Seminar CS715

Large-Scale Data Integration

Content and Dynamics of the Web of Data
Hallo

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

- **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
You and Your Experience

- A Short Round of Introductions
  - What are you studying?
  - Which DWS courses did you attend?
  - What kind of experience do you have with Data Science/Data Engineering projects?

- Participants
  1. Loos, Lukas Michael
  2. Tseng, Yen-Chun
  3. Zyberaj, Lonora
  4. Rösel, Marvin Mike
  5. Joshi, Rahul S
  6. Agarwal, Mayank
  7. Böckling, Martin Giovanni
Agenda of Today’s Kickoff Meeting

1. Seminar organization
2. Seminar topics
3. How to structure your seminar paper / presentation?
4. Questions and guidance
1. Organization
Learning Targets

- Writing a seminar thesis as an exercise for your master thesis
- Searching and citing scientific papers / journal articles
- How to derive summarization statistics about large data spaces
- How to structure your thesis and presentation
- How to argue, how to explain, how to write!
- How to write a nicely formatted paper using LaTeX
<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday, 19.09.2021</td>
<td>Send list of preferred topics via eMail</td>
</tr>
<tr>
<td>Thursday, 23.09.2021, 15:30</td>
<td><strong>Kick-off meeting</strong></td>
</tr>
<tr>
<td></td>
<td>1. Read provided papers about your topic</td>
</tr>
<tr>
<td></td>
<td>2. Search for additional literature</td>
</tr>
<tr>
<td></td>
<td>3. Gather and analyze data</td>
</tr>
<tr>
<td>Until Friday 15.10.2021</td>
<td>Meet with your mentor to discuss your results and presentation</td>
</tr>
<tr>
<td></td>
<td>Prepare draft of your presentation</td>
</tr>
<tr>
<td>Until Sunday 5.11.2021</td>
<td>Send draft presentation to your mentor</td>
</tr>
<tr>
<td></td>
<td>Finalize your presentation</td>
</tr>
<tr>
<td>Friday, 19.11.2021 (10:00-12:30)</td>
<td>Presentation and discussion of your topic (30 % of your final grade)</td>
</tr>
<tr>
<td></td>
<td>Write seminar thesis</td>
</tr>
<tr>
<td>Sunday, 16.01.2021</td>
<td>Submission of your seminar thesis (70 % of your final grade)</td>
</tr>
</tbody>
</table>
Formal Requirements

- **Presentation**
  - 15 minutes + 10 minutes discussion
  - should be 100% understandable for all participants

- **Written report (paper)**
  - 12-15 pages single column
    - including abstract and appendixes
    - not including bibliography
    - every additional page reduces your grade by 0.3
  - written in English
  - use latex template of Springer Computer Science Proceedings

- **Final grade**
  - 70% written report
  - 30% presentation
Which template to use?

http://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines
2. Seminar Topic
Content and Dynamics of the Web of Data

The Classic Document Web

The Web of Data

Deep Web (via APIs and forms)

1. HTML-embedded Data
2. Web APIs
3. Linked Data
More and more websites semantically markup the content of their HTML pages using standardized markup formats.

```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>
```
HTML-embedded Data

- used for embedding data into the HEAD of HTML pages
- putting data in 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>
```
Schema.org

- 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
Usage of Schema.org Data @ Google

Product offers

Job offers

https://developers.google.com/search/docs(guides/search-gallery
Bevölkerungsbestand in Mannheim 2013-2018

**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

**Dataset updated** 15.07.2019

**License**
dl-de-by-2.0

**Available download formats from providers**
excel, csv, json

**Description**
Web Data Commons

- Extracts structured data from the Common Crawl since 2009
  - Microformats, Microdata, RDFa, JSON-LD
  - Schema.org

- Analyzes and provides the extracted data for download
  [http://webdatacommons.org/structureddata/index.html](http://webdatacommons.org/structureddata/index.html)

- Provides class-specific Schema.org data as RDF quads
  [http://webdatacommons.org/structureddata/2020-12/stats/schema_org_subsets.html](http://webdatacommons.org/structureddata/2020-12/stats/schema_org_subsets.html)

- Provides class-specific Schema.org data as one table per website
  [http://webdatacommons.org/structureddata/schemaorgtables/](http://webdatacommons.org/structureddata/schemaorgtables/)
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 via the Web.
  - HTML interfaces
  - Web APIs
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 20,000 Web APIs
  - lists over 6,800 mashups
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
The Linked Open Data Cloud

1,239 datasets connected by 16,147 sets of RDF links (as of March 2019)

https://lod-cloud.net/
Questions To be Answered

- Q1: What data is available?
  - Which entities are described?
  - What attributes are used to describe the entities? (Do they differ from the Google recommendation?)

- Q2: Who is publishing the data?
  - What types of publishers do provide data? And why?
  - Are the main players (most popular websites) in the domain contributing?

- Q3: How has the data changed over the last years?
  - Growth? Change in topic or depth? Is the data maintained?

- Q4: Which applications use the data?
  - Who else beside of Google/Bing/Yandex uses the data?

- Q5: How difficult is it to integrate the data?
  - Match entities? Categorize entities? Fuse data from important attributes?

- Q6: What empirical research has been published about these questions?
Input for your Seminar Thesis

- Scientific literature and web pages about your topic
- Statistics from WebDataCommons, the LOD Cloud, Programmable Web
- List of the most visited websites per domain
- If information is missing in order to answer question Q1-Q6, try to produce the statistics yourself by analyzing content from the respective sources
  - If you run into performance problems, analyze a sample of the data
    • Sources: most visited sites + randomly selected
    • Data: randomly selected instances of popular classes.
  - Restrict yourself to at most 10 sources

- Meusel, Robert, Christian Bizer, and Heiko Paulheim. „A web-scale study of the adoption and evolution of the schema.org vocabulary over time.“ Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics. 2015.


- https://developers.google.com/search/docs/advanced/structured-data/local-business

- http://webdatacommons.org/structureddata/schemaorgtables/
2. Schema.org Job Posting Data – Structure, Dynamics, and Applications (Marvin Mike, mentor: Anna)

- Meusel, Robert, Christian Bizer, and Heiko Paulheim. „A web-scale study of the adoption and evolution of the schema.org vocabulary over time.“ Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics. 2015.

- https://developers.google.com/search/docs/advanced/structured-data/job-posting

- http://webdatacommons.org/structureddata/schemaorgtables/

- Meusel, Robert, Christian Bizer, and Heiko Paulheim. „A web-scale study of the adoption and evolution of the schema.org vocabulary over time.“ Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics. 2015.


- https://developers.google.com/search/docs/advanced/structured-data/event

- http://webdatacommons.org/structureddata/schemaorgtables/
Topics

4. Schema.org Data Set Metadata – Structure, Dynamics, and Applications (Mayank, mentor: Anna)


- https://developers.google.com/search/docs/advanced/structured-data/dataset

- http://webdatacommons.org/structureddata/schemaorgtables/
5. Schema.org Product Data – Structure, Dynamics, and Applications (Rahul, mentor: Anna)

- Meusel, Robert, Christian Bizer, and Heiko Paulheim. „A web-scale study of the adoption and evolution of the schema.org vocabulary over time.“ Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics. 2015.


- https://developers.google.com/search/docs/advanced/structured-data/product

- http://webdatacommons.org/structureddata/schemaorgtables/
6. Web APIs – Topics, Dynamics, and Applications (Lonora, mentor: Chris)


- https://www.programmableweb.com/api-research
7. Linked Data – Topics, Dynamics, Best Practices, and Applications (Lukas Michael, mentor: Chris)


3. How to Structure Your Paper / Presentation
Structure Your Paper along the 6 Questions

- **Q1: What data is available?**
  - Which entities are described?
  - What attributes are used to describe the entities? (Do they differ from the Google recommendation?)

- **Q2: Who is publishing the data?**
  - What types of publishers do provide data? And why?
  - Are the main players (most popular websites) in the domain contributing?

- **Q3: How has the data changed over the last years?**
  - Growth? Change in topic or depth? Is the data maintained?

- **Q4: Which applications use the data?**
  - Who else beside of Google/Bing/Yandex uses the data?

- **Q5: How difficult is it to integrate the data?**
  - Match entities? Categorize entities? Fuse data from important attributes?

- **Q6: What empirical research has been published about these questions?**
Citing different Types of Publications

- Journal article
  - Good to cite, current research results
  - Survey articles (very good for an overview)

- Conference and workshop paper
  - Good to cite, current research results

- Books (sometimes cited)
  - Textbooks
  - Collections of articles/papers => Cite specific paper in book

- Websites
  - Prefer scientific paper over web pages.
    - If a webpage has a author and a title (like a blog post), cite it in the references list. If this is not the case, cite via a footnote.
    - If you find the same information in a scientific paper, cite the paper.
    - Do not cite Wikipedia, ever!

- Slide sets
  - Never cite!
How to Find Relevant Publications?

- Use Standard Search Engines
- Use Google Scholar
  - we use it a lot ourselves
- Search Engines of the University’s library
  - see slides from the library course
- Exploit references: Given a relevant document \( x \)
  - Follow references in the past: papers \( y \) that \( x \) has cited
  - Follow references in the future: papers \( y \) that cited \( x \)
    ("cited by" functionality in Google scholar)
4. Questions?