15th International Semantic Web Conference (ISWC 2016)
Kobe, Japan, 10/20/2016

Is the Semantic Web what we Expected?

Deployment Patterns and Data-driven Challenges

Prof. Dr. Christian Bizer

UNIVERSITÄT
MANNHEIM

Ian's Keynote Last Year in Bethlehem

Build it, and they will come:
Applications of semantic technology

Ian Horrocks
Information Systems Group
Department of Computer Science
University of Oxford

http://videolectures.net/
iswc2015_horrocks_semantic_technology/

This Year

Use cases

- on the public Web

- many data sources

- no central control

Outline

1. What did we expect the Semantic Web to be?

2. What does the Semantic Web actually look like?

    1. Linked Data

    2. HTML-embedded Data

3. Why is this the case?

4. What does this mean for Semantic Web applications?

1. What did we expect the Semantic Web to be?

2001 Article: The Semantic Web

Envisions three things to happen:

– people publish structured data on the Web

– ontologies are used to enable shared understanding

– people implement cool applications that do smart things with the available data

Tim Berners-Lee, James Hendler, Ora Lassila:
The Semantic Web. Scientific American, May 2001.

Expectation: Hyperlinks are Set on Data Level

Tim Berners-Lee
Date: 2006-07-27, last change: $Date: 2009/06/18 18:24:33 $
Status: personal view only. Editing status: imperfect but published.
Up to Design Issues

Linked Data

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF,. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:

1. Use URIs as names for things

2. Use HTTP URIs so that people can look up those names.

3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)

4. Include links to other URIs. so that they can discover more things.

https://www.w3.org/DesignIssues/LinkedData.html

https://www.w3.org/History/1989/proposal.html

Information Management: A Proposal

Tim Berners-Lee, CERN
March 1989, May 1990

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext system.
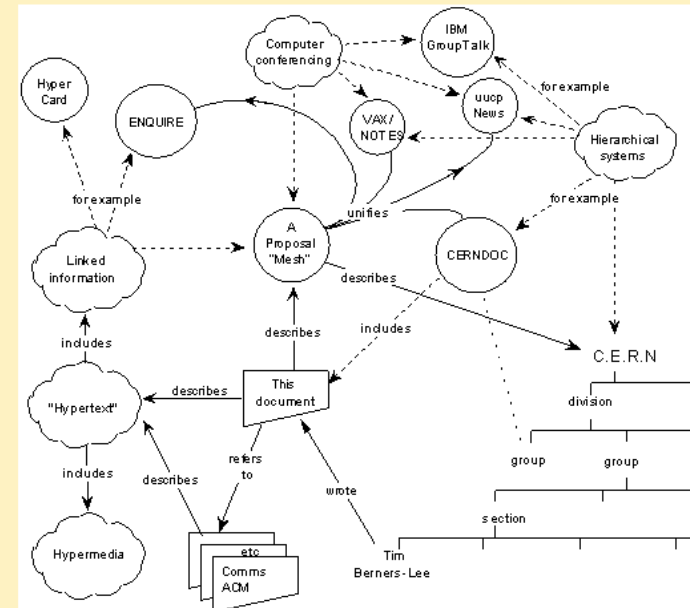
Overview

Many of the discussions of the future at CERN and the LHC era end with the question - ªYes, but how will we ever keep track of such a large project?º This proposal provides an answer to such questions. Firstly, it discusses the problem of information access at CERN. Then, it introduces the idea of linked

Expectation: High Quality Content / Provenance Metadata

- Publishers provide high quality content

- Publishers support applications in determining trustworthiness
  - by providing provenance metadata
 - using digital signatures
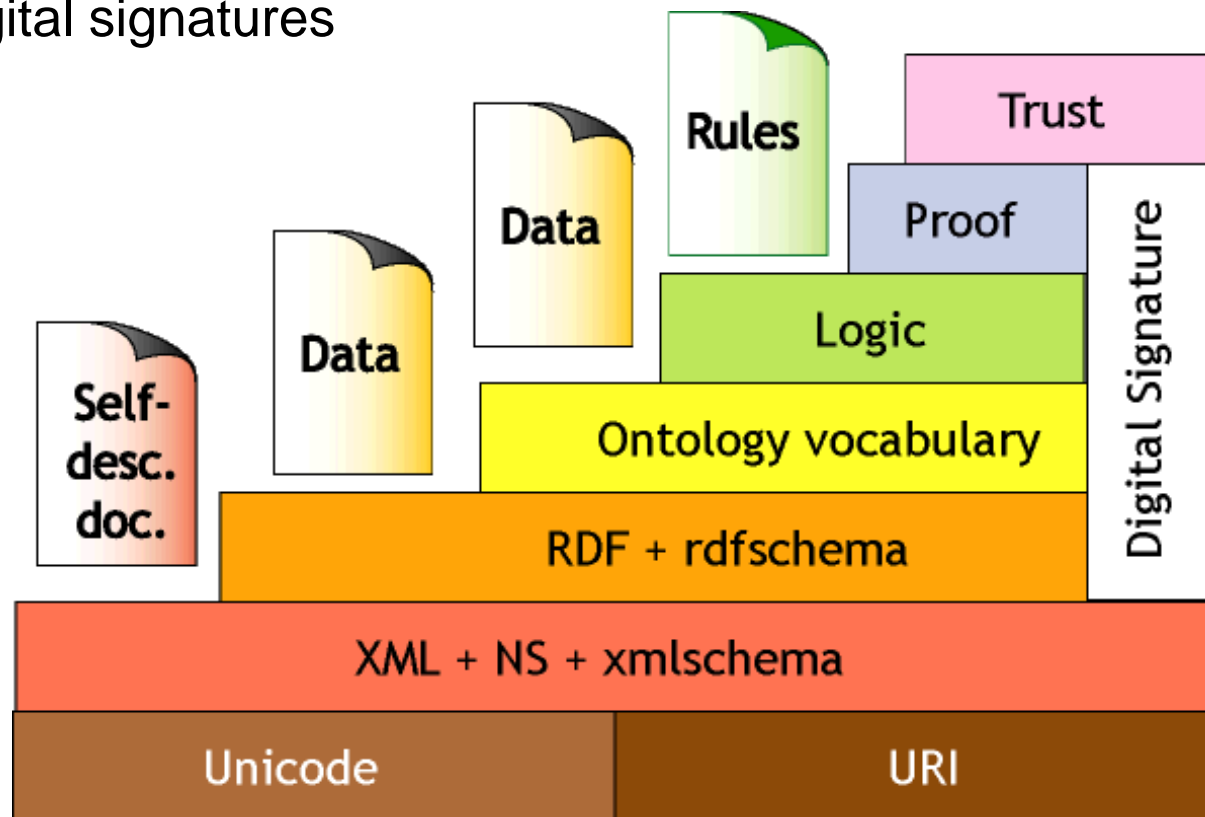
Layer Cake, 2001

Check List: Our Expectations about the Semantic Web

1. People publish structured data on the Web

2. Ontologies are used to enable shared understanding

3. Hyperlinks are set on data level

4. People publish high quality content / metadata

5. Cool applications do smart things with the data

2. What does the Semantic Web actually look like?

Linked Data
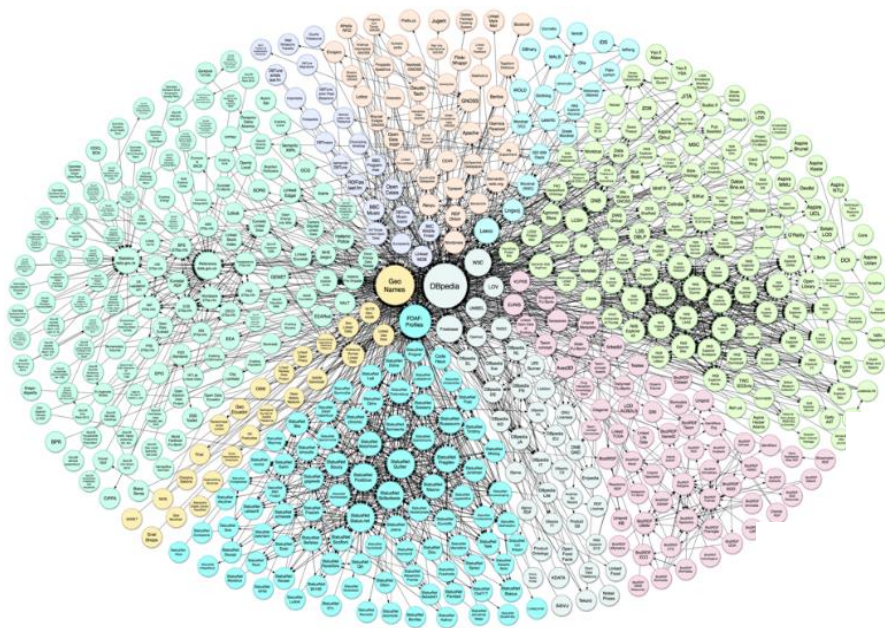
HTML-embedded Data

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 %
Unser-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%
foaf	701	69.13%
dcterm	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%
bio	125	12.32%
cube	114	11.24%
rss	99	9.76%
odc	86	8.48%
w3con	77	7.60%
doap	65	6.41%
bibo	62	6.11%
dcat	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.

RDFa

Microdata

Microformats

JSON-LD

Overall Adoption 2015

Web Data Commons, 2015:

– **2.72 million pay-level-domains (PLDs) out of the 14.41 million PLDs provide HTML-embedded data (19%)**

– **540 million HTML pages out of the 1.7 billion pages provide HTML-embedded data (30%)**

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

Guha/Brickley/Macbeth, 2015:
12 million websites provide schema.org data

Guha/Brickley/Macbeth: Schema.org. ACM queue, 2015

Number of PLDs providing HTML-embedded Data

Legend:
- RDFa (blue)
- Microdata (red)
- hCard (yellow)
- JSON-LD (green)

Y-axis: 1.5M, 1M, 500K, 0

X-axis: 2012, 2013, 2014, 2015

http://webdatacommons.org/structureddata/
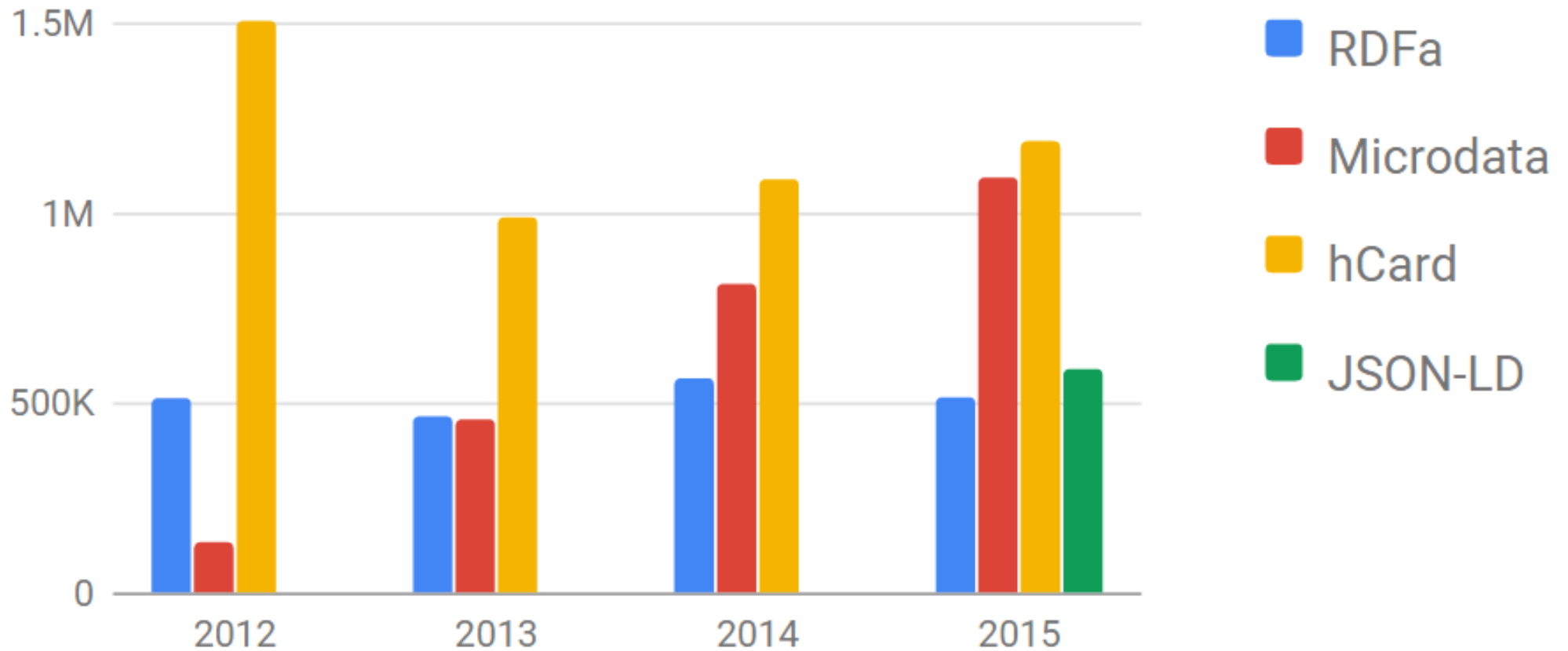
Widely-used Classes

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/

Adoption by E-Commerce Websites

Alexa Top-15 Website	schema:Product
Amazon.com	☒
Ebay.com	✓
NetFlix.com	☒
Amazon.co.uk	☒
Walmart.com	✓
etsy.com	☒
Ikea.com	✓
Bestbuy.com	✓
Homedepot.com	✓
Target.com	✓
Groupon.com	☒
Newegg.com	✓
Lowes.com	☒
Macys.com	✓
Nordstrom.com	✓

Adoption:

60 %

Properties used to Describe Products

Properties	PLDs	
	#	%
schema:Product/name	78,292	87 %
schema:Product/image	59,445	66 %
schema:Product/description	58,228	65 %
schema:Product/offers	57,633	64 %
schema:Offer/price	54,290	61 %
schema:Offer/availability	36,789	41 %
schema:Offer/priceCurrency	30,610	34 %
schema:Product/url	23,723	26 %
schema:Product/aggregateRating	21,166	24 %
schema:Product/manufacturer	10,150	11 %
schema:Product/brand	9,739	11 %
schema:Product/productID	9,221	10 %
schema:Product/sku	7955	9 %
schema:Product/gtin13	935	1 %

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/

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: Product Categorization

1. Small amount of websites publishing categorization information

Properties	PLDs	
	#	%
schema:Offer/category	2200	2 %
schema:WebPage/breadcrumb	460	0.4 %

2. Heterogeneity of the product taxonomies

Home > Shop > Outdoor & Garden > Barbecues & Outdoor Living > Garden Furniture > Tables > Dining Tables

Philadelphia Eagles > Philadelphia Eagles Mens > Philadelphia Eagles Mens Jerseys > over $60

Meusel, Primpeli, Meilicke, Paulheim, Bizer: Exploiting Microdata Annotations to Consistently Categorize Product Offers at Web Scale. EC-Web 2015.

Adoption by Travel Websites

Top 15 Travel Websites	schema:Hotel
Booking.com	✓
TripAdvisor	✓
Expedia	✓
Agoda	✓
Hotels.com (uses OGP)	☒
Kayak	✓
Priceline	✓
Travelocity	✓
Orbitz	✓
ChoiceHotels	✓
HolidayCheck	✓
ChoiceHotels	✓
InterContinental Hotels Group	✓
Marriott International	✓
Global Hyatt Corp.	☒

Adoption:
86 %

Kärle, Fensel, Toma, Fensel: Why Are There More Hotels in Tyrol than in Austria? Analyzing Schema. org Usage in the Hotel Domain. ICTT 2016.

Properties used to Describe Hotels

Top 10 Properties	PLDs	
	#	%
schema:Hotel/name	4173	88,35 %
schema:Hotel/address	3311	70,10 %
schema:Hotel/telephone	2488	52,68 %
schema:PostalAddress/streetAddress	2362	50,01 %
schema:PostalAddress/addressLocality	2231	47,24 %
schema:Hotel/url	2102	44,51 %
schema:PostalAddress/postalCode	2096	44,38 %
schema:AggregateRating/ratingValue	1952	41,33 %
schema:Hotel/aggregateRating	1866	39,51 %
schema:AggregateRating/bestRating	1697	35,93 %

Might improve in the future as new schema.org accommodation vocabulary was released August 2016.

Adoption by Job Portals

Top-10 Employment Websites	schema:JobPosting
Indeed.com	✓
Monster.com	✓
Careerbuilder.com	✓
Snagajob.com	✓
Jobsdb.com	✓
Jobsearch.about.com	☒
Jobs.net	✓
Internships.com	✓
Jobs.aol.com	☒
Quintcareers.com	☒

Adoption:
70 %

Properties used to Describe Job Postings

Properties	PLDs	
	#	%
JobPosting/title	2588	91 %
JobPosting/hiringOrganization	1412	49 %
JobPosting/description	1192	41 %
JobPosting/jobLocation	1062	37 %
Organization/name	862	30 %
JobPosting/datePosted	793	27 %
Place/address	471	16 %
JobPosting/baseSalary	227	8 %
JobPosting/industry	209	7 %
JobPosting/educationRequirements	145	5 %
JobPosting/occupationalCategory	105	0.3 %
JobPosting/skills	56	0.2 %

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 Naturalsound Records
Awards: Grammy Award for Best Rock Album, more

Upcoming events

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

Cool Applications

- Open Graph Protocol allows site owners to determine how entities are displayed in Facebook

- uses RDFa for marking up data in HTML pages

- used by over 200,000 websites (WDC 2015)

Our Expectations Revisited

| 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

3. Why is this the Case?

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

Benefits and Costs for Data Providers

Making the Web a better place isn't enough motivation for most data providers.

Benefits of Publishing HTML-Embedded Data

Get richer visibility in search results and potentially more clicks.
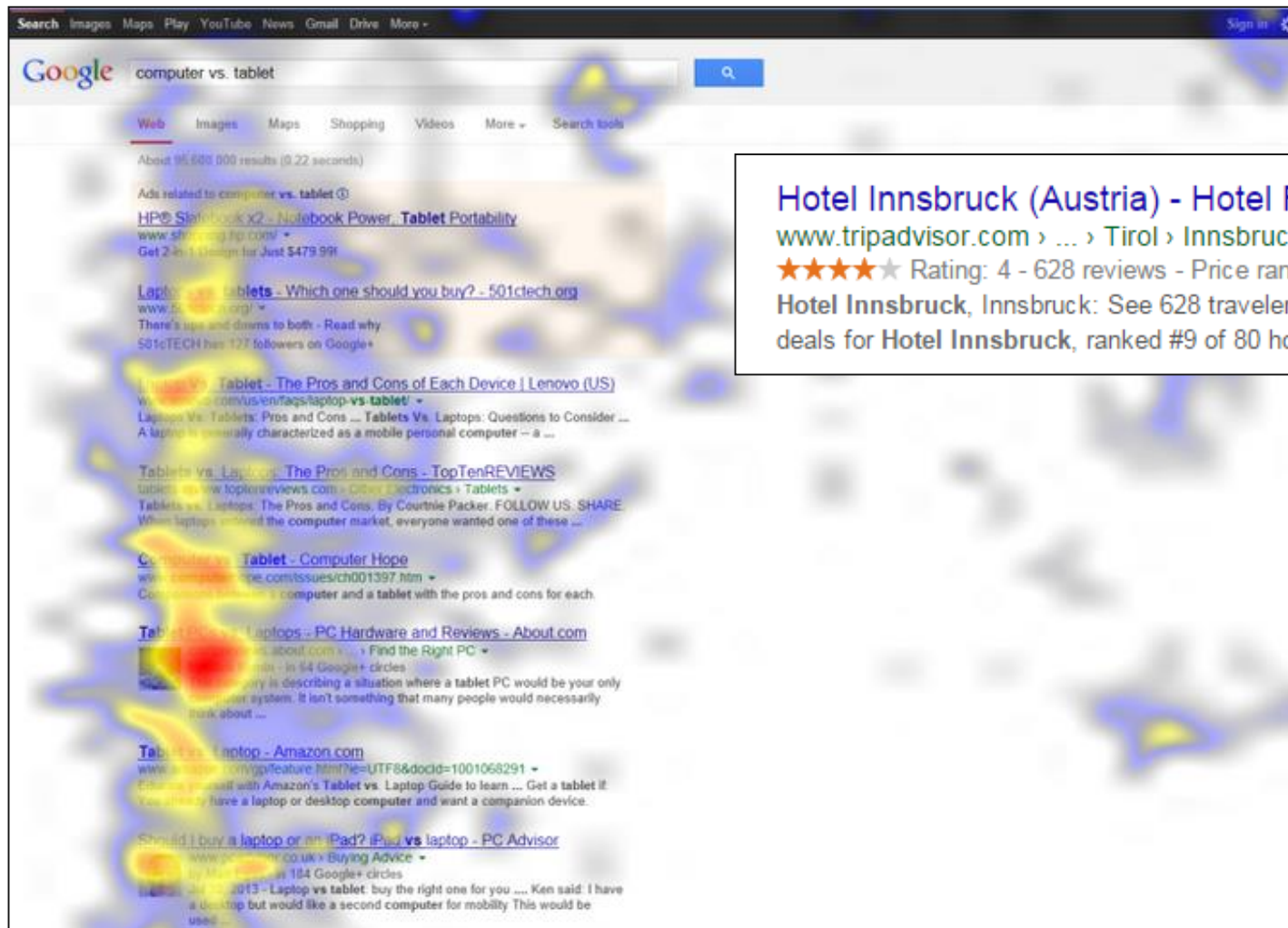
Effort for Publishing HTML-Embedded Data

– Most providers just change their HTML templates

```
<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>
```

– If more effort is required, **most providers just do not do it**

  1. Annotate specific product features
     - given free text descriptions in the backend database
  2. Map to common product taxonomy like GS1 GPC
     - given local categorizations
  3. Annotate skills and occupational categories
     - given free text descriptions in the backend database

Effort for Setting Data Links

<http://dbpedia.org/resource/Berlin> owl:sameAs
<http://sws.geonames.org/2950159> .

- Effort:

  1. Decide which data sources to link to

  2. Compare schemata and develop a matching rule for each class

  3. Run link generation algorithm

 4. Publish resulting link set on the Web

- Benefits:
  - You increase the value of your data as it becomes easier to use it together with data from other sources
  - You reduce the integration costs for the data consumer

Data publisher
provides links

Effort
Distribution

Data consumer
generates links
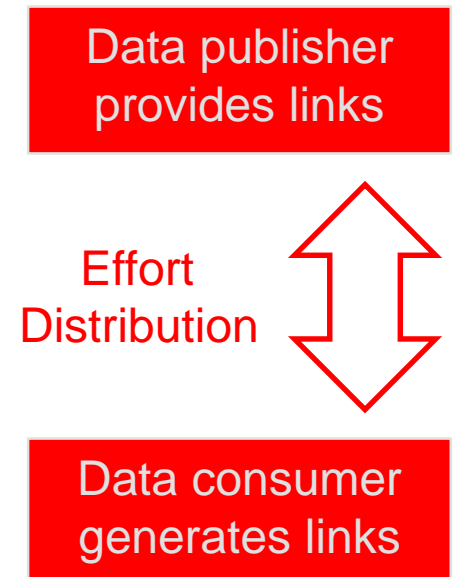
For Whom does the Linking Effort pay off?

– Scientists
  - Innovation becomes possible by connecting datasets
  - My impact / prestige grows if my data is used for cool things
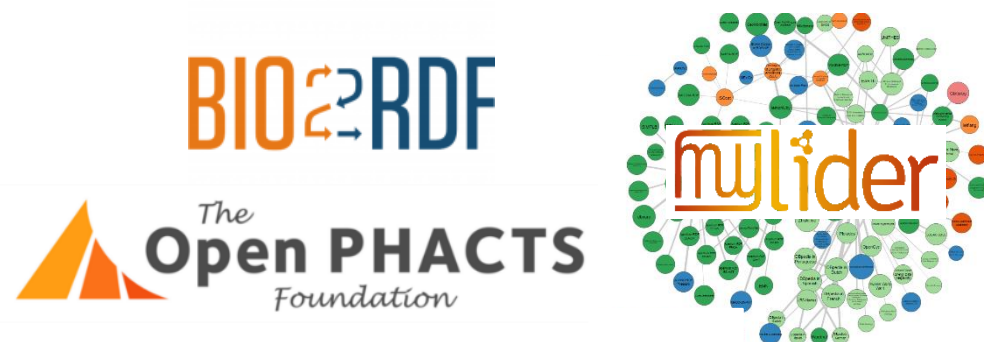
– Librarians
  - Have the mission to catalog artefacts
  - Traditionally use shared identifiers

– E-Commerce Vendors
  - Benefits of setting data links are unclear
  - Just want to look nice on Google
  - Might not want to be comparable on price portals

Effort of Maintaining Links

- We want to be nice!
  - we want to link to everybody

- We set instance- and schema-level links!
  - created and collected 37 link sets
  - over 20 million RDF links
  - http://wiki.dbpedia.org/services-resources/interlinking
    https://github.com/dbpedia/links

- We would likely need a full-time volunteer to maintain all these links

- Result: Many dead links
  1. because target data source has changed
  2. because we used bad linkage rules due to insufficient domain knowledge

Hypothesis

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

- Missing links and shared identifiers

- Flat data structures

- Heterogeneity of taxonomies

- Mixed data quality

4. What does this mean for Semantic Web Applications?

4. 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

Applications should be happy about …

… all effort that data providers put into setting data links

- but treat links with caution as they might be wrong / outdated

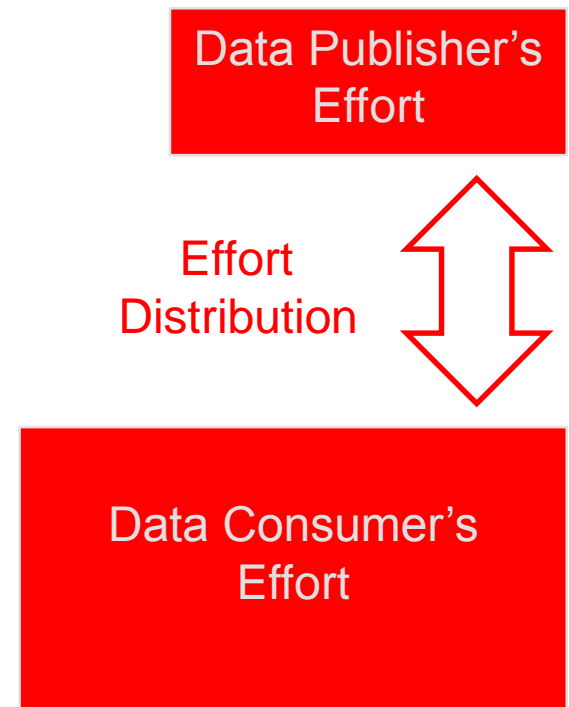… all effort that data providers put into using common vocabularies

- but still try to understand proprietary vocabularies / taxonomies

… all effort that data providers put into structuring their data
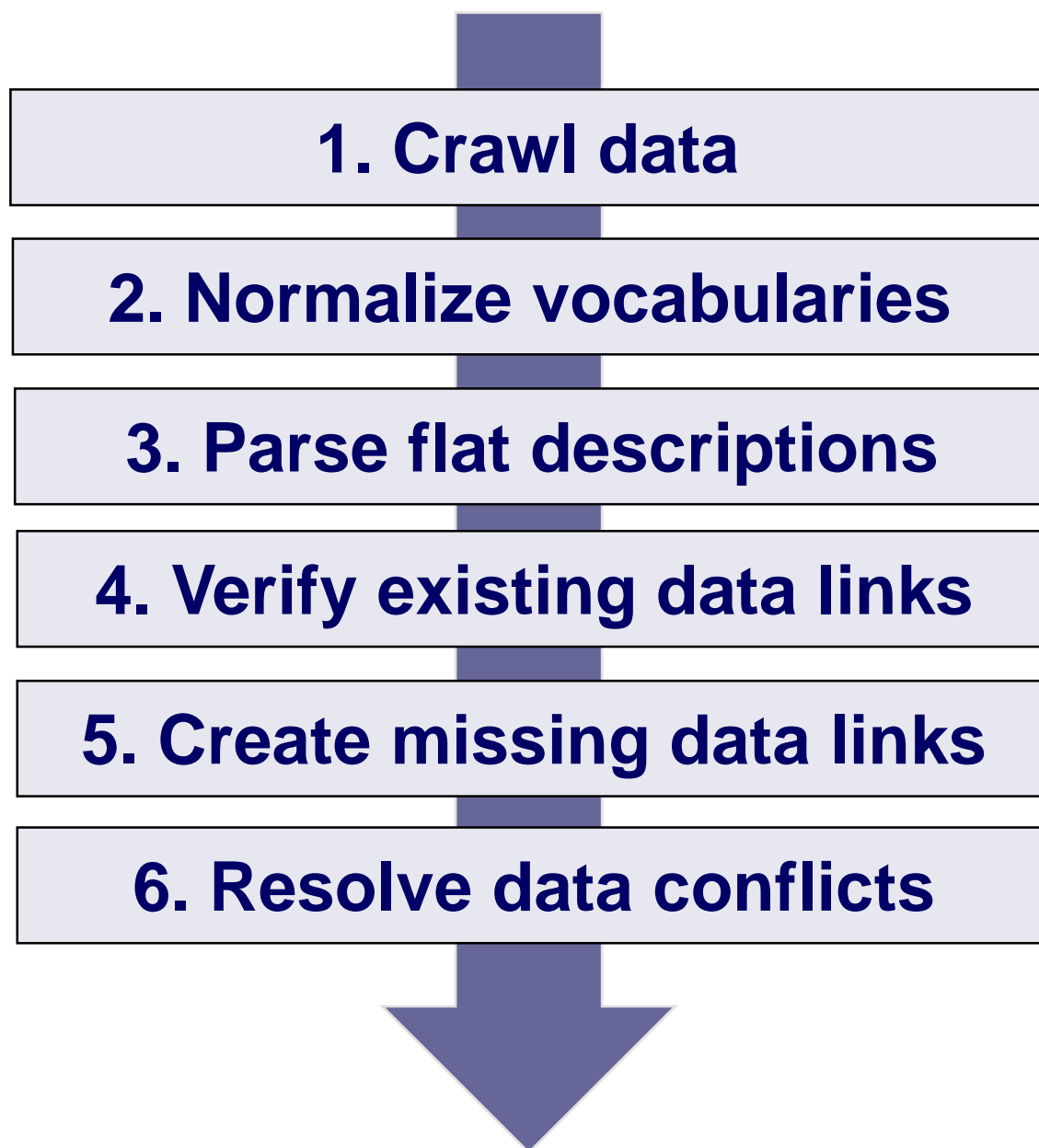
- but still try to understand flat free-text descriptions

Treat all statements on the Web as a claims

- whose trustworthiness needs to be verified

Data Publisher's Effort

Effort Distribution

Data Consumer's Effort

Semantic Web Clients need to be FAT Clients

There are no shortcuts!

1. Crawl data
2. Normalize vocabularies
3. Parse flat descriptions
4. Verify existing data links
5. Create missing data links
6. Resolve data conflicts

Parsing Flat Descriptions

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

Parser

Dictionary
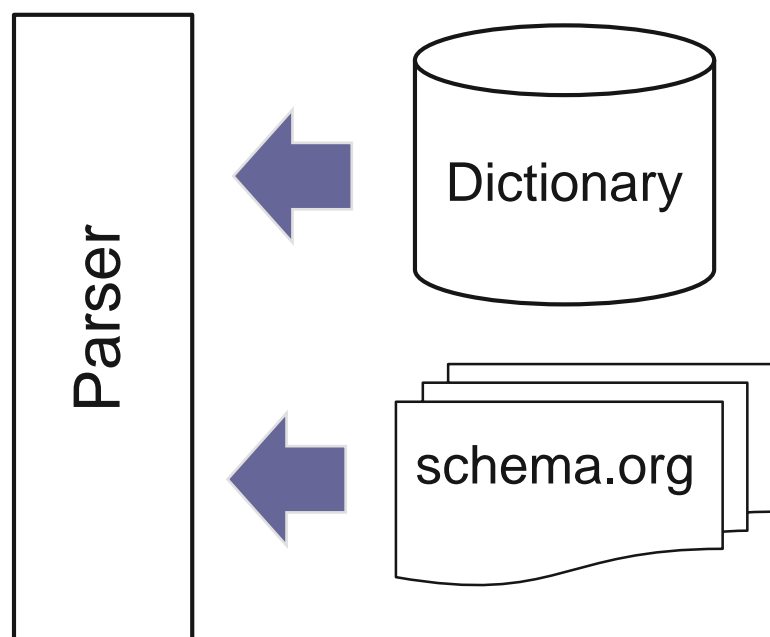
schema.org

schema.org data suitable as distant supervision:

- schema:Product/brand
- schema:Product/manufacturer
- 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.

Create and Verify Data Links

- Supervised learning of detailed matching rules leads to F1>95% (e.g. Silk and LIMES frameworks)

- Sources of supervision
  1. Data links and shared identifiers
     - owl:sameAs
     - schema:Product/productID
     - schema:Product/gtin13
  2. Human guidance via active learning

- How to generalize matching rules to data from multiple sources?
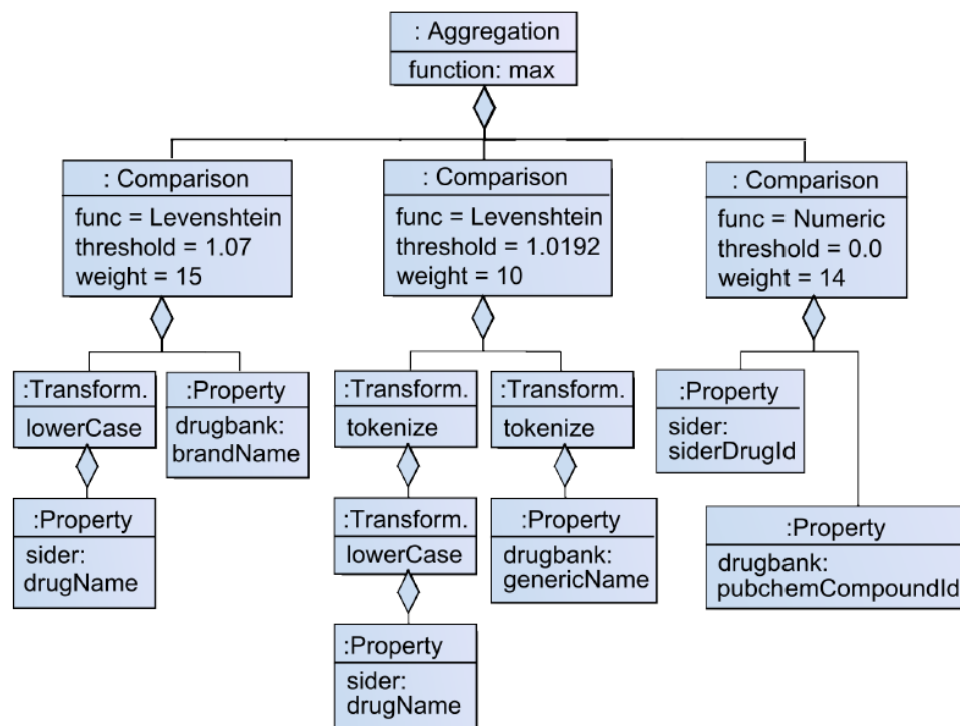
Isele, Bizer: Active Learning of Expressive Linkage Rules using Genetic Programming. JWS 2013.
Stonebraker, et al.: Data Curation at Scale: The Data Tamer System. CIDR 2013.

Resolve Data Conflicts

– Do you have some data that you already trust?

– Knowledge-based Trust

   • determine trustworthiness of a data source by comparing its content with trusted data (ground truth)

   • outperforms PageRank and voting

Web Data Source	
Country	**City**
Germany	Berlin
France	Paris
United Kingdom	London
Canada	Ottawa
USA	Washington D.C.
Mexico	Ecatepec

⟺

✓
?
✓
?
✓
✗

Trusted Data	
Country	**Capital**
Germany	Berlin
France	
United Kingdom	London
Canada	
USA	Washington D.C.
Mexico	Mexico City

Dong, et al.: Knowledge-based Trust: Estimating the Trustworthiness of Web Sources. VLDB 2015.

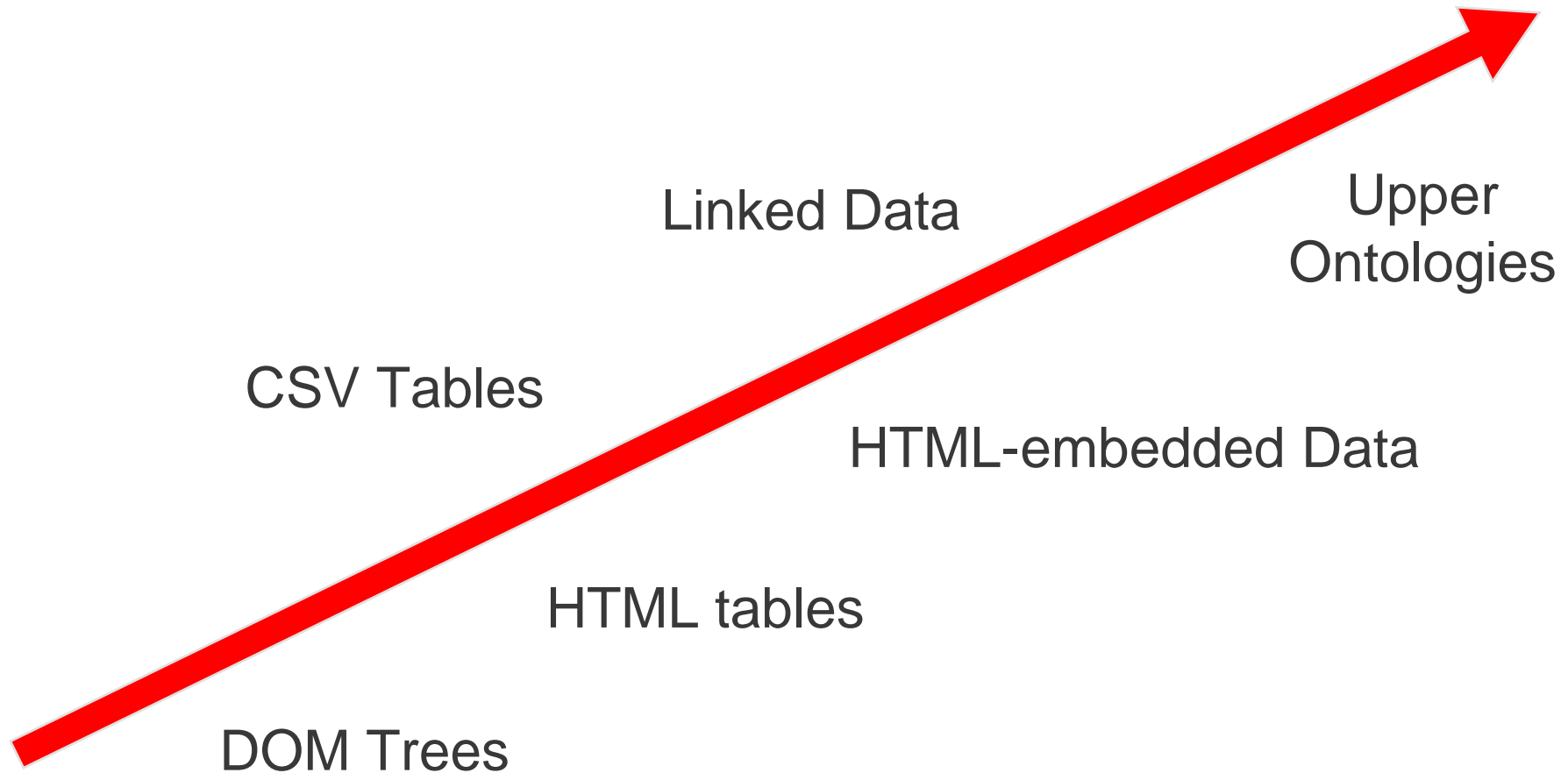# Google Knowledge Vault

- 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 ranking

- Results:

  - 271 million facts with confidence >90%

  - 90 million facts not in Freebase before
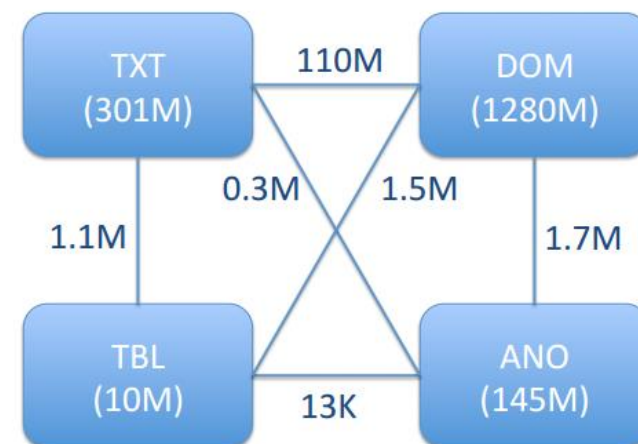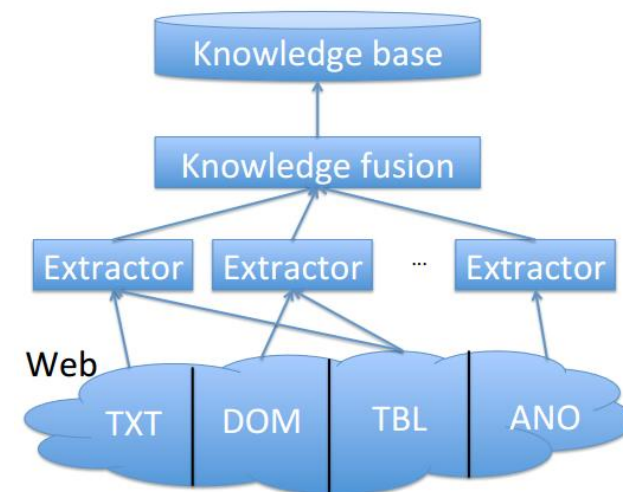
Dong, et al.: Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. SIGKDD 2014.

The Structural Continuum

Be open to different forms of "structured" Web content.

Linked Data

Upper Ontologies

CSV Tables

HTML-embedded Data

HTML tables

DOM Trees

Exploit Schema.org and HTML Tables Together

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.

Conclusions

The Semantic Web is Huge

− The Semantic Web contains more data than most people like

Billions of product offers

A description of every hotel in the world

Lots of data about local businesses

− exciting test-bed for research on data profiling, cleansing and integration

− endless data pool for commercial applications (product comparison, business listings, job search, …)

We will keep on seeing Similar Adoption Patterns

- as we need to be realistic about the effort spent by data publishers

- be happy about any semantic clues (integration hints) provided

- design algorithms to work despite the scarcity and noisiness of clues
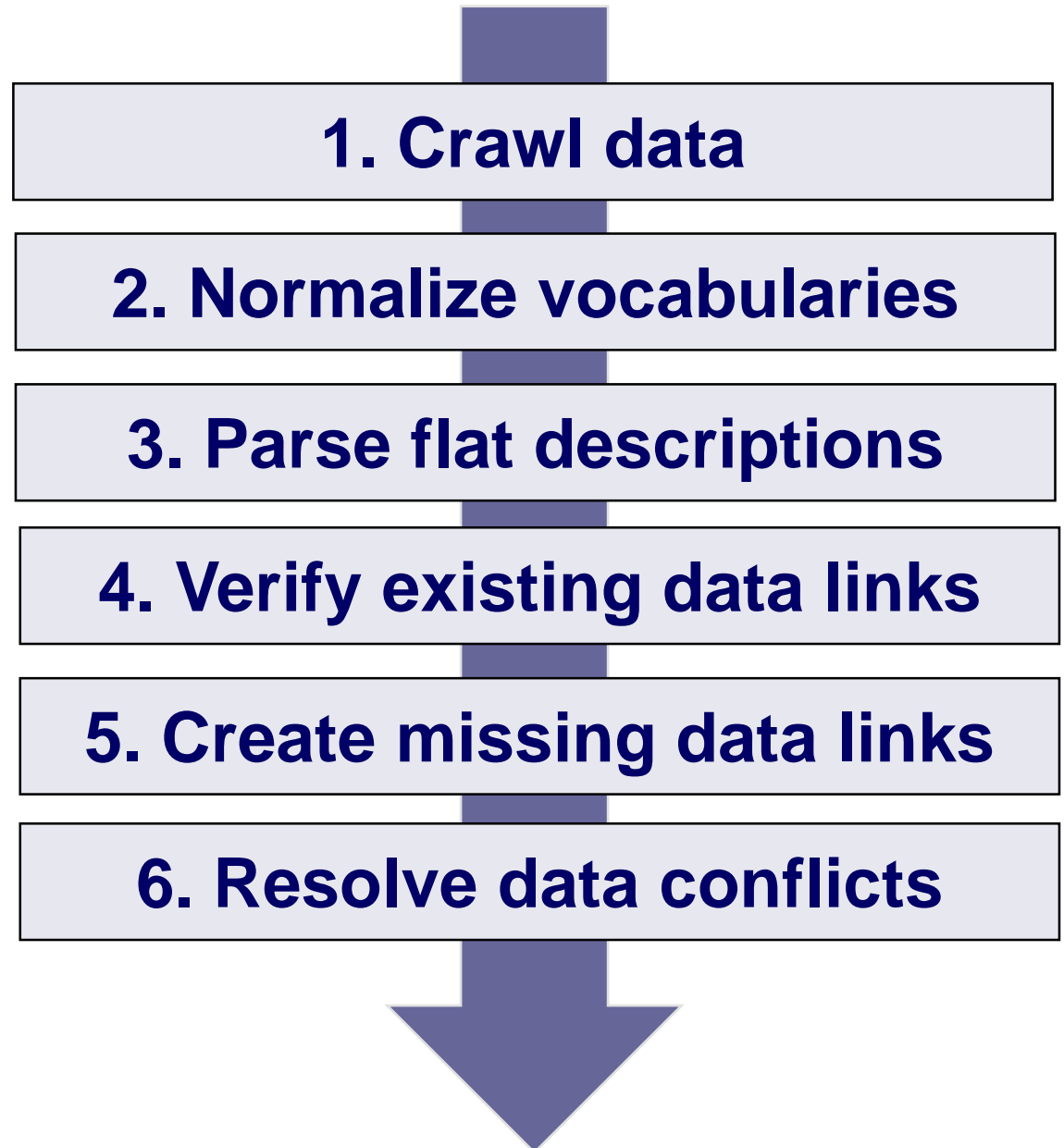
Missing Links and Identifiers

Flat Data Structures

Heterogeneity of Taxonomies

Mixed Data Quality

Semantic Web Clients need to be FAT Clients

There are no shortcuts!

1. Crawl data

2. Normalize vocabularies

3. Parse flat descriptions

4. Verify existing data links

5. Create missing data links

6. Resolve data conflicts

Thank you!