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

- List of Data Catalogs
  - http://data.wu.ac.at/portalwatch/portalslist
Main Types of Shared Data

- **Public Sector Data**
  - Goal: Make data publicly accessible which has been generated by public sector institutions.
  - Laws in many western countries require institutions to publish data
  - Types of data: maps, population statistics, economic data, health data

- **Research Data**
  - Goal: Accelerate innovation by sharing research data and making research results reproducible
  - Institutional repositories, national research data infrastructures, topical portals

- **Commercial Data**
  - Goal: Earn money by collecting, cleaning, and integrating data
  - Types of data: data about consumers, business partners, locations
  - Examples of commercial providers: data.world, Foursquare
Example: Government Data Portal

Das Datenportal für Deutschland

Open Government: Verwaltungsdaten transparent, offen und frei nutzbar

62465 Datensätze
24 Anwendungen
106 Blogbeiträge

DCAT-AP.de Version 2.0 veröffentlicht

https://www.govdata.de/
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 domnierende Baueinheit im Kartenausschnitt ist der Oberrheingraben. Er durchzieht von Südsüdwest...

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: -
Example: Portal aggregating Metadata from other Portals

The official portal for European data

173 Catalogues  36 Countries  1,433,981 Datasets

Search datasets

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

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

![MADATA](https://madata.bib.uni-mannheim.de/)

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

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Product Datasets from the MWPD2020 Challenge at the ISWC2020 Conference (Task 1)</td>
</tr>
<tr>
<td>Alternative Title</td>
<td>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</td>
</tr>
<tr>
<td>Date</td>
<td>November 2020</td>
</tr>
<tr>
<td>Creator</td>
<td>Bizer, Christian ; Peeters, Ralph ; Primpl, Anna</td>
</tr>
<tr>
<td>Divisions</td>
<td>School of Business Informatics and Mathematics &gt; Wirtschaftsinformatik V (Bizer)</td>
</tr>
<tr>
<td>DDC Classification</td>
<td>004 Computer science, Internet</td>
</tr>
<tr>
<td>Keywords</td>
<td>schema org ; product matching ; entity matching ; identity resolution ; record linkage ; e-commerce</td>
</tr>
</tbody>
</table>

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 match”).

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

Example: Data Portal facilitating the Replication of Research

https://paperswithcode.com/datasets
The FAIR Data Principles

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

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

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

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

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

Crawls dataset metadata from the Web.

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 separate data spaces that might be partly accessible from the Web via
  - HTML interfaces
  - Web APIs
Example: Size of Facebook Social Graph

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

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

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

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

- ProgrammableWeb API Catalog
  - lists over 24,000 Web APIs
  - lists over 6,800 mashups
Most Popular APIs

https://www.programmableweb.com/news/which-are-developers-favorite-apis/research/2019/10/24
Mashups are based on a fixed set of data sources

Web APIs expose proprietary interfaces

→ Not index-able by generic web crawlers
→ No automatic discovery of additional data sources
→ No single global data space
Web APIs slice the Web into Data Silos
3. Alternative Approach: Linked Data

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

HTTP URIs take the role of global primary keys.

```
pd:cygri = http://richard.cyganiak.de/foaf.rdf#cygri
dbpedia:Berlin = http://dbpedia.org/resource/Berlin
```
URIs can be looked up on the Web

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

Chris Bizer

picture:

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)

Hands-on: How to get the Data?

- **Download the LOD-a-lot Dataset**
  - 28 billion RDF triples (524 GB zipped)
  - crawled from the public Web of Linked Data in 2017
  - [http://lod-a-lot.lod.labs.vu.nl/](http://lod-a-lot.lod.labs.vu.nl/)

- **Download the Billion Triples Challenge Dataset**
  - 4 billion RDF triples (52 GB gzipped, 1.1 TB uncompressed)
  - crawled from the public Web of Linked Data in 2014

- **Use SPARQL endpoints of individual data sets**
  - Endpoint list: [http://sparqles.ai.wu.ac.at/availability](http://sparqles.ai.wu.ac.at/availability)
4. HTML-embedded Data

1. Webpages traditionally contain structured data in the form of **HTML tables** as well as **template data**

2. More and more websites semantically markup the content of their HTML pages using **standardized markup formats**

Microformats

**Microdata**

**RDFa**

**JSON-LD**
4.1 Microformats

- Microformat effort dates back to 2003
- Small set of fixed formats
  - hcard: people, companies, organizations, and places
  - XFN: relationships between people
  - hCalendar: calendaring and events
  - hListing: small-ads; classifieds
  - hReview: reviews of products, businesses, events
- Shortcoming of Microformats
  - can not represent any kind of data.
- indexed by Google and Yahoo since 2009
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
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=""
    itemprop="http://schema.org/PostalAddress">
    <span itemprop="streetAddress">Parkring 12a</span>
    <span itemprop="addressLocality">Vienna</span>
  </span>
  <div itemprop="aggregateRating" itemscope=""
    itemprop="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>
```
问网站管理者自2011年以来注释数据以丰富搜索结果

- 675类型：事件、本地业务、产品、评论、工作机会

- 编码：Microdata, RDFa, JSON-LD
Usage of Schema.org Data @ Google

Data snippets within search results
Local businesses on maps
Data snippets within info boxes
Usage of Schema.org Data @ Google

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

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

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

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

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

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

http://webdatacommons.org/structureddata/2020-12/stats/schema_org_subsets.html
### 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>schema:Product/name</td>
<td>99 %</td>
</tr>
<tr>
<td>schema:Product/offers</td>
<td>94 %</td>
</tr>
<tr>
<td>schema:Offer/price</td>
<td>95 %</td>
</tr>
<tr>
<td>schema:Offer/priceCurrency</td>
<td>95 %</td>
</tr>
<tr>
<td>schema:Product/description</td>
<td>84 %</td>
</tr>
<tr>
<td>schema:Offer/availability</td>
<td>72 %</td>
</tr>
<tr>
<td>schema:Product/sku</td>
<td>56 %</td>
</tr>
<tr>
<td>schema:Product/brand</td>
<td>30 %</td>
</tr>
<tr>
<td>schema:Product/image</td>
<td>26 %</td>
</tr>
<tr>
<td>schema:Product/aggregateRating</td>
<td>17 %</td>
</tr>
<tr>
<td>schema:Product/mpn</td>
<td>6.3 %</td>
</tr>
<tr>
<td>schema:Product/productID</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...

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

<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>
</table>

- 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

<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>Dusseldorf</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bremen</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 150 INTERNATIONAL AFFAIRS

<table>
<thead>
<tr>
<th>Year</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>End Funding for the United Nations Development Program (UNDP)</td>
</tr>
<tr>
<td>2016</td>
<td>End Funding for the U.N. Intergovernmental Panel on Climate Change (IPCC)</td>
</tr>
<tr>
<td>2016</td>
<td>Eliminate the U.S. Trade and Development Agency (USTDA)</td>
</tr>
<tr>
<td>2016</td>
<td>Reform Food Aid Programs</td>
</tr>
<tr>
<td>2016</td>
<td>Eliminate Export-Import Bank</td>
</tr>
<tr>
<td>2016</td>
<td>Eliminate the Overseas Private Investment Corporation (OPIC)</td>
</tr>
<tr>
<td>2016</td>
<td>Eliminate Funding for the United Nations Population Fund (UNFPA)</td>
</tr>
</tbody>
</table>

#### 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
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. Lmfaoo Feat. Lauren Bennett And Goon Rock - Party Rock Anthem
13. Jackson, Michael - Billie Jean
14. DJ Casper - Cha Cha Slide
Types of Web Tables

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

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

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 1,377 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 Kwan 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

[Diagram showing DBpedia data at https://databus.dbpedia.org]

Previous Snapshot 2021-09

About

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

https://databus.dbpedia.org

[Link to DBpedia SNORQL at https://dbpedia.org/snorql]
Knowledge Graphs

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

- Google Knowledge Graph
  - development started 2012, builds on Freebase
  - 570 million objects described by over 18 billion facts (2012)
  - 1500 classes, 35,000 properties

- Microsoft Satori Knowledge Base
  - revealed to the public in mid-2013

- Yahoo Knowledge Graph
  - revealed to the public early-2014

- Knowledge Graphs employ RDF-style graph data models
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

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/LVNO1 ... 6 hours ago

People also search for

Banana Republic
Nordstrom
Anthropologie
Ann Inc.
Vineyard Vines

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

- Web data integration might involve millions of data sources
  - integration in corporate data lakes often remains below 20 sources

- The topics of the published data partly correlate with the publication methods used:
  - Data Portals: public sector data, statistical data, research data
  - Web APIs: user generated content, location-related data, weather data
  - Schema.org data: e-commerce, local business, event, job data
  - Linked Data: library data, research data, government data
  - Wikipedia, HTML tables: General knowledge

- The Web is the perfect playground for researching and applying Big Data Integration techniques
  - tough challenges concerning heterogeneity, volume, and data quality
  - rewarding if challenges can be handled, e.g. web-scale queries and mining
5. References

- Linked Data

- RDFa, Microdata and Microformats

- Extracting HTML Table Data

- Wrapper Induction

- Knowledge Graphs