalbert_of_Prussia_(1811%E2%80%931873) ↗
"Adam Heinrich Müller"@en	"1779-06-30"^^xsd:date	"1829-01-17"^^xsd:date	:Adam_M%C3%BCller ↗
"Adam Müller"@en	"1779-06-30"^^xsd:date	"1829-01-17"^^xsd:date	:Adam_M%C3%BCller ↗
"Adolf Christen"@en	"1811-08-07"^^xsd:date	"1883-07-13"^^xsd:date	:Adolf_Christen ↗
"Adolf Heinrich von Arnim-Boitzenburg"@en	"1803-04-10"^^xsd:date	"1868-01-08"^^xsd:date	:Adolf_Heinrich_von_Arnim-Boitzenburg ↗
"Adolf Heinrich von Arnim-Boitzenburg"@en	"1876-12-25"^^xsd:date	"1959-06-09"^^xsd:date	:Adolf_Otto_Reinhold_Windaus ↗
"Adolf von Baeyer"@en	"1835-10-31"^^xsd:date	"1917-08-20"^^xsd:date	:Adolf_von_Baeyer ↗
"Agnes of Brandenburg"@en	"1584-07-17"^^xsd:date	"1629-03-26"^^xsd:date	:Agnes_of_Brandenburg_(1584-1629) ↗
"Albert Frederick of Brandenburg-Schwedt"@en	"1672-01-24"^^xsd:date	"1731-06-21"^^xsd:date	:Margrave_Albert_Frederick_of_Brandenburg-Schwedt ↗

 DEVELOP 

GETTING STARTED DATASETS SUPPORT GITHUB ONTOLOGY

DOWNLOADS 2016-10

Downloads 2016-10

This pages provides all DBpedia datasets as links to files in [bzip2](#) compression. The DBpedia datasets are licensed under the terms of the [Creative Commons Attribution-ShareAlike License](#) and the [GNU Free Documentation License](#).

In addition to the RDF version of the data, we also provide a tabular version of some of the core DBpedia data sets as CSV and JSON files. See [DBpediaAsTables](#).

See also the [change log](#) for recent changes and developments.

Contents

- [Wikipedia Input Files](#)
- [Ontology](#)
- [Dataset Table](#)
- [Links to other datasets](#)

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

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, ...

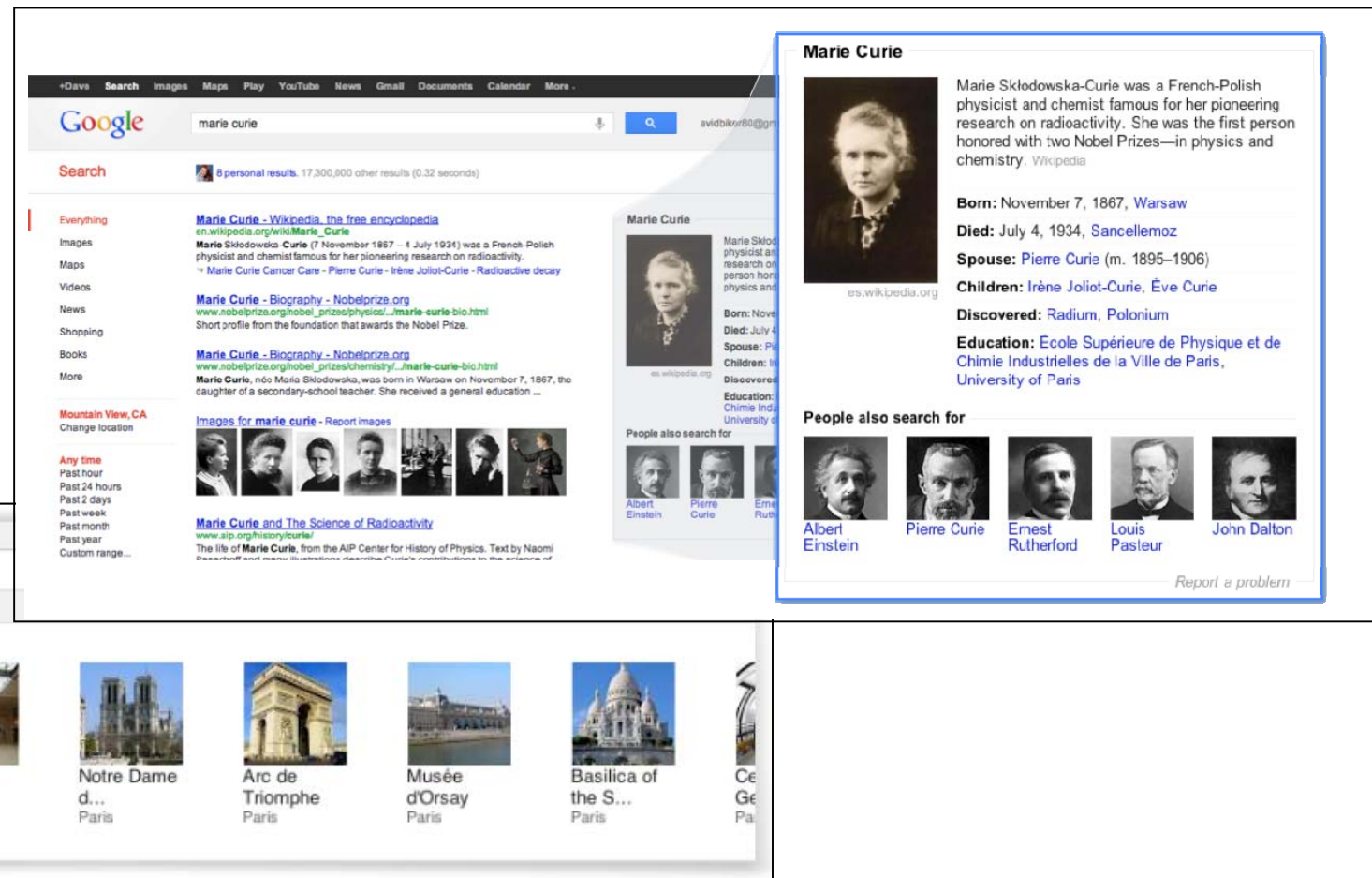
4. schema.org annotations in web pages

- 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: Fulfill information need without having users navigate to other websites



Applications of the Google Knowledge Graph

1. Answer fact queries: “birthdate michael douglas”
2. Compare things: ”compare eiffel tower vs empire state building”

September 25, 1944 (age 69 years)
Michael Douglas, Date of birth



 Catherine Zeta-Jones
Spouse
September 25, 1969

 Kirk Douglas
Father
December 9, 1916

 Morgan Freeman
June 1, 1937

[Feedback/More info](#)

	 Eiffel Tower	 Empire State Building
Height	324 m	381 m
Floors	3	102
Architect	Stephen Sauvestre	William F. Lamb
Construction started	January 28, 1887	1929
Architecture firms	Eiffel & Cie Barbier, Benard and Turenne	Shreve, Lamb & Harmon
Function	Observation tower Transmitter station	Tower block
Opened	March 31, 1889	May 1, 1931
Engineering firm	Eiffel & Cie	Starrett Brothers and Eken

[Feedback/More info](#)

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.

*Source: Nicolas Torzec, Yahoo 2014

SEO Topic: How to influence Knowledge Graphs?



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

Summary

- In addition to text, the Web contains a vast amount of structured data.
- The topics of the published data partly correlate with the publication methods used:
 - Web APIs: User generated content, geographic data
 - Schema.org data: E-commerce, local business, event, job data
 - Linked Data: E-government data, library data, research data
 - Wikipedia, HTML labels: 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.