

Seminar CS715

Large-Scale Data Integration

Content and Dynamics of the Web of Data



- **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

Schedule

Date	Session
Sunday, 19.09.2021	Send list of preferred topics via eMail
Thursday, 23.09.2021, 15:30	Kick-off meeting
	1. Read provided papers about your topic 2. Search for additional literature 3. Gather and analyze data
Until Friday 15.10.2021	Meet with your mentor to discuss your results and presentation
	Prepare draft of your presentation
Until Sunday 5.11.2021	Send draft presentation to your mentor
	Finalize your presentation
Friday, 19.11.2021 (10:00-12:30)	Presentation and discussion of your topic (30 % of your final grade)
	Write seminar thesis
Sunday, 16.01.2021	Submission of your seminar thesis (70 % of your final grade)

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
 - <http://www.springer.com/de/it-informatik/Incs/conference-proceedings-guidelines>
- Final grade
 - 70% written report
 - 30% presentation

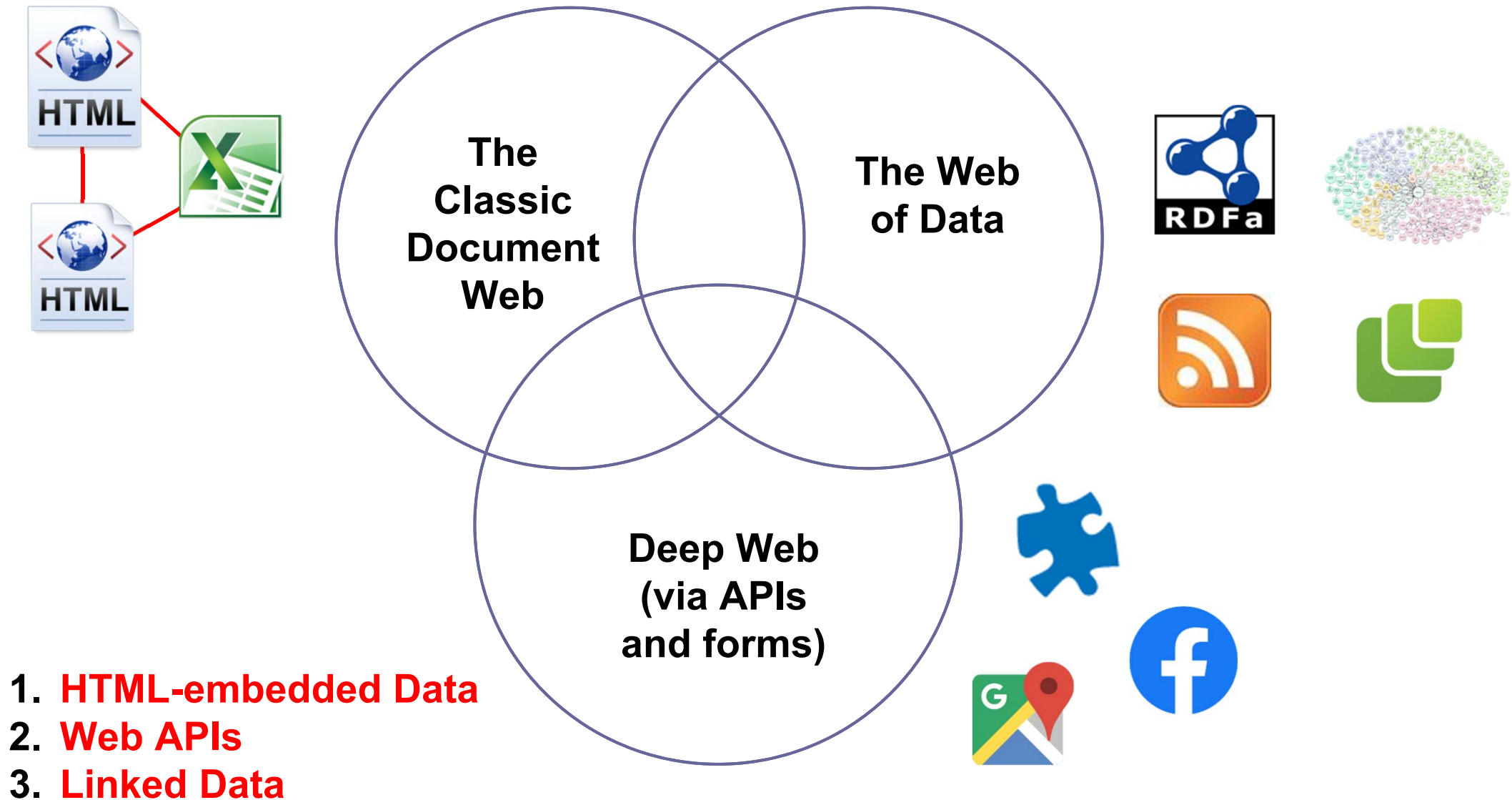
Which template to use?



<http://www.springer.com/de/it-informatik/Incs/conference-proceedings-guidelines>

2. Seminar Topic

Content and Dynamics of the Web of Data



HTML-embedded Data

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



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



```
<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>
```



- 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

schema.org **Search**

[Home](#) [Schemas](#) [Documentation](#)

Thing > Organization > LocalBusiness

A particular physical business or branch of an organization. Examples of LocalBusiness include a restaurant, a particular branch of a restaurant chain, a branch of a bank, a medical practice, a club, a bowling alley, etc.

Property	Expected Type	Description
Properties from <u>Thing</u>		
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
url	URL	URL of the item.
Properties from <u>Place</u>		
address	<u>PostalAddress</u>	Physical address of the item.
aggregateRating	<u>AggregateRating</u>	The overall rating, based on a collection of reviews or ratings, of the item.
containedIn	<u>Place</u>	The basic containment relation between places.



Usage of Schema.org Data @ Google

Gramercy Tavern - Flatiron - New York, NY | Yelp

www.yelp.com > Restaurants > American (New) ▾

★★★★★ Rating: 4.5 - 1,288 reviews - Price range: \$\$\$\$

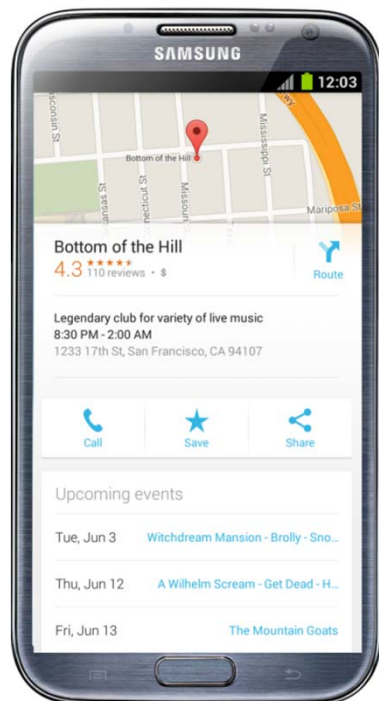
Jeff C and I were in New York for vacation, and I wanted to treat him to a nice dinner for Gramercy Tavern is certainly a legendary NY dining establishment.

Gramercy Tavern Restaurant - New York, NY | OpenTable

www.opentable.com > ... > Gramercy restaurants ▾

★★★★★ Rating: 4.7 - 508 reviews - Price range: \$50 and over

Book now at Gramercy Tavern in New York, explore menu, see photos and read 508 reviews: "The menu was so limited but it was worth trying, food was deli..."



% <Q NIKK@N
R ICDI !
N@<M>C M@NPCN

-J >G=PNDI@NN@!
JI H <KN

% <Q NIKK@N
R ICDI !
DI AJ !=JS@N



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

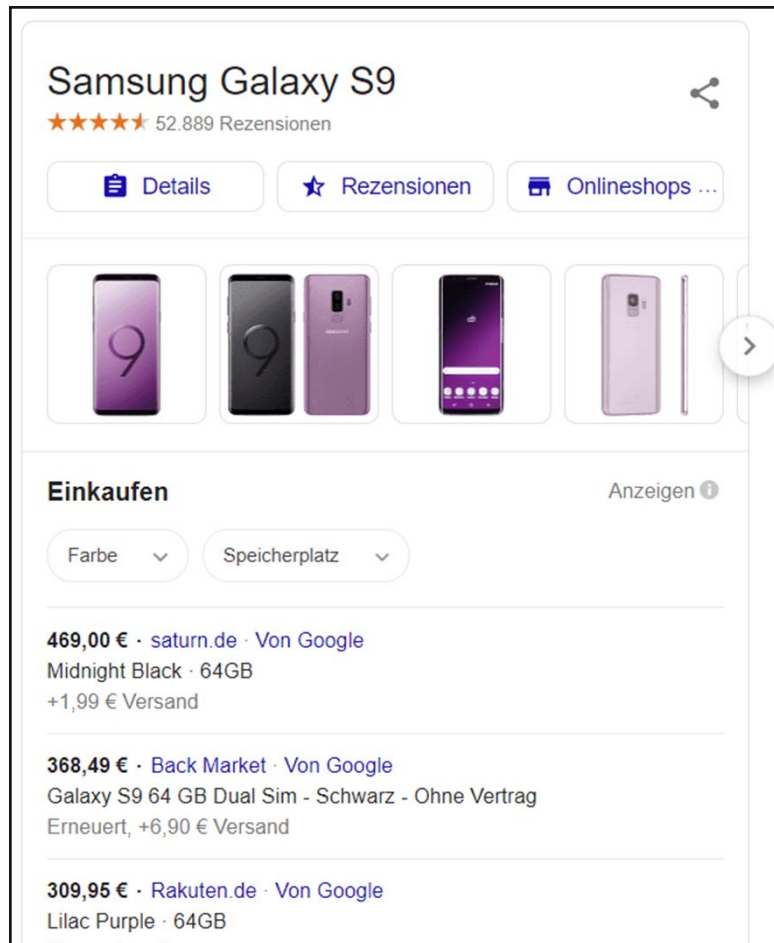
Record labels: Fat Possum Records, Nonesuch Records, V2 Records, Alive Natural Sound Records

Awards: Grammy Award for Best Rock Album, more

Upcoming events

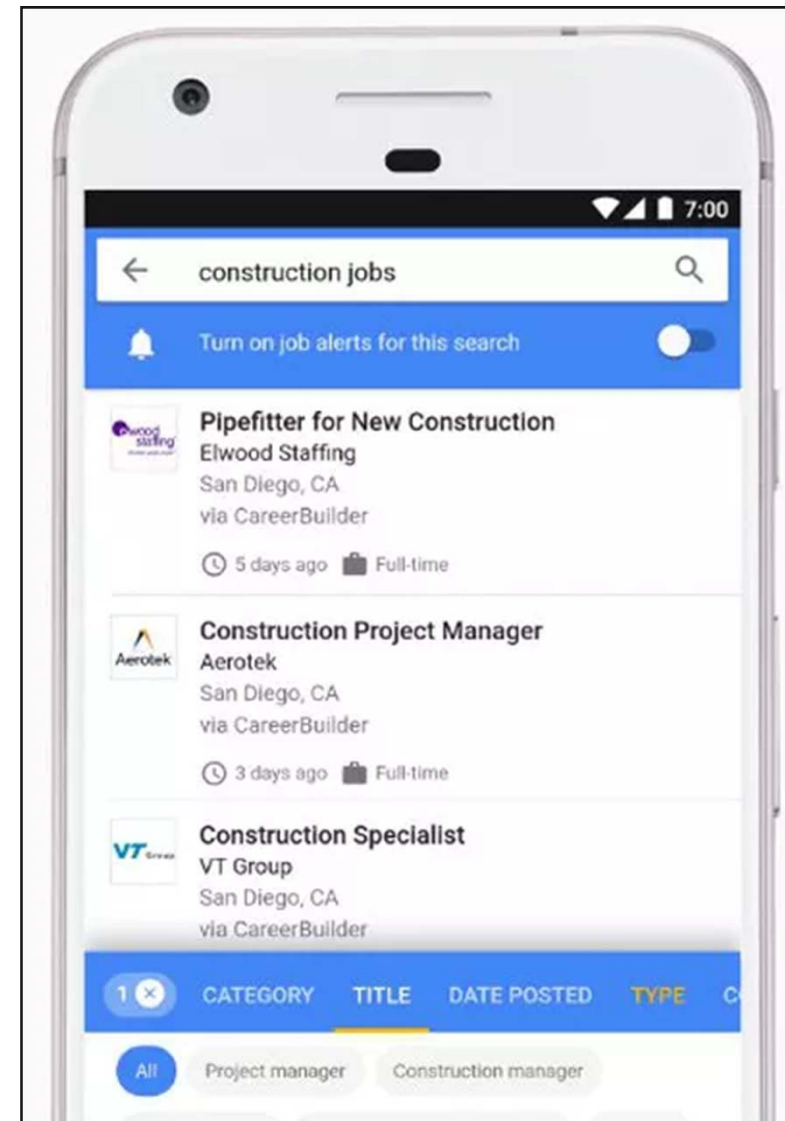
Jun 20 Fri	The Black Keys Neuhausen ob Eck (near you)
May 16 Fri	The Black Keys Gulf Shores, AL
Jun 22 Sun	The Black Keys Scheeßel

Usage of Schema.org Data @ Google



1MJ?P>O!
JAA@N

+J=!
JAA@N



<https://developers.google.com/search/docs/guides/search-gallery>

100+ results found



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

Bevölkerungsbestand in Mannheim 2013-2018

[Explore at mannheim.opendatasoft.com](https://mannheim.opendatasoft.com)

Dataset updated 15.07.2019

License

dl-de-by-2.0

Available download formats from providers

excel, csv, json

Description

Bevölkerung in Mannheim und den Stadtteilen nach Jahr, Wohnstatus, Geschlecht und Migrationshintergrund Mannheim ist Standort einer Landeserstaufnahmeeinrichtung (LEA) sowie Bedarfsorientierter Erstaufnahmeeinrichtungen (BEAs) für Flüchtlinge. Die Bewohnerinnen und Bewohner dieser Einrichtungen sind meldepflichtig und fließen als Hauptwohnbevölkerung in die Statistik der Stadt Mannheim ein.

<https://toolbox.google.com/datasetsearch>

Web Data Commons

- Extracts structured data from the Common Crawl **since 2009**

- Microformats, Microdata, RDFa, JSON-LD
- Schema.org
- Covers several million websites



- Analyzes and provides the extracted data for download

<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

- Provides class-specific Schema.org data as one table per website

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

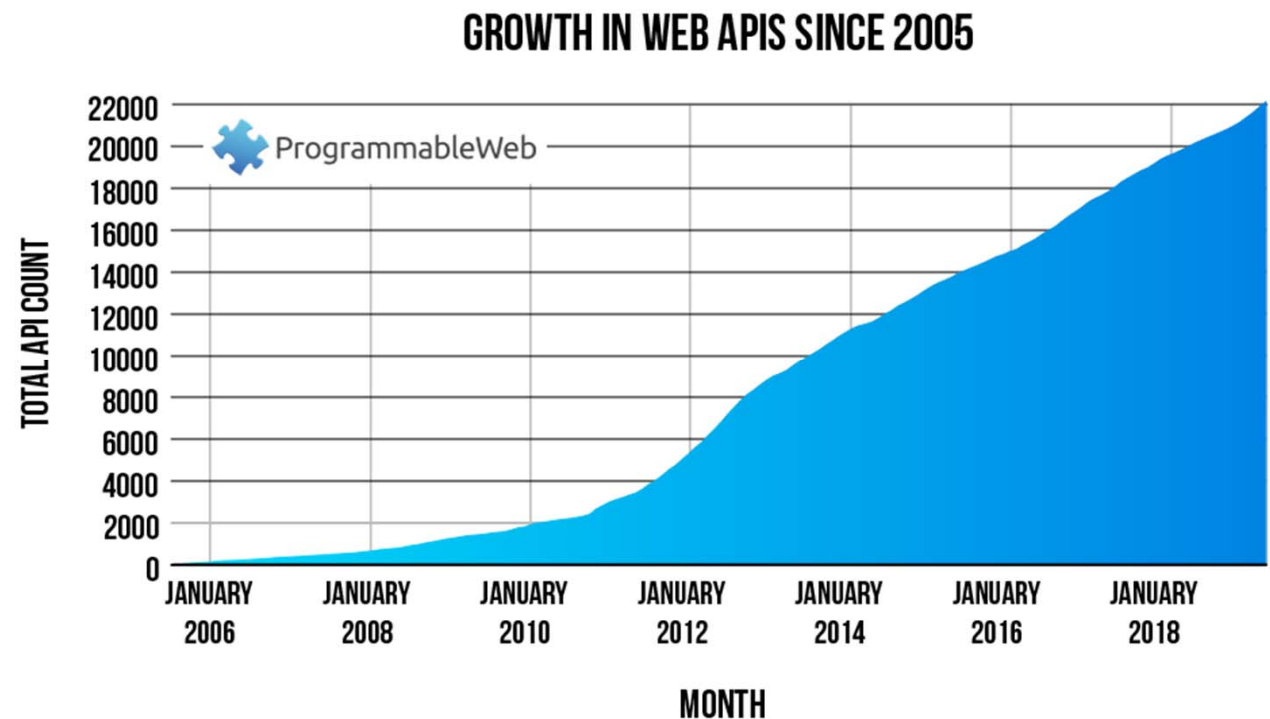
Web APIs

- 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



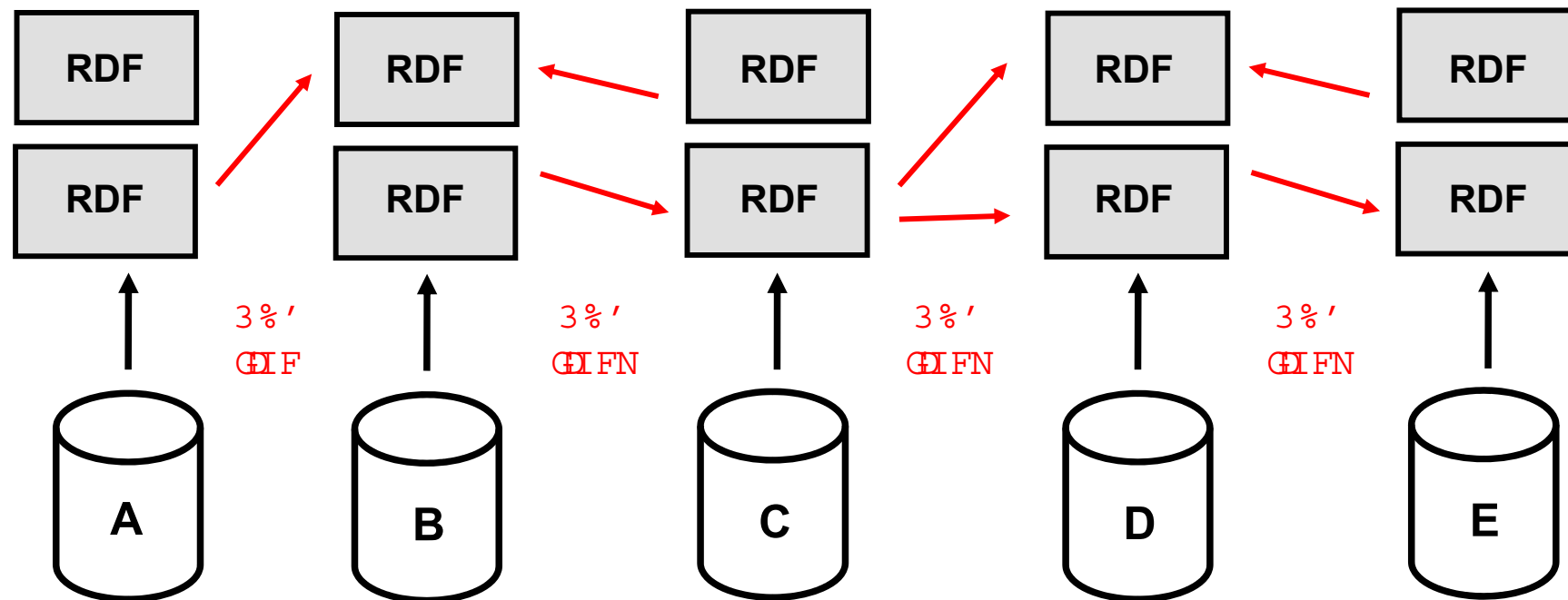
Web APIs

- 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
 - analyzes the APIs and published reports





- 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)

Legend

Cross Domain

Geography

Government

Life Sciences

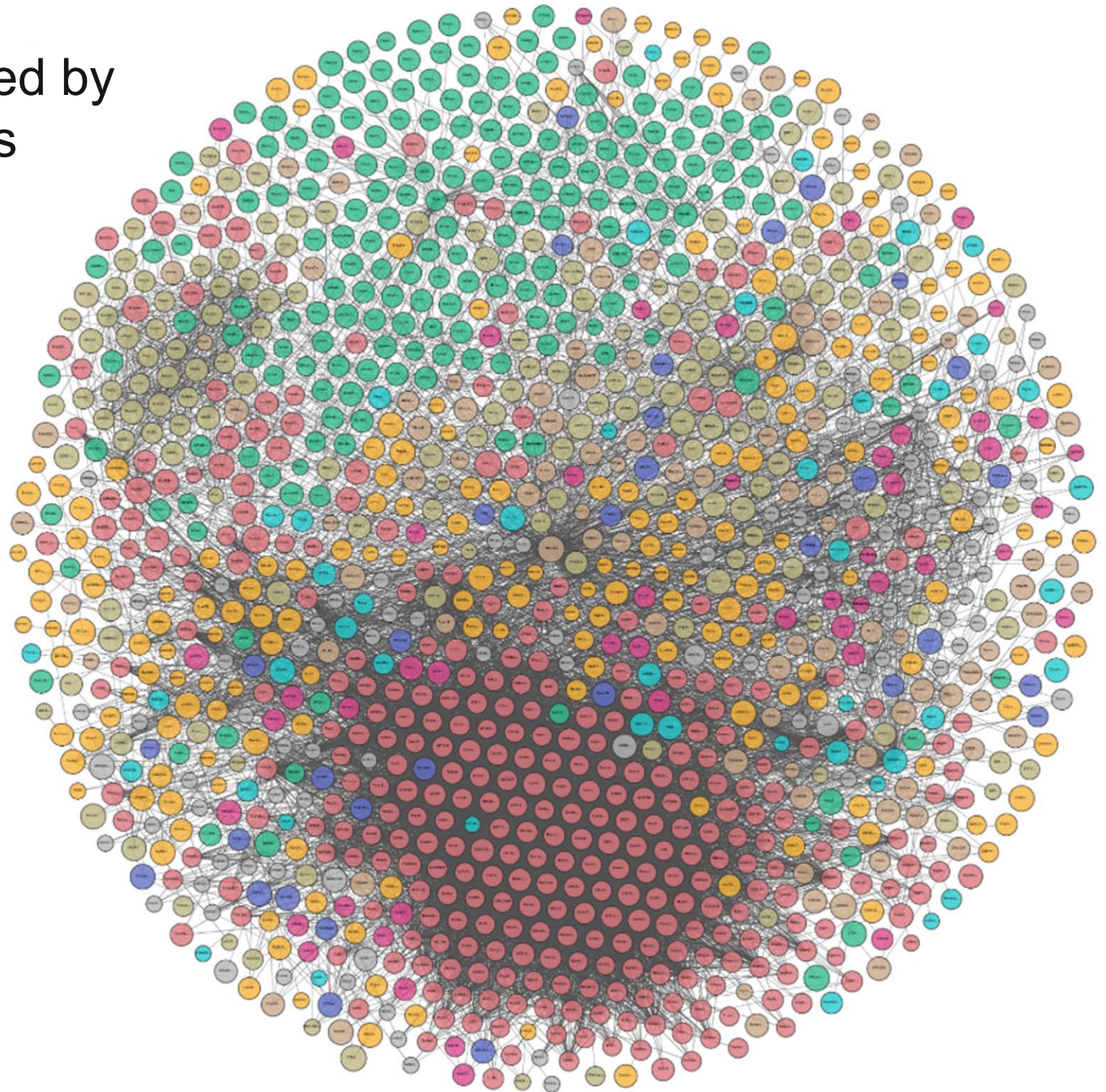
Linguistics

Media

Publications

Social Networking

User Generated



<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/Bing/Yandex webmaster recommendations?
- 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 topic
- 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 10 - 20 sources

1. Schema.org Local Business/Hotel Data – Structure, Dynamics, and Applications (Yen-Chun, 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.
- Kärle E., Fensel A., Toma I., Fensel D. (2016) Why Are There More Hotels in Tyrol than in Austria? Analyzing Schema.org Usage in the Hotel Domain. In: Information and Communication Technologies in Tourism 2016.
- <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/>

3. Schema.org Event Data – Structure, Dynamics, and Applications (**Martin Giovanni**, 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.
- Foley J, Bendersky M, Josifovski V (2015) Learning to extract local events from the web. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 423–432
- <https://developers.google.com/search/docs/advanced/structured-data/event>
- <http://webdatacommons.org/structureddata/schemaorgtables/>

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

- Brickley, Dan, Matthew Burgess, and Natasha Noy. „Google Dataset Search: Building a search engine for datasets in an open Web ecosystem.“ The World Wide Web Conference. 2019.
- Benjelloun, Omar, Shiyu Chen, and Natasha Noy. „Google dataset search by the numbers.“ International Semantic Web Conference. Springer, Cham, 2020.
- Chapman, Adriane, et al. „Dataset search: a survey.“ The VLDB Journal 29.1 (2020): 251-272.
- <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.
- Peeters, Ralph, et al. „Using schema. org annotations for training and maintaining product matchers.“ Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics. 2020.
- Bizer, C., Primpeli, A. und Peeters, R. (2019). Using the Semantic Web as a source of training data. Datenbank-Spektrum, 19, 127-135.
- <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**)

- M. Maleshkova, C. Pedrinaci, and J. Domingue, “Investigating web APIs on the world wide web,” in Proc. of European Conference on Web Services (ECOWS), 2010.
- Sohan, S. M., Craig Anslow, and Frank Maurer. „A case study of web API evolution.“ 2015 IEEE World Congress on Services. IEEE, 2015.
- <https://www.programmableweb.com/api-research>

7. Linked Data – Topics, Dynamics, Best Practices, and Applications (**Lukas Michael**, mentor: **Chris**)

- Debattista, Jeremy, et al. „Evaluating the quality of the LOD cloud: An empirical investigation.“ Semantic Web 9.6 (2018): 859-901.
- Debattista, Jeremy, et al. „Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud.“ Companion Proceedings of The 2019 World Wide Web Conference. 2019.
- Herrera, et al. „BTC-2019: The 2019 Billion Triple Challenge Dataset.“ Proceedings of the International Semantic Web Conference. 2019.
- Schmachtenberg, Max, Christian Bizer, and Heiko Paulheim. „Adoption of the linked data best practices in different topical domains.“ International Semantic Web Conference. Springer, Cham, 2014.

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/Bing/Yandex webmaster recommendations?
- 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?
(Discuss in each chapter)

Citing different Types of Publications

- Journal article
 - Good to cite, peer-reviewed research results
 - Survey articles (very good for an overview)
- Conference and workshop paper
 - Good to cite, peer-reviewed current research results
- Books (sometimes cited)
 - Textbooks
 - Collections of articles/papers => Cite specific paper in book
- Web Pages
 - 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?

