

Vrije Universiteit Amsterdam
19/06/2018

Exploiting Semantics Where We Find Them

A Bottom-up Approach to the Semantic Web

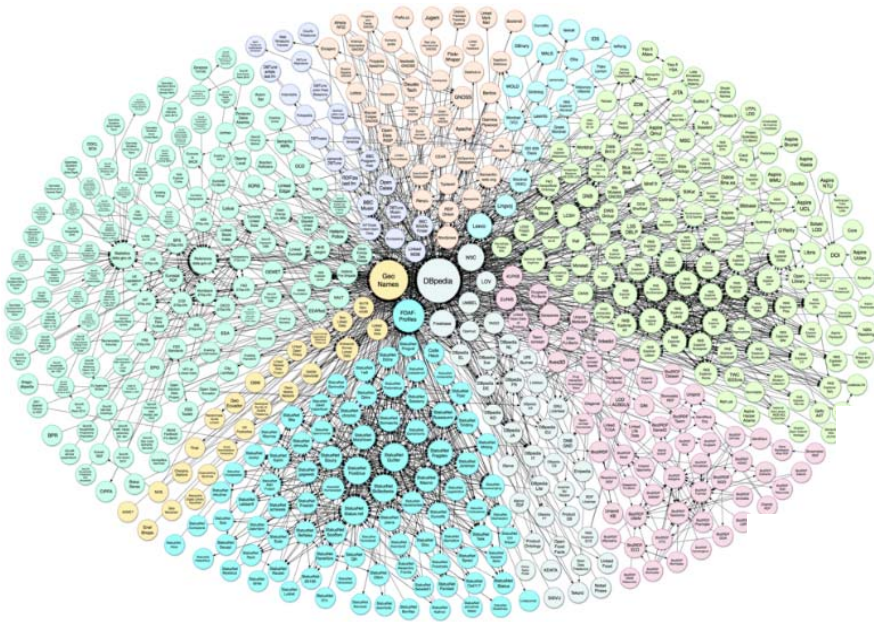
Prof. Dr. Christian Bizer

Outline

1. What does the Semantic Web actually look like?
 1. Linked Data
 2. HTML-embedded Data
2. Page- and Site-Structure as Semantic Clues
 1. HTML tables and lists
 2. Web site structure
3. What does this mean for Semantic Web applications?

1. What does the Semantic Web actually look like?

Linked Data



HTML-embedded Data



2.1 Linked Data Deployment

LOD Cloud 2014: 1091 datasets

Topics	# of Datasets	Percentage
Media	24	2 %
Government	199	18 %
Publications/Library	138	13 %
Geographic	27	2 %
Life Sciences	85	8 %
Cross-domain	47	4 %
User-generated Content	51	5 %
Social Networking	520	48 %

LODStats 2016: 2740 datasets

Schmachtenberg, Bizer, Paulheim: Adoption of the Linked Data Best Practices. ISWC2014.
Ermilov, Lehmann, Martin, Auer: LODStats: The Data Web Census Dataset. ISWC 2016.

Ontological Agreement

- Strong agreement on some vocabularies
 - base terminology, people, publications
- Proprietary vocabularies are used in addition to common ones, as data is often very specific

Prefix	Occurrence	Quota
rdf	996	98.22%
rdfs	736	72.58%
<u>foaf</u>	701	69.13%
dcterms	568	56.01%
owl	370	36.49%
wgs84	254	25.05%
sioc	179	17.65%
admin	157	15.48%
skos	143	14.11%
void	137	13.51%
<u>bio</u>	125	12.32%
<u>cube</u>	114	11.24%
rss	99	9.76%
odc	86	8.48%
w3con	77	7.60%
doap	65	6.41%
<u>bibo</u>	62	6.11%
<u>dcat</u>	59	5.82%

<http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/>
<https://lov.okfn.org/dataset/lov/>

RDF Links

- Some datasets put a lot of effort into linking
- Many datasets only link to a small number of other datasets or do not set RDF links at all

Link to	# of Datasets	Percentage
more than 10 datasets	79	8 %
6 to 10 datasets	81	8 %
5 datasets	31	3 %
4 datasets	42	4 %
3 datasets	54	5 %
2 datasets	106	10 %
1 datasets	176	17 %
0 datasets	445	44 %

} 16 %

} 71 %

<http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/>

Cool Applications

Prototypes of Semantic Web browsers and search engines.

Many of them are discontinued!

The image displays three screenshots of Semantic Web applications, illustrating prototypes of Semantic Web browsers and search engines.

Falcons: A search interface for the "Chicago" object. It shows a list of objects (1-10 of 63,109) and a sidebar with various types (Any type, Abstraction, Agent, etc.). The main content area displays information about "Chicago" and "Chicago Cubs players".

marbles: A browser interface for the "Tim Berners-Lee" object. It shows a list of objects (1-10 of 63,109) and a sidebar with various types (Any type, Abstraction, Agent, etc.). The main content area displays information about "Tim Berners-Lee" and "Tim Berners-Lee's FOAF file".

SIG.MA: A browser interface for the "Chris Bizer" object. It shows a list of objects (1-10 of 63,109) and a sidebar with various types (Any type, Abstraction, Agent, etc.). The main content area displays information about "Chris Bizer" and "Chris Bizer's FOAF file".

2.2 HTML-embedded Data

Microdata



RDFa



Microformats



JSON-LD



Overall Adoption 2017

Web Data Commons, 2017:

- 7.4 million pay-level-domains (PLDs) out of the 26 million PLDs provide HTML-embedded data (28%)
- 1.2 billion HTML pages out of the 3.2 billion pages provide HTML-embedded data (38.9%)

<http://webdatacommons.org/structureddata/2017-12/>

1000 times more PLDs compared to Linked Data

Widely-used Classes

Class	# of Domains WDC 2015	# of Domains schema.org 2015
schema:PostalAddress	124,000	> 1,000,000
schema:Product	108,000	> 1,000,000
schema:Offer	82,000	> 1,000,000
schema:LocalBusiness	77,000	500,000 - 1,000,000
schema:Person	74,000	> 1,000,000
schema:Review	28,000	250,000 - 500,000
schema:GeoCoordinates	17,000	100,000 - 250,000
schema:Event	12,000	100,000 - 250,000
schema:Hotel	5,300	10,000 - 50,000
schema:Restaurant	3,800	10,000 - 50,000
schema:JobPosting	3,600	10,000 - 50,000

<http://webdatacommons.org/structureddata/2015-11/>

Challenge: Flat Data Structures

Websites do not explicitly annotate product features but mention them in product names and descriptions.

```
<div itemtype="http://schema.org/Product">  
  <span itemprop="name">Apple MacBook Air A1370 Intel Core i5  
    1.60GHz 64GB SSD 11.6 Laptop  
  </span>  
  <span itemprop="description"> Catch up on work, school, or socializing  
    on the Apple MacBook Air A1370 11.6-inch laptop. This handy  
    computer features 2GB DDR3 RAM, an Intel Core i5 560UM  
    processor, 64GB hard drive, and the Mac OS ...  
  </span>  
</div>
```

Petrovski, Bryl, Bizer: Integrating Product Data from Websites Offering Microdata Markup. DEOS 2014.

Challenge: Small Amount of Identifiers and Data Links

1. Small amount of product identifiers

Properties	PLDs	
	#	%
schema:Product/productID	9,221	10 %
schema:Product/gtin13	935	1 %

2. Hardly any schema:sameAs links

Definition: URL of a reference Web page that unambiguously indicates the item's identity.

Properties	PLDs	
	#	%
schema:Product/sameAs	85	0.07 %
schema:LocalBusiness/sameAs	655	0.8 %
schema:Organization/sameAs	3,900	5 %

<http://webdatacommons.org/structureddata/2015-11/>

Cool Applications

1. Rich snippets within search results

Hotel Innsbruck (Austria) - Hotel Reviews - TripAdvisor
www.tripadvisor.com > ... > Tirol > Innsbruck > Innsbruck Hotels ▾
★★★★★ Rating: 4 - 628 reviews - Price range: \$\$
Hotel Innsbruck, Innsbruck: See 628 traveler reviews, 261 candid photos, and great deals for Hotel Innsbruck, ranked #9 of 80 hotels in Innsbruck and rated 4 of ...

2. Knowledge graph panels

<https://developers.google.com/structured-data/>

Starbucks
Coffee company

Starbucks Corporation, doing business as Starbucks Coffee, is an American global coffee company and coffeehouse chain based in Seattle, Washington. [Wikipedia](#)

Stock price: SBUX (NASDAQ)
\$97.81 +0.44 (+0.45%)
Mar 24, 3:28 PM EDT - Disclaimer

Customer service: 1 (800) 782-7282

CEO: Howard Schultz

Founded: March 30, 1971, Pike Place Market, Seattle, WA

Headquarters: Seattle, WA

Founders: Gordon Bowker, Jerry Baldwin, Zev Siegl

Profiles

Facebook Twitter Instagram LinkedIn Google+



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 Schneßel



Linked Data vs. HTML-embedded Data

Expectation	Linked Data	HTML-embedded Data
1. People publish structured data	> 1000 sources, wide range of specific topics	Millions of sources, focused on search engines and Facebook
2. Ontologies enable understanding	Partial agreement, complex data structures	Strong agreement, flat data structures
3. Hyperlinks on data level	Some data links	Hardly any data links
4. High quality content	Web quality, partly outdated	Web quality, some SPAM
5. Cool applications	various application prototypes	strong application pull by search engines

2. HTML Page- and Site-Structure as Semantic Clues

Web Data Commons – Web Tables Corpus

233 million relational Web tables from 15 million pay-level-domains (PLDs)

■ 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

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

1.74 billion rows

253,000,000 different subject labels

Template-Structure of HTML Pages

- Most websites are generated from databases using HTML-templates.
- Wrapper induction
 - Supervised learning
- Problem:
 - Wrappers are site-specific
 - Large scale extraction noisy

The screenshot shows the Amazon product page for a Samsung UN40EH5300 40-Inch 1080p 60Hz LED HDTV, Black. The page layout includes a top navigation bar with the Amazon logo, a search bar, and links for 'Bizer's Amazon.com', 'Today's Deals', 'Gift Cards', 'Sell', and 'Help'. Below the navigation bar, there's a 'Shop by Department' section with 'Electronics' selected. The main product image shows the TV displaying a 'SMART HUB' interface. To the right of the image, the product title is 'Samsung UN40EH5300 40-Inch 1080p 60Hz LED HDTV, Black' by Samsung, with a 4.5-star rating and 786 customer reviews. The list price is \$799.99, and the current price is 'To see our price, add this item to your cart. You can always remove it later'. The product is 'In Stock' and ships from and sold by Amazon.com. A countdown timer indicates it will be available on Thursday, Aug. 22, with 7 hours and 43 minutes remaining. Below the image, there are links to 'Click for larger image and other views', 'See 1 customer image', and 'Share your own customer images'. The 'Product Features and Technical Details' section lists features like Smart TV, Signature Services, Smart Hub, Built-in Wi-Fi, TV with stand (36.5-Inch x 23.9-Inch x 9.8-Inch), AllShare Play, Clear Motion Rate: 120, and Built-in WiFi. Technical details include Brand Name: Samsung, Model: UN40EH5300, Display Technology: LED-lit, Display Size: 40 inches, Image Aspect Ratio: 16:09, Height: 23.90 inches, and Width: 36.50 inches.

We will keep on seeing similar adoption patterns, as we need to be realistic about the effort spent by data publishers.

- Some providers spend a lot of effort in publishing their data.
- Most providers are not willing to spend a lot of effort.
 - Small amounts of links and shared identifiers
 - Flat data structures
 - Heterogeneity of vocabularies
- The wider the coverage we want, the less structured the content we get

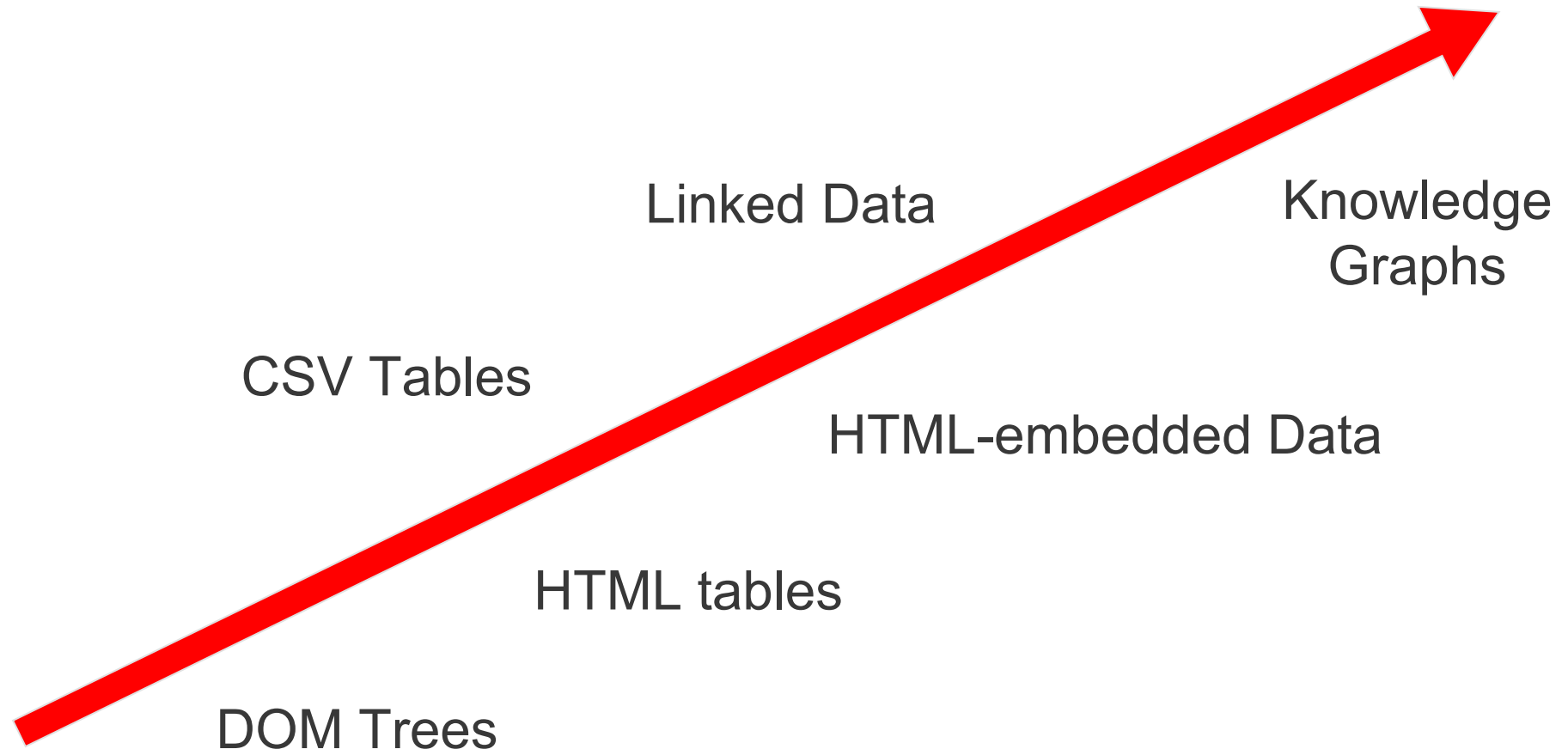
3. What does this mean for Semantic Web Applications?

**Be happy about all semantic clues
(integration hints) provided**

But do not expect the clues to be perfect

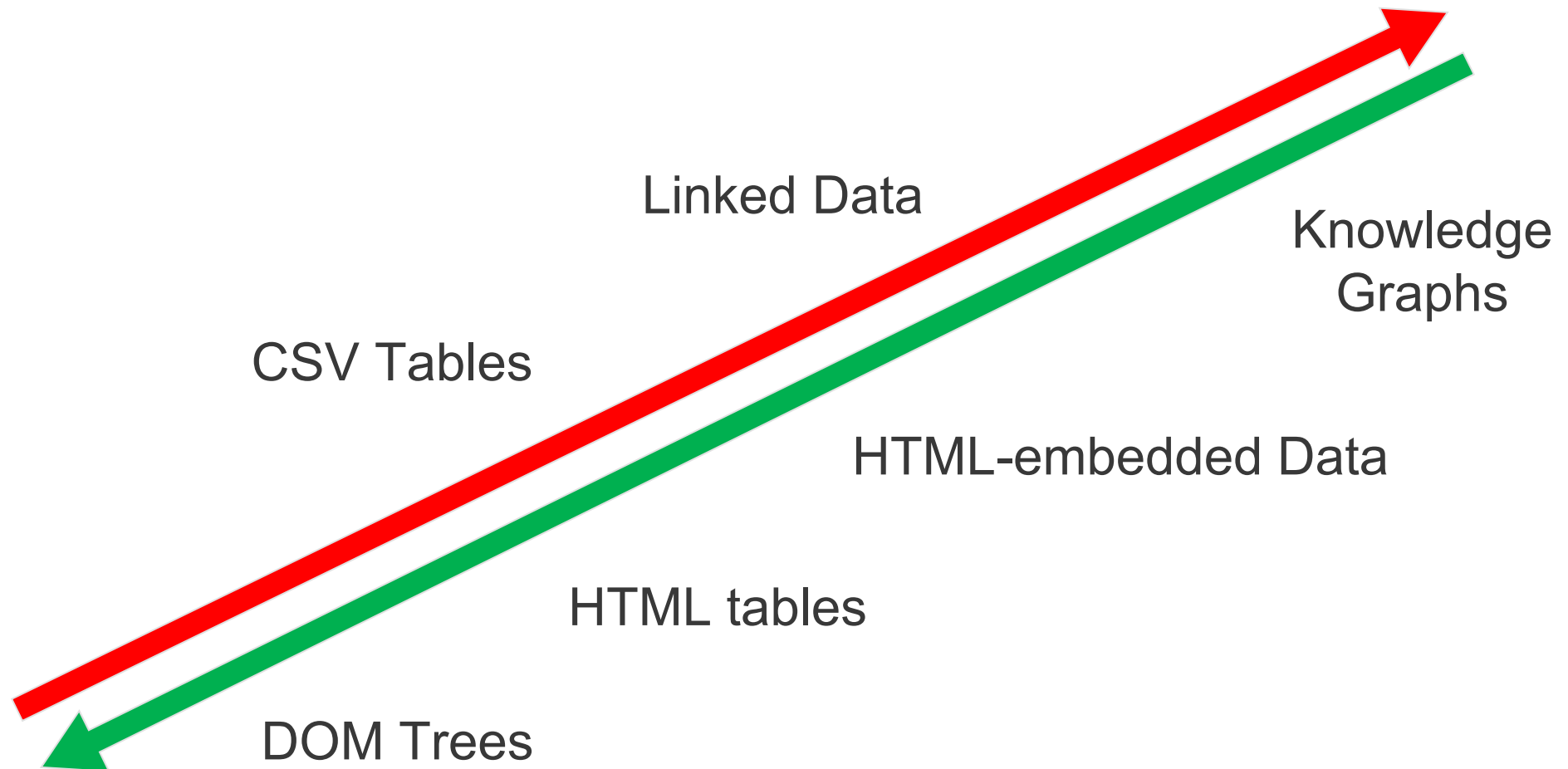
The Structural Continuum

**Be open to different forms of
“structured” web content.**



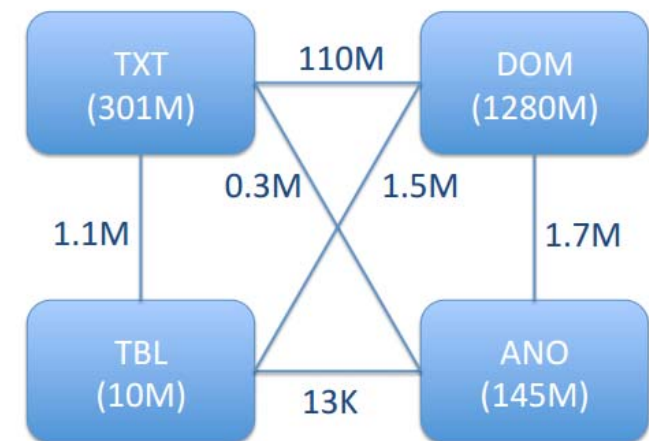
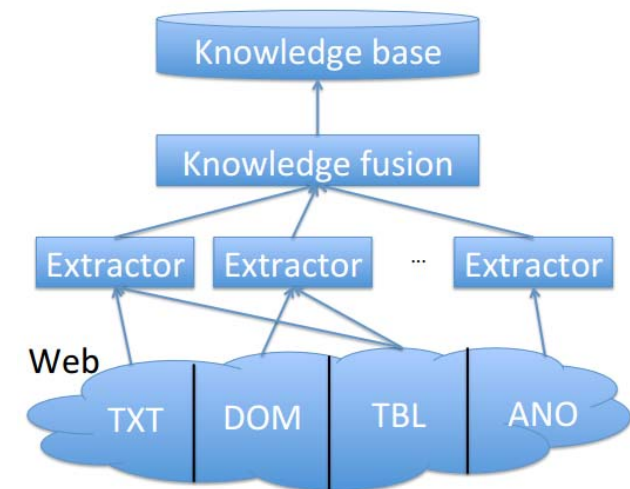
Exploiting the Structural Continuum

Use the small amounts of clean content to understand the larger amounts of less structured content.



Google Knowledge Vault

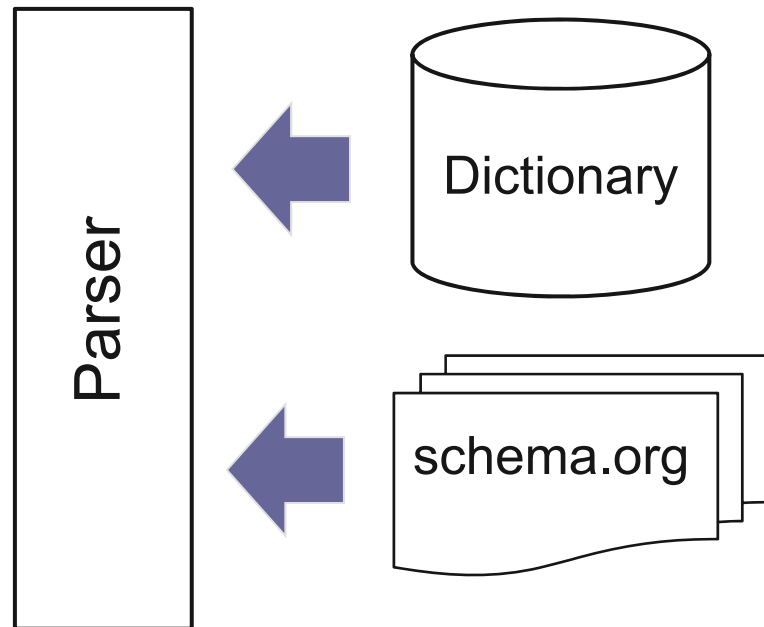
- Extends Freebase with data from one billion web pages
 1. Web text (TXT): Entity linking, relationship extraction
 2. HTML trees (DOM): Wrapper induction
 3. HTML tables (TBL): Relational tables
 4. Semantic Annotations (ANO): schema.org, OGP
- Employs knowledge-based trust for data fusion
- Results:
 - 271 million facts with confidence >90%
 - 90 million facts not in Freebase before



Dong, et al.: Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. SIGKDD 2014.

Parsing Flat Descriptions

New Samsung Galaxy S4 GT-1950S 16GB 5.0 inches Android
Smartphone with 2-Year Sprint Contract - White Frost



schema.org data suitable as distant supervision:

- schema:Product/brand
- schema:Product/manufacture
- schema:JobPosting/industry
- Schema:JobPosting/skills
- schema:Event/name

Ristoski, Mika: Enriching Product Ads with Metadata from HTML Annotations. ESWC 2016.
Foley, et al.: Learning to Extract Local Events from the Web. SIGIR 2015.

Exploit Schema.org and HTML Tables Together

s:breadcrumb

s:name

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

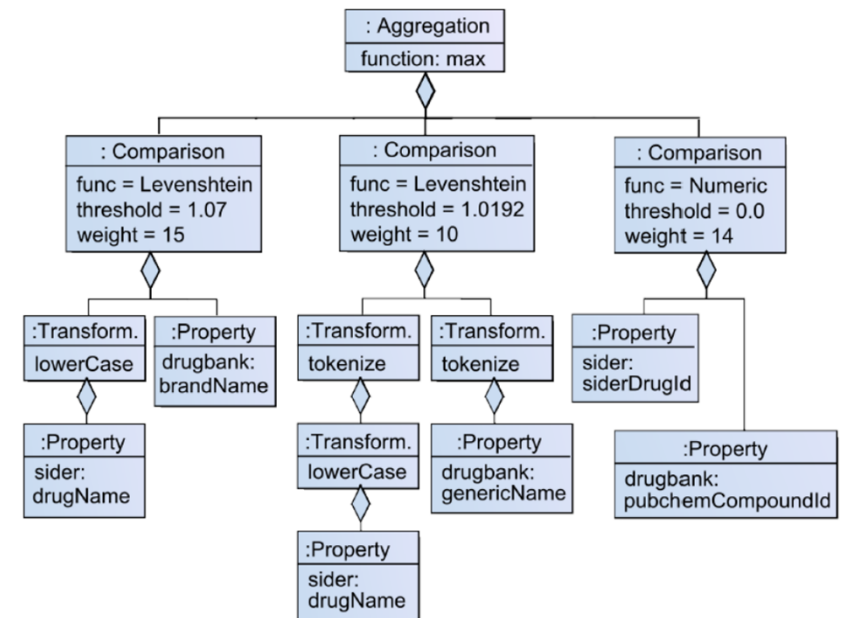
Qty: 1 **Add to Cart**
Add to List Add to Registry

s:price

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.

Create and Verify Data Links

- Supervised learning of detailed matching rules leads to $F1 > 95\%$ (e.g. Silk and LIMES frameworks)
- Sources of distant supervision
 1. RDF links
 - owl:sameAs
 2. shared identifiers
 - schema:Product/productID
 - schema:Product/gtin13



- How good can we get in large-scale settings using this supervision?
- How to generalize matching rules to data from multiple sources?

Qiu, Srivastava, et al.: Big Data Linkage for Product Specification Pages. SIGMOD 2018.

Stonebraker, et al.: Data Curation at Scale: The Data Tamer System. CIDR 2013.

Conclusions

The Semantic Web is Huge

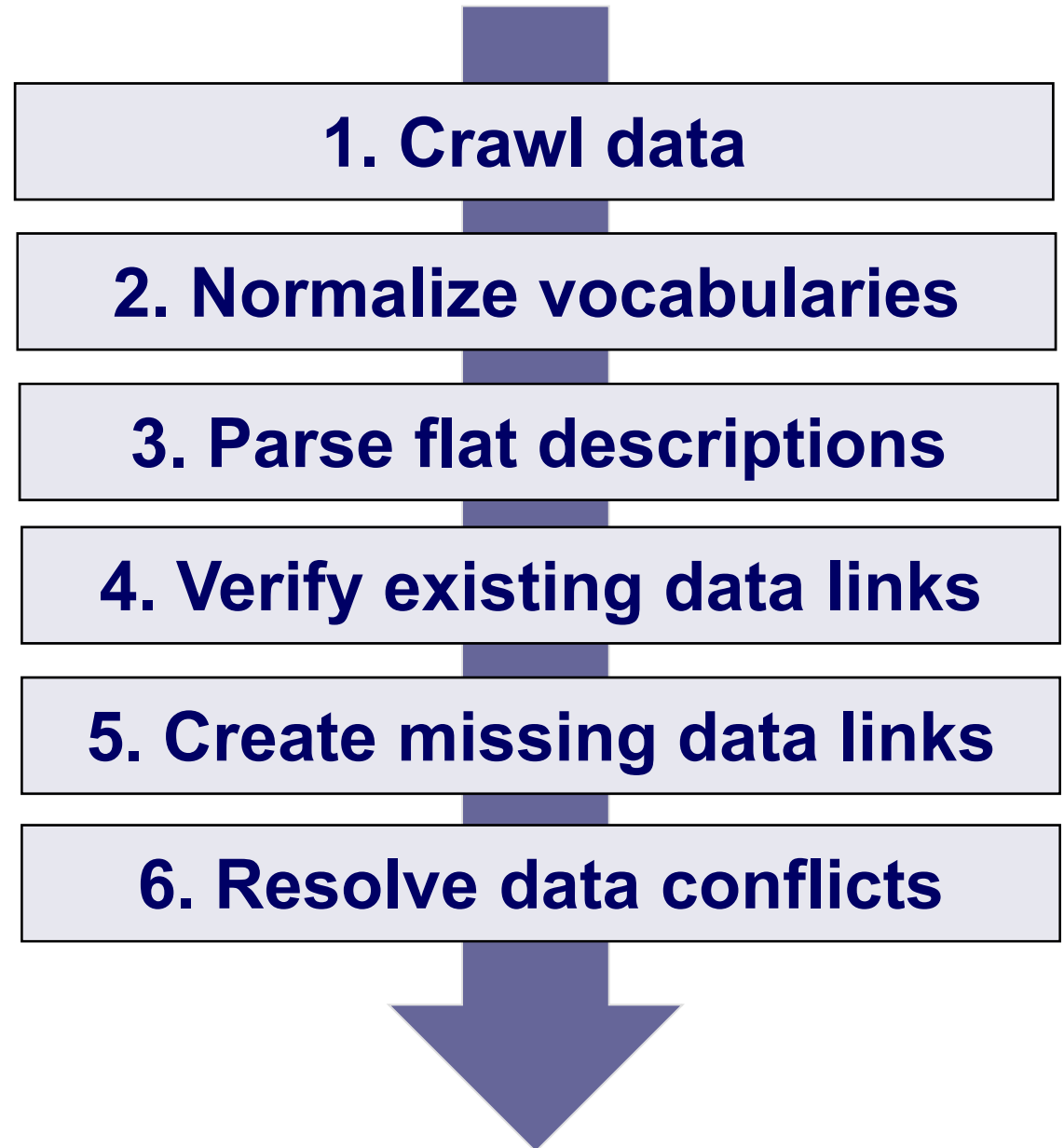
- The Semantic Web contains more data than most people like



- exciting test-bed for research on data profiling, cleansing and integration
- Large data pool for commercial applications (product comparison, business listings, job search, ...)

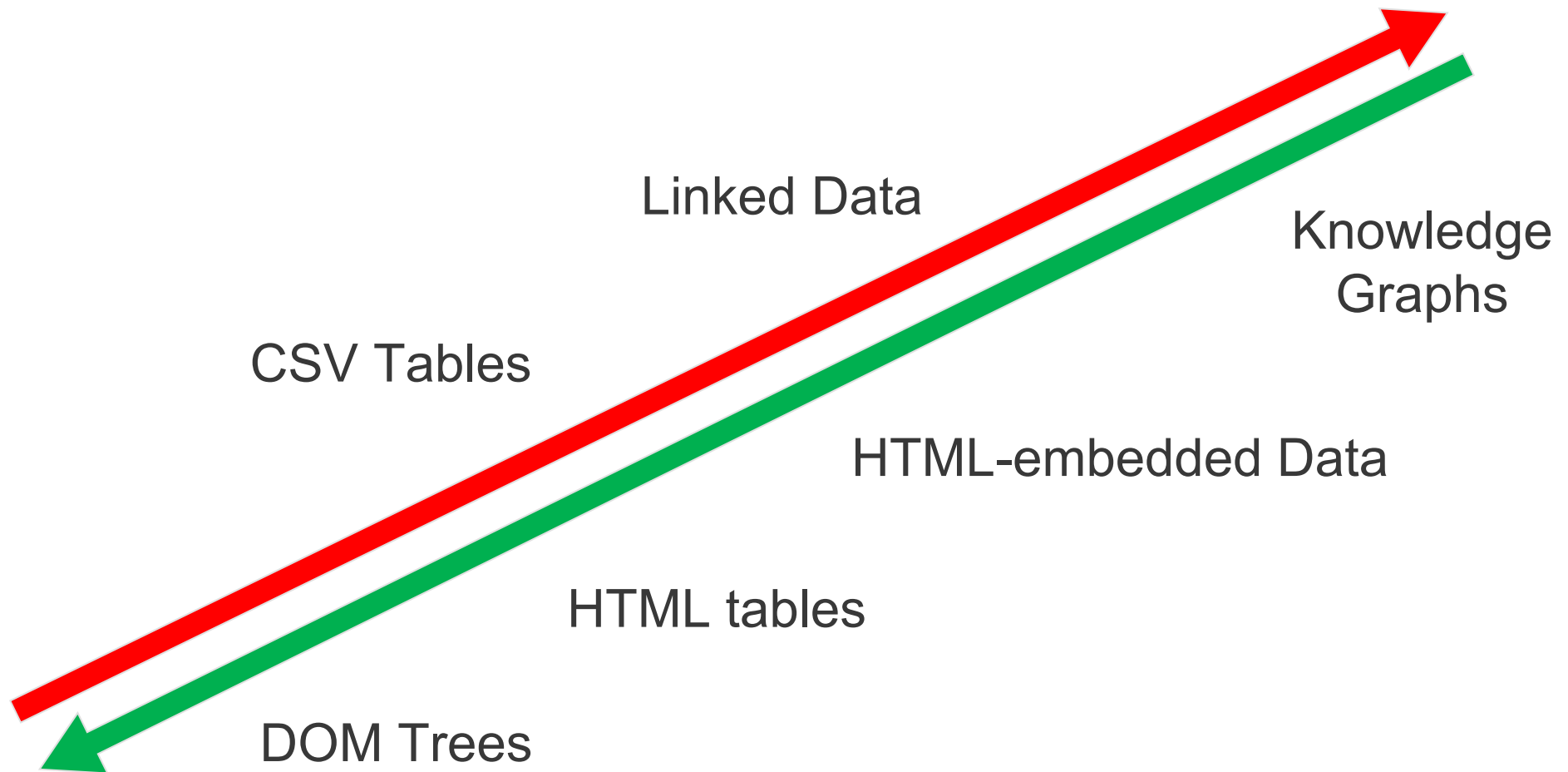
Semantic Web Clients need to be FAT Clients

There are no shortcuts!



Challenge and Opportunity: The Structural Continuum

Use the small amounts of clean content to understand the larger amounts of less structured content.



Thank you
