

Team Project HWS 2019

Integrating Product Specifications from the Web



- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
 - Web Data Integration
 - Data and Web Mining
 - Adoption of Data Web Technologies
 - Knowledge Base Construction
- Room: B6 - B1.15
- eMail: chris@informatik.uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30



- **Anna Primpeli**
- Graduate Research Associate
- Research Interests:
 - Data Extraction
 - Web Data Integration
 - Active Learning
 - Structured Data on the Web
- Room: B6, 26, C 1.04
- eMail: anna@informatik.uni-mannheim.de



Agenda of Today's Kickoff Meeting

1. You and Your Experience
2. Motivation and Project Goals
3. The WDC Training Dataset for Large-Scale Product Matching
4. Organization
5. Specific Subtasks
6. Schedule
7. Formal Requirements

Motivation of the Team Project

The Web is a rich source of product information

- same product is described by 100s of websites
 - merchants, producer, consumers
- different websites describe different aspects of a product
 - technical spec vs consumer experience
- there are plenty of offers for a product online
 - we can collect price information on global scale
- many websites point us at similar products



Using information about products from the Web, we can

- build comprehensive product catalogues and search engines
- construct global price comparison engines
- understand consumer and market behavior

Google Shopping: A Comprehensive Product Catalog

Groups product offers from different sources that describe the same product entity.

Aggregates the offers' features and maps them to a category-specific schema

Presents a structured product catalog to the user



Samsung S7 - 32 GB - Schwarz - Ohne SIM-Lock

€259 online

#2 in Mobiltelefone > Samsung > Galaxy > Galaxy S > Galaxy S7 ★★★★★ 42 110 product reviews

Samsung - Galaxy - Galaxy S - Galaxy S7 - Galaxy S7 (Original) - Android - 5,1 Zoll Display - Fingerabdruck, 5 Megapixel Frontkamera - 12 Megapixel

Samsung entdeckt eine neue Dimension für Smartphones. Eine Dimension voller atemberaubender Möglichkeiten. Neuer Wege, Samsungs Erfahrungen und Erinnerungen mit anderen zu teilen. Eine neue Zeit bricht an und von Anfang an dabei sein. Samsung hat Ihnen zugehört ... [more](#)

Schwarz - €259

Online stores Related items Reviews Details Accessories

Online stores Your location: Mannheim

☐ Free delivery ☐ Refurbished / used

Sellers	Seller Rating	Details	Base Price
europhones.de	No rating		€259.00 Free shipping
real.de - Spreewald Kontor	No rating		€369.95 (€2,433.88 / 1 kg) +€4.89 shipping
JACOB	90% positive (242)		€375.64 Free shipping
Onlineshopping.de + Show all 4	No rating		€259.00 Free shipping
Rakuten.de - ...	No rating		€490.11 +€6.90 shipping

Details

Allgemein

Produkttyp	Smartphone
Hersteller-Modellnummer	SM-G930F
Formfaktor	Touch
Schutz	Staubgeschützt, wasserfest
Integrierte Komponenten	Rückwärtige Kamera, Frontkamera, Audi
Breite	69.6 mm
Tiefe	7.9 mm
Höhe	142.4 mm
Gewicht	152 g
Gehäusefarbe	Schwarz
Gehäusematerial	Metall, Glas
SAR-Wert	0,621 W/kg (Körper) / 0,406 W/kg (Kopf)

Mobil

Technologie	WCDMA (UMTS) / GSM
Mobil-Breitband-Generation	4G
Service Provider	Nicht angegeben
Betriebssystem	Android 6.0 (Marshmallow)
Benutzeroberfläche	Samsung TouchWiz
Steuerung	GPS, A-GPS, GLONASS

Project Goal

... drive a well structured, comprehensive product catalogue from the Web

.... similar to the Google Shopping catalogue

... by:

1. Find clusters of offers from e-Shops that refer to the same product
2. Extract and normalize product features (key value pairs) from these clusters
3. Assemble a clean product catalogue for two product categories
 1. Single integrated schema per category
 2. Clean set of values per attribute

Details	
Allgemein	
Produkttyp	Smartphone
Hersteller-Modellnummer	SM-G930F
Formfaktor	Touch
Schutz	Staubgeschützt, wasserfest
Integrierte Komponenten	Rückwärtige Kamera, Frontkamera, Audi
Breite	69.6 mm
Tiefe	7.9 mm
Höhe	142.4 mm
Gewicht	152 g
Gehäusefarbe	Schwarz
Gehäusematerial	Metall, Glas
SAR-Wert	0,621 W/kg (Körper) / 0,406 W/kg (Kopf)
Mobil	
Technologie	WCDMA (UMTS) / GSM
Mobil-Breitband-Generation	4G
Service Provider	Nicht angegeben
Betriebssystem	Android 6.0 (Marshmallow)
Benutzeroberfläche	Samsung TouchWiz
Steuerung	GPS, A-GPS, GLONASS

How to find many offers of the same product?

Not an easy task!

- 
1. Which sources to consider?
 2. Which data to extract?
 3. How to recognize identical products?
 4. How to categorize products?
- 

OR...

Use the **WDC Training Dataset for Large-Scale Product Matching**

<http://webdatacommons.org/largescaleproductcorpus/index.html>

How to determine product features and extract feature values?

- Specification Tables and Lists
 - Detailed features as key/value pairs
- Product Titles
 - Product name plus selected features
- Product Descriptions
 - long free texts

Samsung Galaxy S4 Verizon AT&T T-Mobile GSM
Unlocked Smartphone SRF



Produkttyp	Smartphone
Formfaktor	Touch
Integrierte Komponenten	Rückwärtige Kamera, Frontkamera
Breite	70 mm
Tiefe	7.9 mm
Höhe	137 mm
Gewicht	130 g
Gehäusefarbe	White Frost

Before we can use these features:

- values need to be cleansed and normalized
- we might want to apply information extraction in order to increase the structuredness and density of the data

Das Samsung Galaxy S4 ist der unterhaltsame und hilfreiche Begleiter für Ihr mobiles Leben. Es verbindet Sie mit Ihren Liebsten. Es lässt Sie gemeinsam unvergessliche Momente erleben und festhalten. Es vereinfacht Ihren Alltag.

The WDC Training Dataset for Large-Scale Product Matching

1. Semantic Annotations in HTML Pages
2. Web Data Commons Project
3. Web Data Commons – Training Dataset for Large-Scale Product Matching



Semantic Annotation of HTML Pages: Schema.org



- ask site owners since 2011 to annotate data for enriching search results
- 675 Types: Event, Place, Local Business, **Product, Review**, Person
- Encoding: Microdata, RDFa, JSON-LD

The screenshot shows the schema.org website. At the top is a red header with the 'schema.org' logo, a search bar, and a 'Search' button. Below the header is a navigation bar with links for 'Home', 'Schemas', and 'Documentation'. The main content area displays the breadcrumb path 'Thing > Organization > LocalBusiness'. Below this is a description of 'LocalBusiness' as a particular physical business or branch of an organization, with examples like a restaurant, a branch of a bank, a medical practice, a club, or a bowling alley. A table follows, listing properties and their expected types and descriptions. The table is divided into two sections: 'Properties from Thing' and 'Properties from Place'.

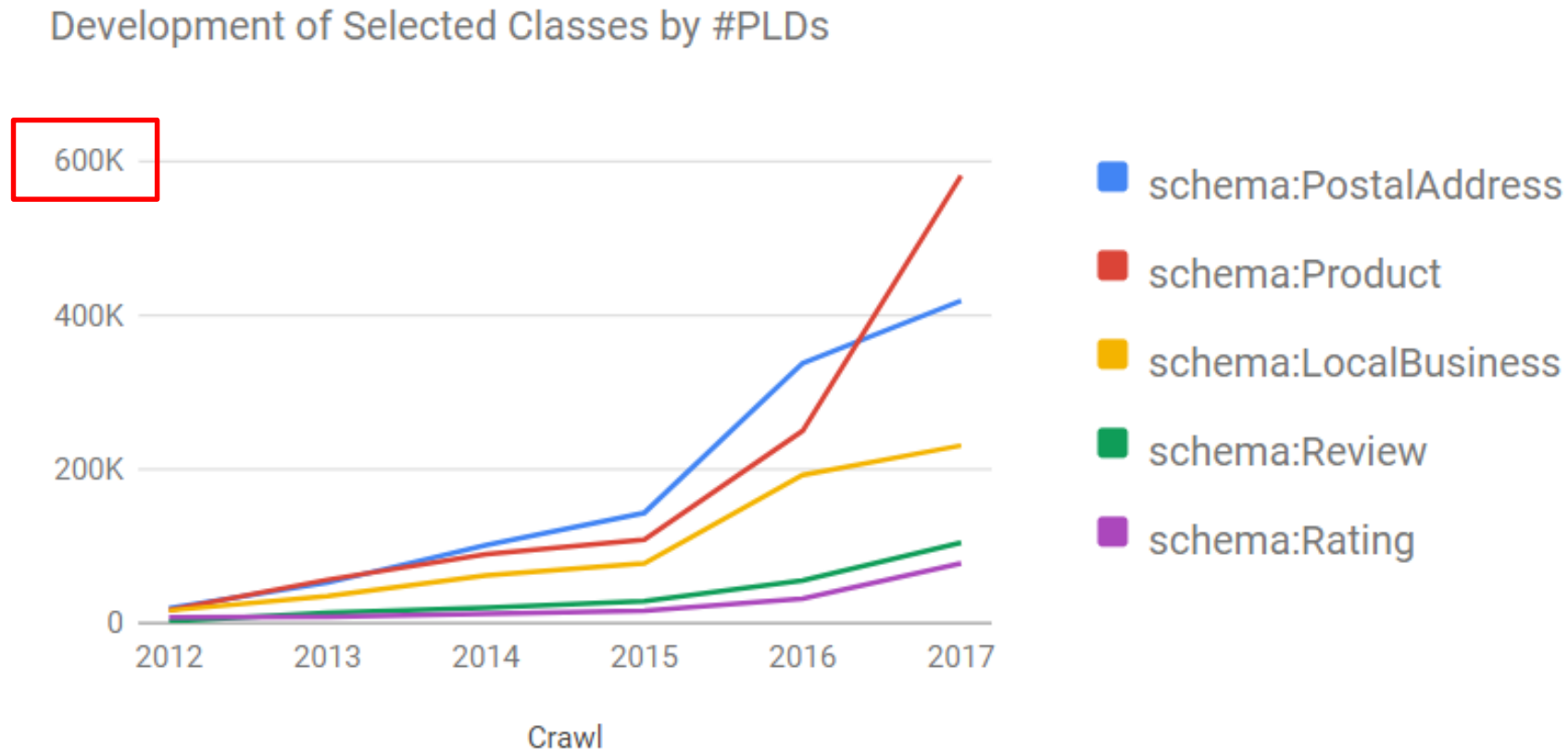
Example: Microdata Annotations in HTML



```
<div itemtype="http://schema.org/Product">
  <span itemprop="name">Sony GTK-XB5L Audiosystem</span>
  <span itemprop="gtin13">04048945021687</span>
  <span itemprop="description">high-power home audio system with Bluetooth technology</span>
</div>

<div itemprop="aggregateRating" itemscope itemtype="http://schema.org/AggregateRating">
  <span itemprop="ratingValue"> 4 </span> stars-based on
  <span itemprop="reviewCount"> 250 </span> reviews.
</div>
```

schema.org Annotations: Most Popular Classes



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

Properties used to Describe Products 2017

Top 15 Properties	PLDs	
	#	%
schema:Product/name	535,625	92%
schema:Offer/price	462,444	80%
schema:Product/offers	462,233	79%
schema:Offer/priceCurrency	430,556	74%
schema:Product/image	419,391	72%
schema:Product/description	377,639	65%
schema:Offer/availability	337,876	58%
schema:Product/url	263,720	45%
schema:AggregateRating/ratingValue	184,004	32%
schema:Product/sku	126,696	22%
schema:AggregateRating/reviewCount	112,408	19%
schema:Product/aggregateRating	101,434	17%
schema:Product/brand	73,934	13%
schema:Product/productID	35,211	6%
schema:Product/manufacturer	21,967	4%

Samsung Galaxy S4 Verizon AT&T T-Mobile GSM
Unlocked Smartphone SRF

Das Samsung Galaxy S4 ist der unterhaltsame und hilfreiche Begleiter für Ihr mobiles Leben. Es verbindet Sie mit Ihren Liebsten. Es lässt Sie gemeinsam unvergessliche Momente erleben und festhalten. Es vereinfacht Ihren Alltag.

UPC 610214632623

UPC 610214632623



Attribute/Value Table together with schema.org Annotations

s:breadcrumb

s:name

Attribute/Value
HTML Table

Walmart

FREE store pickup as soon as today

Search

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

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

★★★★★ Write a review Q&A By: HP

Specifications

Aspect Ratio:	16:9
Graphics Information:	HD Graphics 515
Processor Type:	Intel Core M5-6Y54 Dual-Core Processor
Hard Drive Capacity:	256 GB
Color:	Silver
Display Technology:	LED Backlight, Full HD Display
Resolution:	1080p
Form Factor:	Laptop
Processor Speed:	1.10 GHz
Color Category:	Silver
Contained Battery Type:	Lithium Ion
Maximum RAM Supported:	8 GB

\$1,060.79 + FREE shipping
Only 2 left!

Sold & Shipped by BidDeal

FREE Standard shipping Shipping options

Pickup not available from this seller

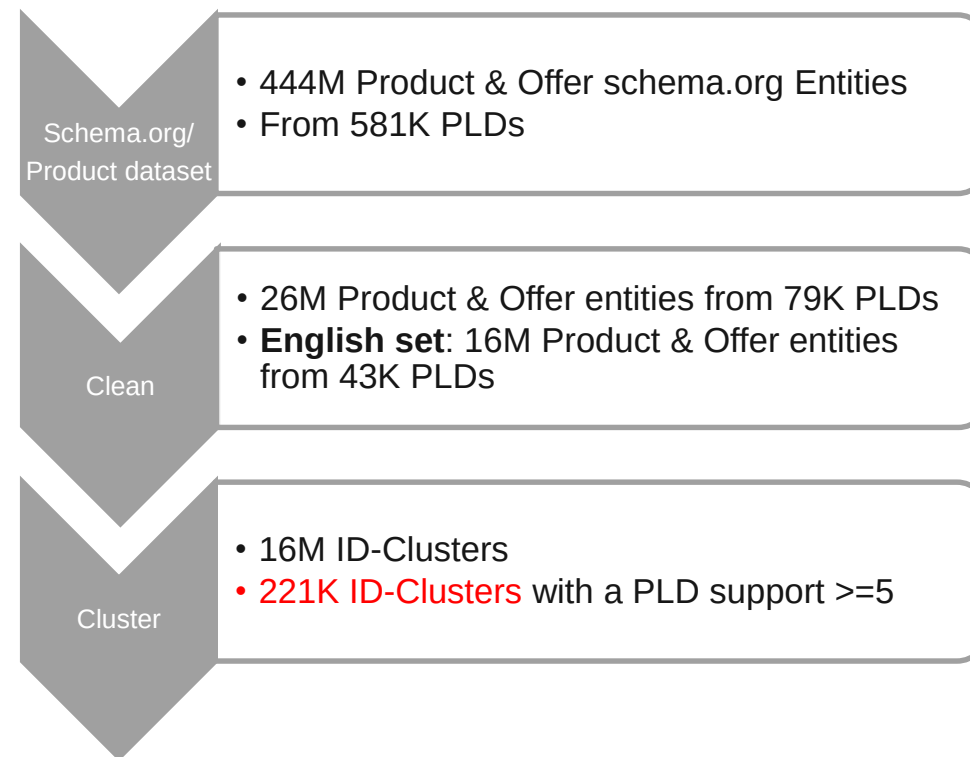
Qty: 1 Add to Cart

Add to List Add to Registry

Qui, et al.: **DEXTER: Large-Scale Discovery and Extraction of Product Specifications on the Web**. VLDB 2015.
Petrovski, P., Bizer, C.: **Extracting attribute-value pairs from product specifications on the web**. In: International Conference on Web Intelligence, 2017.

The WDC Training Dataset for Large-Scale Product Matching

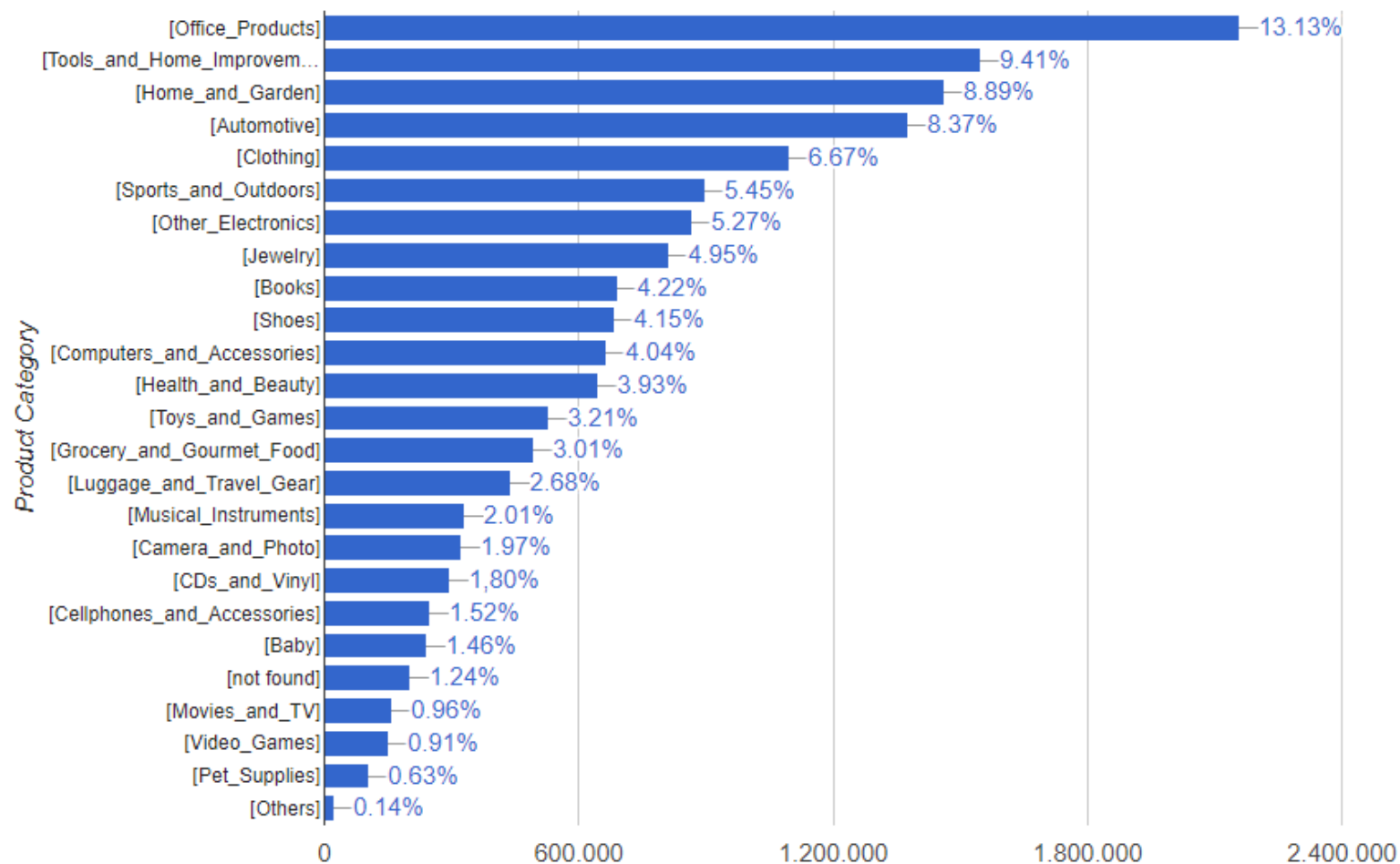
- Training corpus grouping schema.org product/ offer annotations by identifier value.
 - all **WDC 2017 product data** is included that
 - provides **some sort of product ID** (gtin, mpn, sku, identifier)
- Initial cleaning steps are performed
- Clustering of product descriptions from different PLDs (web sites) that share identifier values.



Details and Download:

- <http://webdatacommons.org/largescaleproductcorpus/index.html>
- Primpeli, A., Peeters, R., & Bizer, C.: **The WDC training dataset and gold standard for large-scale product matching**. In: Companion Proceedings of the 2019 World Wide Web Conference. pp. 381-386 ACM (2019).

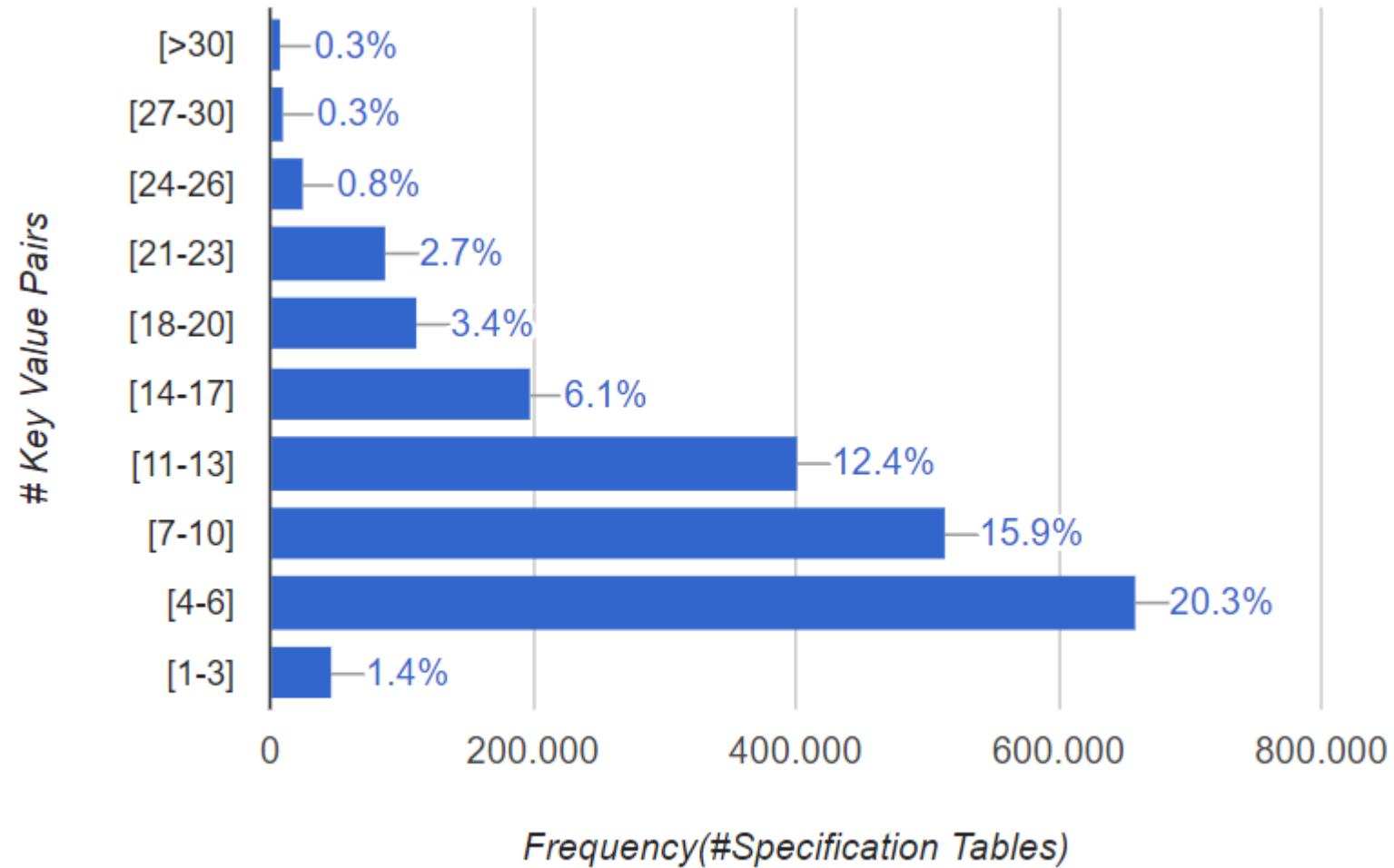
Distribution of Offers per Category in the English Training Set



source: <http://webdatacommons.org/categorization/index.html>

Offer Entities

Number of Key Value Pairs per Specification Table



Organization

Duration: 6 months (27.09.2019 – 29.03.2020)

Participants: 8 people

Type of work: Team and subgroup based

Milestones: 4 project phases

ECTS Points: 12

Evaluation

- Intermediate presentations
- Final report
- Individual contribution to the deliverables

Questions and Subtasks

1. Which categories to consider? → **Data Selection and Profiling**
2. How to integrate the schemata of the different sources? → **Schema matching**
3. How to increase the amount of extracted feature-values? → **Advanced Feature Extraction**
4. How to fuse multiple values of the same product and same feature? → **Data Fusion**
5. How to evaluate the quality of the product catalog? → **Catalog Profiling and Evaluation**

Phase 1: Data Selection, Data Profiling

Participants: all team members

Duration: 27.09. – 25.10.

Deliverables: 20 min. presentation, code & data

Input: WDC English Training Set, Categorization of Offers **V2**, Specification tables

*provided at <http://webdatacommons.org/largescaleproductcorpus/>
and <http://webdatacommons.org/categorization/>*

Tasks

1. Consider the offers with specification tables and
 - 1a. Profile amount of specification tables per category (cluster and offer level)
 - 1b. Profile amount of feature-value pairs per category (cluster and offer level)
3. Based on the profiling results decide on two product categories (or subcategories)
4. Profile the features and the feature values of the offers of the selected categories

Phase 1: How to get started?

"Brand": "Plextor", "Product name": "M5M 128GB", "Productserie": "M5M", "Productgroep": "Hard disks/SSDs", "MPN": "PX-128M5M PX128M5M"

- Get the updated categorization of the corpus

<http://webdatacommons.org/categorization/index.html>

File: [categories_offers_en_clusters.csv.gz](#)

- Get the offers of the english corpus

<http://webdatacommons.org/largescaleproductcorpus/index.html>

File: [offers_english.json.gz](#)

- Get the specification tables per offer

<http://webdatacommons.org/largescaleproductcorpus/index.html>

File: [specTables.json.gz](#)

Phase 1: How to get started?

Example cluster categorization entry
source: categories_offers_en_clusters.csv.gzip

predictedLabel	cluster_id
Computers_and_Accessories	100
Health_and_Beauty	100000
Musical_Instruments	1000000

How many offers per category?

Example offer entry
source: offers_english.json.gz

```
{"url": "http://www.snapweb.net/stdn-23r03-006-4",  
  "nodeID": "node31a031bfe4c4e6c9e9591270e9a169d",  
  "cluster_id": "100",  
  "identifiers": [{"sku": "stdautoqd24wh"}],  
  "schema.org_description": [{"  
    "/name": "[\" M5M 128GB Productserie M5M\"]",  
    "/brand": "[\" Plextor \"]",  
    "/description": "[\" Brand Plextor Product name M5M 128GB  
    Productserie M5M \"<http://www.snapweb.net/it-it/accessori-per-  
    lettori/datalogic/accessori-per-lettori/stdn-23r03-006-4>",  
    "relationToParent": "http://www.w3.org/1999/xhtml/microdata#item""par  
    ent_schema.org_description": [{"  
      "/title": "[\" MPN PX-128M5M \""]"}]}
```

How many offers
have specification
tables?

Example specification table entry
source: specTables.json.gz

How many offers
have how many
key value pairs?

```
{"url": "http://www.snapweb.net/stdn-23r03-006-4.html",  
  "fileName": "crawl-data/CC-MAIN-2017-  
  47/segments/1510934803848.60/2.warc.gz",  
  "offset": "732495936",  
  "specTableContent": "Brand Plextor Product name M5M 128GB  
  Productserie M5M Productgroep Hard disks/SSDs MPN PX-  
  128M5M PX128M5M",  
  "keyValuePairs": {"Brand": "Plextor", "Product name": "M5M  
  128GB", "Productserie": "M5M", "Productgroep": "Hard  
  disks/SSDs", "MPN": "PX-128M5M PX128M5M"}}
```

Aggregate on the **cluster** level to
answer:

- How many clusters have specification tables?
- How many clusters have how many key value pairs?

Group clusters by **category** to
answer:

- How many specification tables are there per category?
- How many key value pairs are there per category?

Phase 2: Schema Matching

Participants: two subgroups (one subgroup per selected category)

Duration: 25.10 – 29.11.

Deliverables: 30 min. Presentation, code & data

Input: Feature-value pairs from the specification tables of product offers for two categories

Tasks:

1. Closed schema matching ($N \rightarrow 1$)

- 1a. Retrieve a dictionary from top attributes using the profiling results of phase 1
- 1b. Perform and **evaluate** label / instance-based schema matching

2. Open schema matching ($N \rightarrow N$)

- 2a. Consider the clustered offers, perform and **evaluate** duplicate-based schema matching

3. Compile dictionary of possible values for each attribute

Phase 3: Data Normalization and Advanced Feature Extraction

Participants: two subgroups (one subgroup per selected category)

Duration: 29.11 – 24.01

Deliverables: 30 min. presentation, code & data

Input: Dictionary from phase 2, all offers from the WDC English Training set

Tasks

Using the dictionary from phase 2:

1. Design/learn data normalization rules considering their datatypes
 - e.g.: convert all dates to years, normalize values of categorical attributes
2. Increase the completeness/accuracy of the catalog by extracting feature-value pairs from offers containing no specification tables using their titles and descriptions.

Phase 4: Data Fusion and Catalog Curation

Participants: two subgroups (one subgroup per selected category)

Duration: 24.01 – 06.03.

Deliverables: 30 min. presentation, code & data

Tasks

1. Compile integrated catalog for each category
 - 1a. Map the feature-values of every product to the catalog
 - 1b. Fuse values (voting-based, source-based, ...)
 - 1c. Evaluate fusion
2. Profile attribute density in the integrated catalog

Schedule

Date	Session
Friday, 27.09.2019	Kickoff meeting (today)
	Phase 1 (all members): Data Selection, Data Profiling
Monday, 14.10.2019, 17:30	Meet Anna and report plan/current results
Friday, 25.10.2019, 9:00	1 st Deliverable: 20 minutes presentation, code & data - Subgroup formation
	Phase 2 (in 2 subgroups): Schema Matching
Friday, 08.11.2019, 17:30	Meet Anna and report plan/current results
Friday, 29.11.2019, 09:00	2 nd Deliverable: 30 minutes presentation, code & data
	Phase 3 (in 2 subgroups): Advanced Feature Extraction and Data Normalization
Friday, 13.12.2019, 17:30	Meet Anna and report plan/current results
Friday, 31.01.2020, 9:00	3 rd Deliverable: 30 minutes presentation, code & data
	Phase 4 (in 2 subgroups): Data Fusion and Catalog Curation
Friday, 07.02.2020, 17:30	Meet Anna and report plan/current results
Friday, 06.03.2020, 9:00	4 th Deliverable: 30 minutes presentation, code & data
Sunday, 29.03.2020	Final Report Submission

Formal Requirements & Consultation

Deliverables

1. On the deliverable dates provide us via e-mail with:

- **Presentation slides**
- **Task to member report:** excel sheet stating which team member conducted which subtask
- **Code/ Data:** link or zipped folder with your code and data

2. Final Report

- **15 pages** including appendices, not including the bibliography
- every additional page reduces your grade by 0.3
- Created with Latex template of the Data and Web Science group (<https://www.uni-mannheim.de/dws/teaching/thesis-guidelines/>)

All e-mails should be sent to Chris & Anna!

Formal Requirements & Consultation

Final grade

- 20% for every phase
- 20% for final report
- Late submission: -0.3 per day

Consultation

- Send one e-mail per team or subgroup stating your questions to Anna

Related Work (1/2)

- Primpeli, A., Peeters, R. and Bizer, C. The WDC training dataset and gold standard for large-scale product matching. In Liu, L., Companion Proceedings of The 2019 World Wide Web Conference (S. 381-386). , ACM 2019.
- Petrovski, Petar, Volha Bryl, and Christian Bizer. "Integrating product data from websites offering microdata markup." *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014.
- Qiu, Disheng, et al. "Dexter: large-scale discovery and extraction of product specifications on the web." *Proceedings of the VLDB Endowment* 8.13 (2015): 2194-2205
- Petrovski, P., Bizer, C.: Extracting attribute-value pairs from product specifications on the web. In: Proceedings of the International Conference on Web Intelligence. pp. 558–565 ACM (2017).
- Ristoski, Petar, and Peter Mika. "Enriching product ads with metadata from HTML annotations." *International Semantic Web Conference*. Springer, Cham, 2016
- Kannan, Anitha, et al. "Matching unstructured product offers to structured product specifications." *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011

Related Work (2/2)

- Petrovski, Petar, Volha Bryl, and Christian Bizer. "Learning regular expressions for the extraction of product attributes from e-commerce microdata." *Proceedings of the Second International Conference on Linked Data for Information Extraction-Volume 1267*. CEUR-WS. org, 2014
- Probst, Katharina, et al. "Semi-Supervised Learning of Attribute-Value Pairs from Product Descriptions." *IJCAI*. Vol. 7. 2007
- Meusel, Robert, et al. "Exploiting microdata annotations to consistently categorize product offers at web scale." *International Conference on Electronic Commerce and Web Technologies*. Springer International Publishing, 2015
- Poon, Hoifung, and Pedro Domingos. "Unsupervised ontology induction from text." Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.
- Navigli, Roberto, Paola Velardi, and Stefano Faralli. "A graph-based algorithm for inducing lexical taxonomies from scratch." *IJCAI*. Vol. 11. 2011.
- Ristoski, Petar, et al. "Large-scale taxonomy induction using entity and word embeddings." Proceedings of the International Conference on Web Intelligence. ACM, 2017.

Questions?

