Web Data Integration

Types of Structured Data on the Web
Topology of the Web Today

The classic Document Web

The Web of Data

Deep Web (via APIs and forms)
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: data.world, Foursquare
Example: Government Data Portal

Das Datenportal für Deutschland
Open Government: Verwaltungsdaten transparent, offen und frei nutzbar

DCAT-AP.de Version 2.0 veröffentlicht


https://www.govdata.de/
Geologische Übersichtskarte der Bundesrepublik Deutschland 1:200.000 (GÜK200) - CC 7110 Mannheim

Auf Blatt Mannheim ist der nördliche Oberrheingraben mit seinen mesozoischen Flanken dargestellt. Die dominierende Baueinheit im Kartenausschnitt ist der Oberrheingraben. Er durchzieht von Südsüdwest...
Example: Portal aggregating Metadata from other Portals

The official portal for European data

173 Catalogues 36 Countries 1,433,981 Datasets

Search datasets

https://data.europa.eu/en
Example: Institutional Research Data Repository

https://madata.bib.uni-mannheim.de/
Example: Focused Research Data Portal

Example: Data Portal facilitating the Replication of Research

https://paperswithcode.com/datasets
Maschine Learning-oriented Data Portals

- Combine sharing data, models, and code.
The FAIR Data Principles

- **Findable**
  1. (Meta)data are assigned a globally unique identifier
  2. Data are described with rich metadata
  3. (Meta)data are registered or indexed in a searchable resource

- **Accessible**
  1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  2. Metadata are accessible, even when the data are no longer available

- **Interoperable**
  1. (Meta)data use a formal, broadly applicable language for knowledge representation
  2. (Meta)data use vocabularies that follow FAIR principles
  3. (Meta)data include qualified references to other (meta)data

- **Reusable**
  1. (Meta)data are released with a clear data usage license
  2. (Meta)data are associated with detailed provenance
  3. (Meta)data meet domain-relevant community standards

https://www.go-fair.org/fair-principles/
Example: Dataset Search Engine

Google Dataset Search

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

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 MigrationshintergrundMannheim ist Standort einer Landeserstautnahmeeinrichtung (LEA) sowie Bedarfsorientierter Erstautnahmeeinrichtungen (BEAs) für Flüchtlinge. Die Bewohnerinnen und Bewohner dieser Einrichtungen sind meldepflichtig und fließen als Hauptwohnbewöl

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

Crawls dataset metadata from the Web.
2. Web 2.0 Applications and Web APIs

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

- Users (September 2018)
  - 2.3 billion monthly active users
  - including 1 billion mobile users
- 740 billion friend connections
- 4 million likes every minute
- 250 billion photos uploaded

- Data Volume
  - 4 Petabyte of new data generated every day
  - over 300 Petabyte in Facebook’s data warehouse

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

- Provide limited access to the collected data
  - restricted to specific queries (canned queries)
  - restricted by number of queries / number of results

- ProgrammableWeb API Catalog
  - listed over 24,000 Web APIs until 2022
  - listed over 6,800 mashups
Most Popular APIs

Top Tracked APIs of All Time

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
  - Principles F1, A1, I1, I2, I3
The Marbles Hyperdata Browser

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

Tim Berners-Lee

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

- Person

http://www.w3.org/2000/10/swa/swa-contact#Male

label

- Tim Berners-Lee

sameAs

- Tim Berners-Lee (also at www.wiwias.tu-berlin.de)

image

Web links

- 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 URL name

- 985d47c5a730db7407a16c4e85374a62635c5e

workplace homepage

- http://www.w3.org/

nickname

- TimBL

pmbox

- mailto:tim@w3.org
The SigMa Linked Data Search Engine

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]
The TriQL.P Browser: Filtering Information using Context-, Content- and Rating-Based Trust Policies. [16]
D2R Server - Publishing Relational Databases on the Semantic Web. [16]
Named Graphs, Provenance and Trust [16]
The Linked Open Data Cloud

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

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)

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**
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

```html
<html xmlns="http://www.w3.org/1999/xhtml"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:foaf="http://xmlns.com/foaf/0.1/">
  ...
  <div about="http://example.com/Peter" typeof="foaf:Person">
    <span property="foaf:name">Peter Smith</span> knows
  </div>
  ...
</html>
```
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)

```html
<div itemprop="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>
</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

```html
<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
Usage of Schema.org Data @ Google

**Data snippets within search results**

**Local businesses on maps**

**Data snippets within info boxes**

---

**Gramercy Tavern - Flatiron - New York, NY | Yelp**

www.yelp.com › Restaurants › American (New)

 فهي نمط مراجعة 4.5 - 1,200 نماذج - نطاق السعر: $500

جفف وك وليام في نيو يورك لأنه كان مهمًا وقليلًا ومثيرًا. كانت غرفة غرف الرفوف في نيو يورك مدهشة ومريحة، وتأكدنا أنها ستعود مرة أخرى. غرف غرفة غرف الرفوف في نيو يورك هي مدهشة ومريحة.

---

**Gramercy Tavern Restaurant - New York, NY | OpenTable**

www.opentable.com › › Gramercy restaurants

 فهي نمط مراجعة 4.7 - 508 نماذج - نطاق السعر: $60 and over

أصلاً، أود أن أقول إن غرف غرفة غرف الرفوف في نيو يورك تتميز بالرقة والروعة، حيث يبرز غرف غرفة غرف الرفوف والمحفظة في غرف غرفة غرف الرفوف في نيو يورك. كانت غرفة غرفة غرف الرفوف في نيو يورك مدهشة ومريحة، وتأكدنا أنها ستعود مرة أخرى. غرفة غرفة غرف الرفوف في نيو يورك مدهشة ومريحة، وتأكدنا أنها ستعود مرة أخرى.

---

**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](https://en.wikipedia.org/wiki/The_Black_Keys)

**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**

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jun 20</td>
<td>The Black Keys</td>
</tr>
<tr>
<td>Fri</td>
<td>Neusaben ob Eck (near you)</td>
</tr>
<tr>
<td>May 16</td>
<td>The Black Keys</td>
</tr>
<tr>
<td>Fri</td>
<td>Gulf Shores, AL</td>
</tr>
<tr>
<td>Jun 22</td>
<td>The Black Keys</td>
</tr>
<tr>
<td>Sun</td>
<td>Scheels</td>
</tr>
</tbody>
</table>
Usage of Schema.org Data @ Google

https://developers.google.com/search/docs/guides/search-gallery
The Web Data Commons Project

- extracts all Microformat, Microdata, RDFa, JSON-LD data from the Common Crawl
- analyzes and provides the extracted data for download
- statistics about some extraction runs
  - 2022 CC Corpus: 3.0 billion HTML pages → 86.4 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 → 800 Euro
- http://webdatacommons.org/structureddata/
Overall Adoption 2022

1.5 billion HTML pages out of the 3.15 billion pages provide semantic annotations (47%).

14.2 million pay-level-domains (PLDs) out of the 33.8 million PLDs (websites) provide semantic annotations (42%).

http://webdatacommons.org/structureddata/2022-12/stats/stats.html
### Frequently used Schema.org Classes (2020)

<table>
<thead>
<tr>
<th>Class</th>
<th># Websites (PLDs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JSON-LD</td>
</tr>
<tr>
<td>schema:WebPage</td>
<td>4,484,026</td>
</tr>
<tr>
<td>schema:Person</td>
<td>3,151,809</td>
</tr>
<tr>
<td>schema:BreadcrumbList</td>
<td>1,688,820</td>
</tr>
<tr>
<td>schema:Article</td>
<td>1,327,578</td>
</tr>
<tr>
<td>schema:Product</td>
<td>1,234,972</td>
</tr>
<tr>
<td>schema:Offer</td>
<td>1,182,855</td>
</tr>
<tr>
<td>schema:PostalAddress</td>
<td>863,243</td>
</tr>
<tr>
<td>schema:BlogPosting</td>
<td>529,020</td>
</tr>
<tr>
<td>schema:LocalBusiness</td>
<td>363,843</td>
</tr>
<tr>
<td>schema:AggregateRating</td>
<td>432,014</td>
</tr>
<tr>
<td>schema:Place</td>
<td>255,139</td>
</tr>
<tr>
<td>schema:Event</td>
<td>194,115</td>
</tr>
<tr>
<td>schema:Review</td>
<td>181,097</td>
</tr>
<tr>
<td>schema:JobPosting</td>
<td>28,759</td>
</tr>
</tbody>
</table>

## Adoption by Travel Websites

<table>
<thead>
<tr>
<th>Top 15 Travel Websites</th>
<th>schema:Hotel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Booking.com</td>
<td>✓</td>
</tr>
<tr>
<td>TripAdvisor</td>
<td>✓</td>
</tr>
<tr>
<td>Expedia</td>
<td>✓</td>
</tr>
<tr>
<td>Agoda</td>
<td>✓</td>
</tr>
<tr>
<td>Hotels.com</td>
<td>✓</td>
</tr>
<tr>
<td>Kayak</td>
<td>✓</td>
</tr>
<tr>
<td>Priceline</td>
<td>✓</td>
</tr>
<tr>
<td>Travelocity</td>
<td>✓</td>
</tr>
<tr>
<td>Orbitz</td>
<td>✓</td>
</tr>
<tr>
<td>ChoiceHotels</td>
<td>✓</td>
</tr>
<tr>
<td>HolidayCheck</td>
<td>✓</td>
</tr>
<tr>
<td>ChoiceHotels</td>
<td>✓</td>
</tr>
<tr>
<td>InterContinental Hotels Group</td>
<td>✓</td>
</tr>
<tr>
<td>Marriott International</td>
<td>✓</td>
</tr>
<tr>
<td>Global Hyatt Corp.</td>
<td>✗</td>
</tr>
</tbody>
</table>

Adoption: 93 %
## Properties used to Describe Products (2020)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>% of PLDs</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>schema:Product/name</code></td>
<td>99 %</td>
</tr>
<tr>
<td><code>schema:Product/offers</code></td>
<td>94 %</td>
</tr>
<tr>
<td><code>schema:Offer/price</code></td>
<td>95 %</td>
</tr>
<tr>
<td><code>schema:Offer/priceCurrency</code></td>
<td>95 %</td>
</tr>
<tr>
<td><code>schema:Product/description</code></td>
<td>84 %</td>
</tr>
<tr>
<td><code>schema:Offer/availability</code></td>
<td>72 %</td>
</tr>
<tr>
<td><code>schema:Product/sku</code></td>
<td>56 %</td>
</tr>
<tr>
<td><code>schema:Product/brand</code></td>
<td>30 %</td>
</tr>
<tr>
<td><code>schema:Product/image</code></td>
<td>26 %</td>
</tr>
<tr>
<td><code>schema:Product/aggregateRating</code></td>
<td>17 %</td>
</tr>
<tr>
<td><code>schema:Product/mpn</code></td>
<td>6.3 %</td>
</tr>
<tr>
<td><code>schema:Product/productID</code></td>
<td>4.7 %</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Total Number of</th>
<th>Top Classes (Entity Count)</th>
<th>Total File Size</th>
<th>Quad File</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://schema.org/AdministrativeArea">http://schema.org/AdministrativeArea</a></td>
<td>Quads: 1,724,657</td>
<td><a href="http://schema.org/AdministrativeArea">http://schema.org/AdministrativeArea</a> (100,671)</td>
<td>23 MB</td>
<td>schemaorgAdministrativeArea.nq.gz (sample)</td>
</tr>
<tr>
<td></td>
<td>URLs: 85,625</td>
<td><a href="http://schema.org/GeoCoordinates">http://schema.org/GeoCoordinates</a> (84,152)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hosts: 63</td>
<td><a href="http://schema.org/Country">http://schema.org/Country</a> (83,851)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://schema.org/">http://schema.org/</a> Continent (83,567)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://schema.org/Hotel">http://schema.org/Hotel</a></td>
<td>Quads: 148,211,253</td>
<td><a href="http://schema.org/Rating">http://schema.org/Rating</a> (7,007,590)</td>
<td>2,994 MB</td>
<td>schemaorgHotel.nq.gz (sample)</td>
</tr>
<tr>
<td></td>
<td>URLs: 3,136,152</td>
<td><a href="http://schema.org/Hotel">http://schema.org/Hotel</a> (6,335,124)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hosts: 5,337</td>
<td><a href="http://schema.org/Review">http://schema.org/Review</a> (4,408,551)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://schema.org/AggregateRating">http://schema.org/AggregateRating</a> (3,936,372)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>URLs: 2,011,332</td>
<td><a href="http://schema.org/Place">http://schema.org/Place</a> (16,321,339)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hosts: 3,962</td>
<td><a href="http://schema.org/Organization">http://schema.org/Organization</a> (12,164,867)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://schema.org/Postaladdress">http://schema.org/Postaladdress</a> (7,516,387)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://schema.org/PostalAddress">http://schema.org/PostalAddress</a></td>
<td>Quads: 776,573,609</td>
<td><a href="http://schema.org/PostalAddress">http://schema.org/PostalAddress</a> (48,086,763)</td>
<td>14,364 MB</td>
<td>schemaorgPostalAddress.nq.gz (sample)</td>
</tr>
<tr>
<td></td>
<td>URLs: 13,475,056</td>
<td><a href="http://schema.org/LocalBusiness">http://schema.org/LocalBusiness</a> (16,641,260)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hosts: 131,084</td>
<td><a href="http://schema.org/GeoCoordinates">http://schema.org/GeoCoordinates</a> (12,345,942)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://schema.org/Place">http://schema.org/Place</a> (9,071,774)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>URLs: 48,314,143</td>
<td><a href="http://schema.org/Offer">http://schema.org/Offer</a> (221,781,710)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hosts: 104,119</td>
<td><a href="http://schema.org/AggregateRating">http://schema.org/AggregateRating</a> (38,398,548)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://schema.org/Review">http://schema.org/Review</a> (28,209,876)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Only tip of the iceberg, as each website is only partly crawled.
There are hundreds of millions of high-quality HTML tables on the Web and in Wikipedia.

4.2 HTML Tables

---

### Germany - Largest Cities

<table>
<thead>
<tr>
<th>Name</th>
<th>Population</th>
<th>Latitude/Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlin</td>
<td>3,426,354</td>
<td>52.524 / 13.411</td>
</tr>
<tr>
<td>Hamburg</td>
<td>1,739,117</td>
<td>53.575 / 10.015</td>
</tr>
<tr>
<td>Munich</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cologne</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frankfurt am Main</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Essen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stuttgart</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dortmund</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bremen</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

### 150 INTERNATIONAL AFFAIRS

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>End Funding for the United Nations Development Program (UNDP)</td>
</tr>
<tr>
<td>9</td>
<td>End Funding for the U.N. Intergovernmental Panel on Climate Change (IPCC)</td>
</tr>
<tr>
<td>10</td>
<td>Eliminate the U.S. Trade and Development Agency (USTDA)</td>
</tr>
<tr>
<td>11</td>
<td>Reform Food Aid Programs</td>
</tr>
<tr>
<td>12</td>
<td>Eliminate Export-Import Bank</td>
</tr>
<tr>
<td>13</td>
<td>Eliminate the Overseas Private Investment Corporation (OPIC)</td>
</tr>
<tr>
<td>14</td>
<td>Eliminate Funding for the United Nations Population Fund (UNFPA)</td>
</tr>
</tbody>
</table>

---

### Most Requested Songs

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

---

**WEBSITE**
- www.aeroflot.ru/eng/
- www.airfrance.fr

**HUB**
- Moscow  SU
- Paris  AF
- Tokyo  NH
- Seoul  OZ
- Hong Kong  CX
Types of Web Tables

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

Cafarella (2008)

Crestan, Pantel: **Web-Scale Table Census and Classification.** WSDM 2011.
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/
## Attribute Statistics

<table>
<thead>
<tr>
<th>Attribute</th>
<th>#Tables</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>4,600,000</td>
</tr>
<tr>
<td>price</td>
<td>3,700,000</td>
</tr>
<tr>
<td>date</td>
<td>2,700,000</td>
</tr>
<tr>
<td>artist</td>
<td>2,100,000</td>
</tr>
<tr>
<td>location</td>
<td>1,200,000</td>
</tr>
<tr>
<td>year</td>
<td>1,000,000</td>
</tr>
<tr>
<td>manufacturer</td>
<td>375,000</td>
</tr>
<tr>
<td>country</td>
<td>340,000</td>
</tr>
<tr>
<td>isbn</td>
<td>99,000</td>
</tr>
<tr>
<td>area</td>
<td>95,000</td>
</tr>
<tr>
<td>population</td>
<td>86,000</td>
</tr>
</tbody>
</table>

28,000,000 different attribute labels

## Subject Attribute Values

<table>
<thead>
<tr>
<th>Value</th>
<th>#Rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>usa</td>
<td>135,000</td>
</tr>
<tr>
<td>germany</td>
<td>91,000</td>
</tr>
<tr>
<td>greece</td>
<td>42,000</td>
</tr>
<tr>
<td>new york</td>
<td>59,000</td>
</tr>
<tr>
<td>london</td>
<td>37,000</td>
</tr>
<tr>
<td>athens</td>
<td>11,000</td>
</tr>
<tr>
<td>david beckham</td>
<td>3,000</td>
</tr>
<tr>
<td>ronaldinho</td>
<td>1,200</td>
</tr>
<tr>
<td>oliver kahn</td>
<td>710</td>
</tr>
<tr>
<td>twist shout</td>
<td>2,000</td>
</tr>
<tr>
<td>yellow submarine</td>
<td>1,400</td>
</tr>
</tbody>
</table>

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

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
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

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?

J.Crew
- Specially retailer company
- J.Crew Group, Inc., is an American multi-brand, multi-channel, specialty retailer. The company offers an assortment of women’s, men’s and children’s apparel and accessories, including swimwear, outerwear, ... Wikipedia
- Customer service: 1 (800) 562-0258
- Headquarters: New York City, NY
- CEO: Mickey Drexler
- Founder: Emily Scott
- Founded: 1983

Company type: can be influenced by Wikidata or Wikipedia

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

Logo: can be specified by using Organization 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

People also search for
- Banana Republic
- Nordstrom
- Anthropologie
- Ann Inc.
- Vineyard Vines

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

Profiles
- Facebook
- Instagram
- Twitter
- LinkedIn
- Google+

Recent posts on Google+
- J.Crew
- 1,488,424 followers • Shared publicly

Because THIS is the summer you finally learn how to surf. And THESE are the board shorts you’ll be wearing when you catch the first wave.
http://jcrew.co/LVNG1 ... 6 hours ago

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**

- **RDFa, Microdata and Microformats**

- **Extracting HTML Table Data**

- **Wrapper Induction**

- **Knowledge Graphs**