

Analysis and design of a semantic modeling language to describe public data sources

Branislav Vidojevic, 08.04.2019, Munich Final Presentation (Advisor: Patrick Holl)

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de

Motivation

Research Methodology

Key Findings

Implemented Software

Evaluation

Summary & Outlook

Tim Berners–Lee 5-star Open Data



- Make your stuff available on the Web (whatever format) under an open license



- Make it available as structured data (e.g., Excel instead of image scan of a table)



- Make it available in a non-proprietary open format (e.g., CSV instead of Excel)

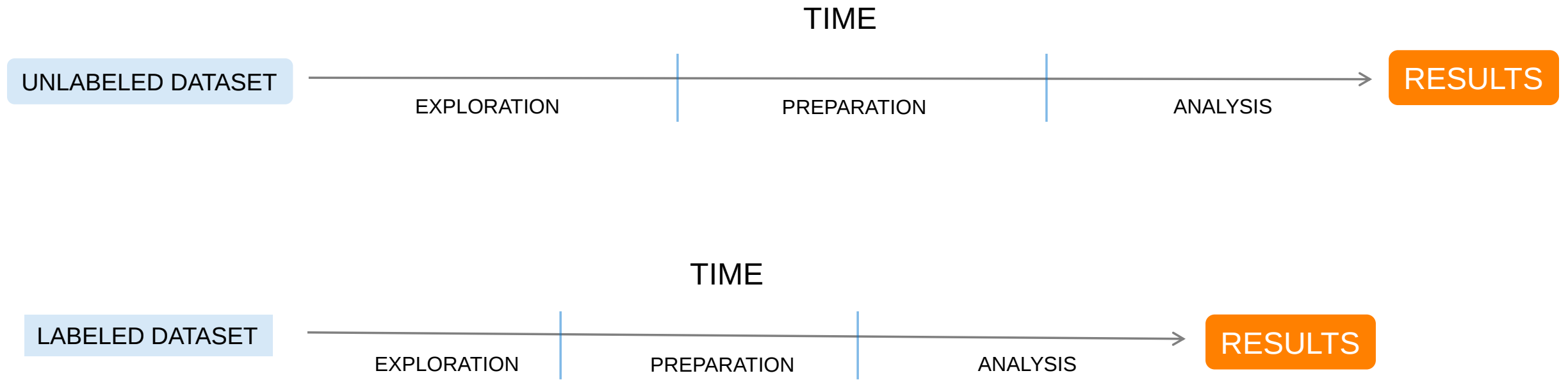


- Use URIs to denote things, so that people can point at your stuff



- Link your data to other data to provide context

Source: <https://5stardata.info/en/>



HOW TO OVERCOME AN EXPLORATION TIME GAP FOR UNLABELED DATASETS?

Motivation

Research Methodology

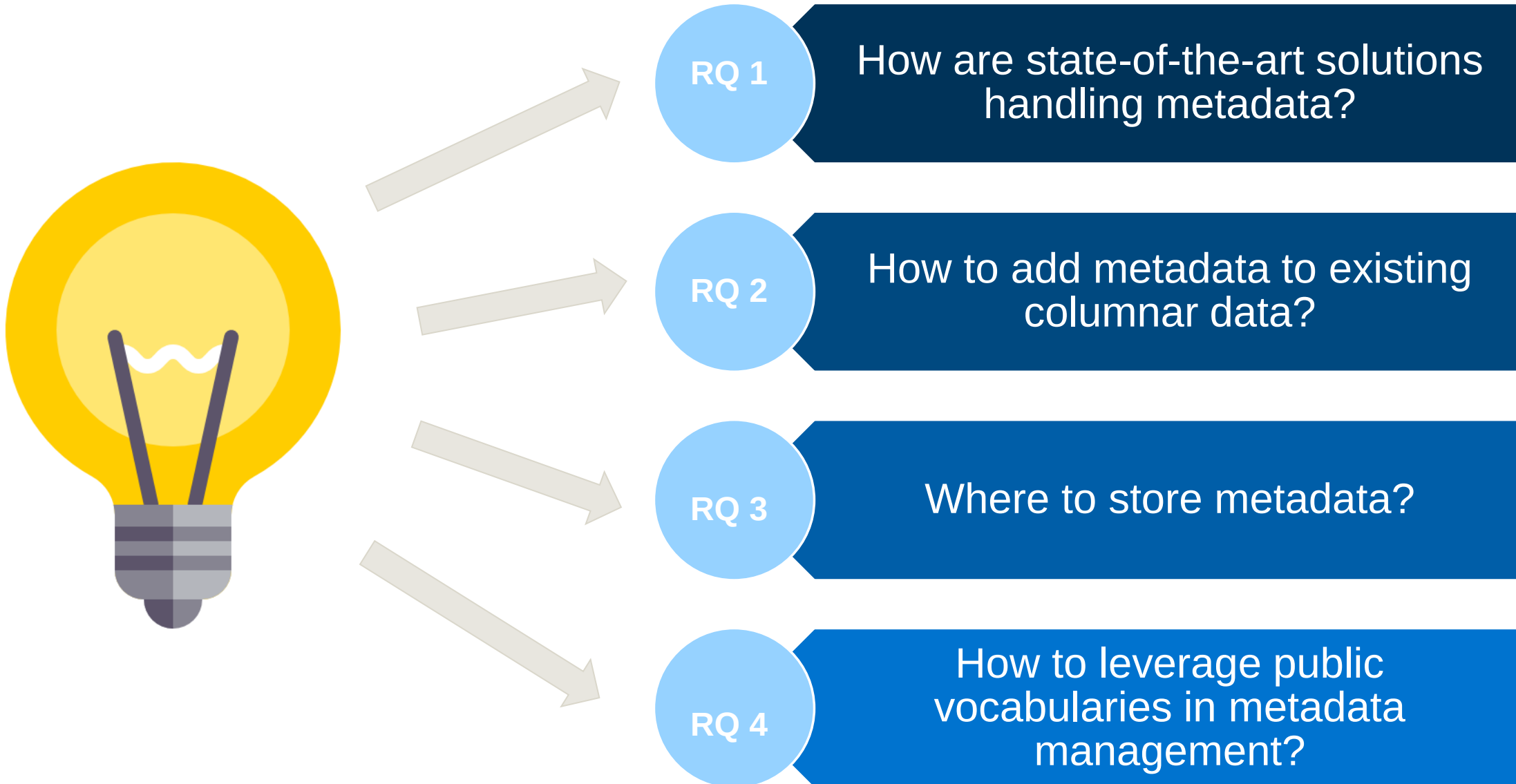
Key Findings

Implemented Software

Evaluation

Summary & Outlook

Research Questions



Analyzed Solutions



Deeper

**OPEN DATA INITIATIVE
(SAP, Microsoft, Adobe)**



Motivation

Research Methodology

Key Findings

Implemented Software

Evaluation

Summary & Outlook

Important Concepts

- User oriented vs. Business oriented
- Modularity
- Interface
- Data formats
- Data connectors
- Semantic Metadata

Design Decisions

- Modular and pluggable library
- Metadata in JSON format
- REST API for metadata management
- Extensible Metadata
- Metadata on dataset and attribute level
- Reuse existing knowledge

Motivation

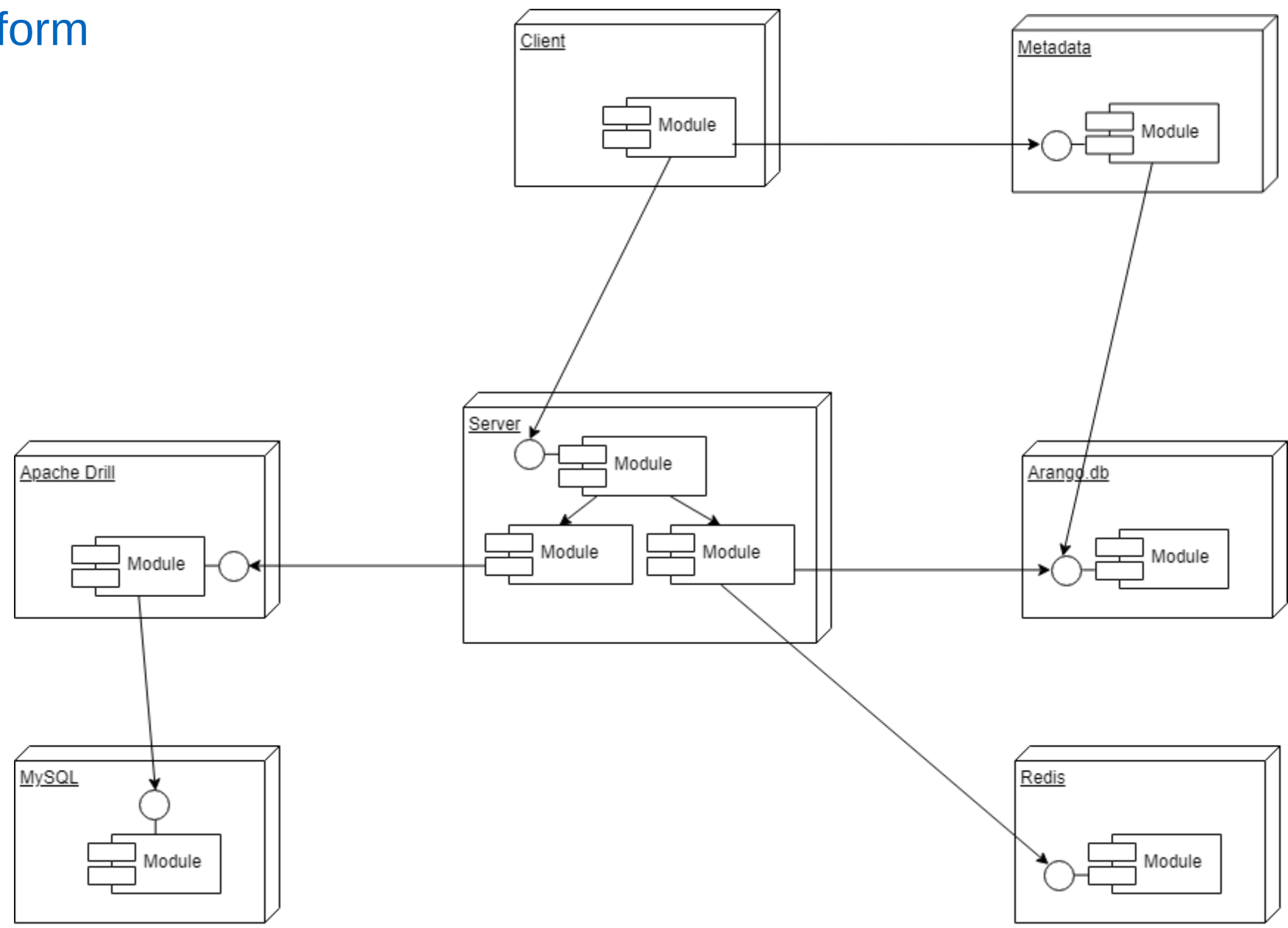
Research Methodology

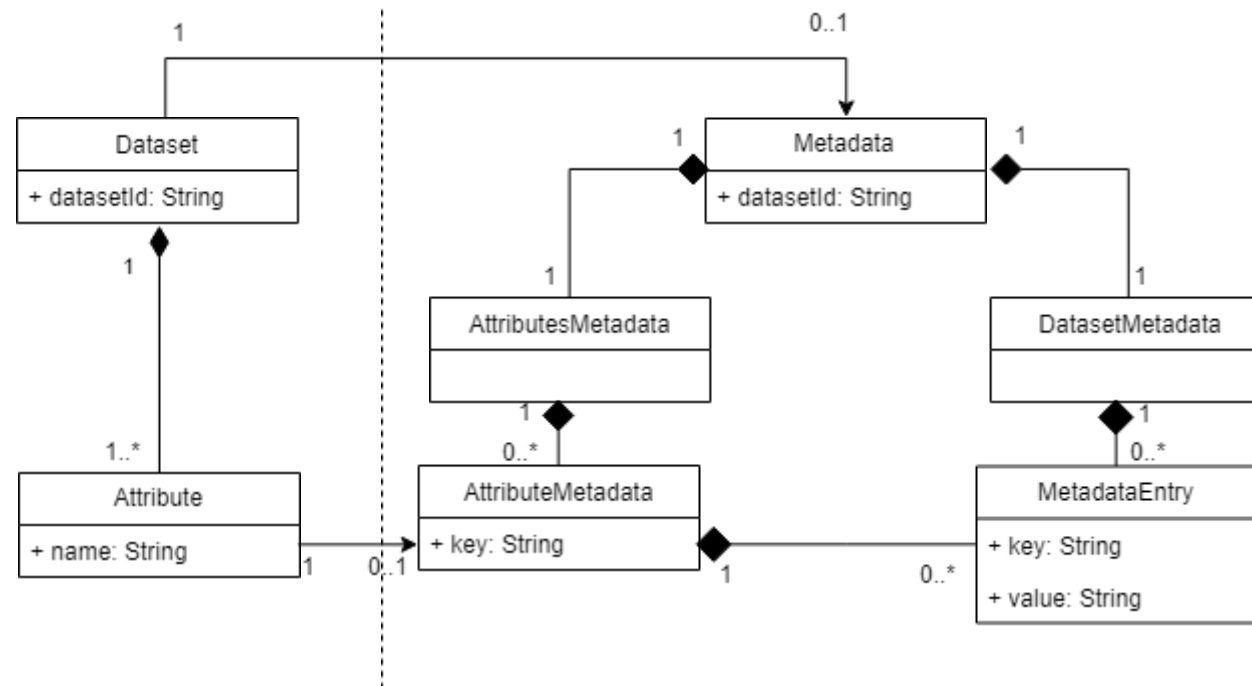
Key Findings

Implemented Software

Evaluation

Summary & Outlook





Midas Metadata Model Features

- Flat Data to JSON format

```
{  
  "name": "John",  
  "lastName": "Doe",  
  "address": "Strasse 12",  
  "postalCode": "12345",  
  "city": "Munich"  
}
```



```
{  
  "person": {  
    "name": "John",  
    "lastName": "Doe"  
  },  
  "location": {  
    "address": "Strasse 12",  
    "postalCode": "12345",  
    "city": "Munich"  
  }  
}
```

- Schema.org Entity Search via REST API
- Metadata enrichment with Schema.org Entity Data
- Create, Read, Update, Delete REST API on top of the Metadata model
- Comprehensive Swagger documentation

entity-search-controller Entity Search Controller		▼
GET	/entities	getEntities
flat-data-to-json-controller Flat Data To json Controller		▼
POST	/flat2json	convert
metadata-controller Metadata Controller		▼
POST	/metadata	createMetadata
GET	/metadata/{dataSourceId}	getMetadata
PUT	/metadata/{dataSourceId}	updateMetadata
DELETE	/metadata/{dataSourceId}	deleteMetadata
POST	/metadata/enrich	enrich

Motivation

Research Methodology

Key Findings

Implemented Software

Evaluation

Summary & Outlook

The 4-level scoring method

A simple method for rating the fitness of semantic annotation to annotate an unlabeled entity.

0

- There was no entity defined in Schema.org vocabulary that could describe an attribute

1

- Entity could be described only by its data type

2

- There is an entity that semantically describes an unlabeled entity, but it does not describe it unambiguously

3

- There is an entity in Schema.org vocabulary that fully describes an unlabeled attribute in the dataset

Example Calculation

Name	Last Name	Address	Postcode	Group Name	Result	Result Percentile
John	Doe	Wonderland 10b	11000	B3	45	20
Alice	Ranger	Moonstreet 21	99123	C1	55	3

Attribute	Score	Schema.org Entity URL
Name	3	https://schema.org/givenName
Last Name	3	https://schema.org/familyName
Address	3	https://schema.org/address
Postcode	3	https://schema.org/postalCode
Group Name	2	https://schema.org/name
Result	1	https://schema.org/Number
Result Percentile	0	/

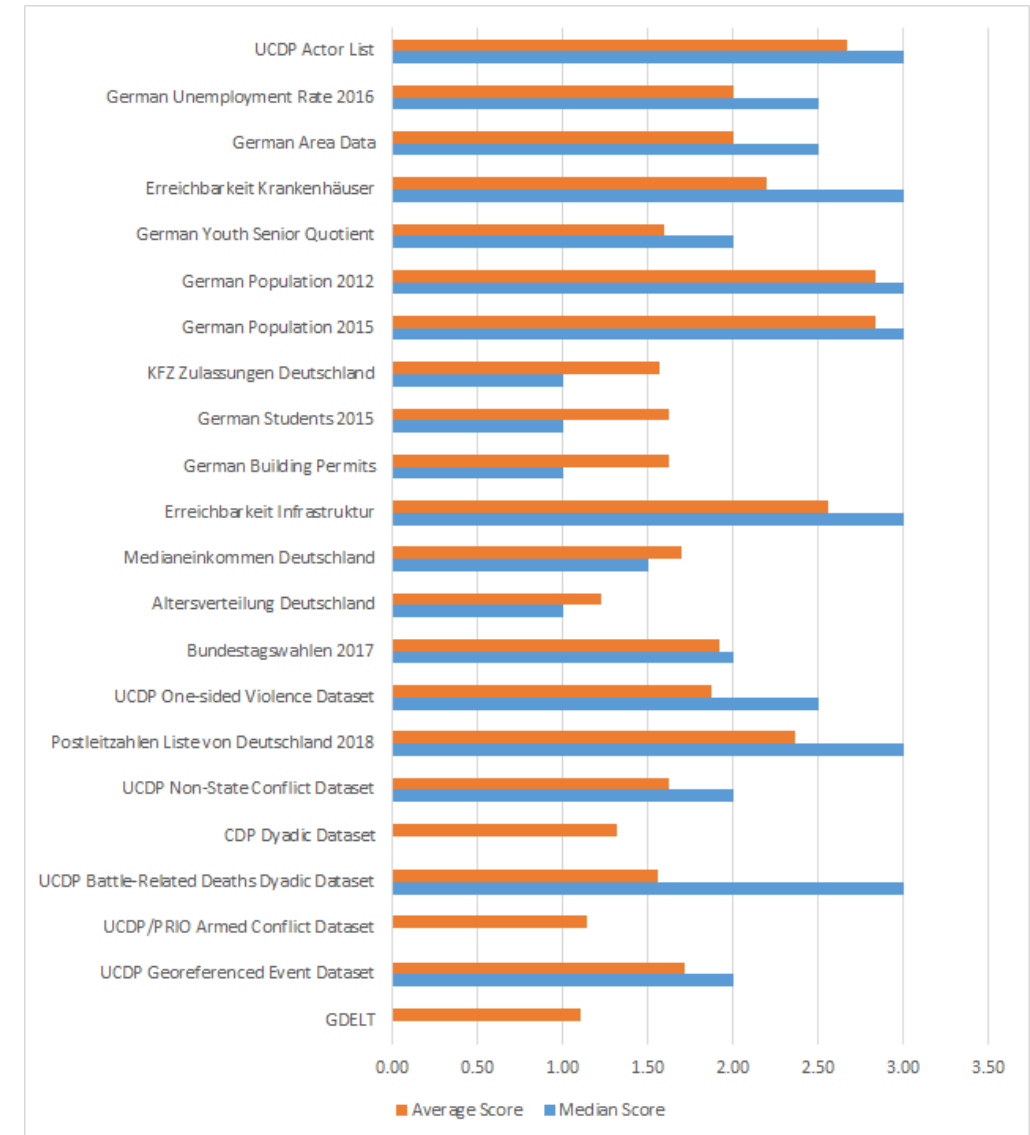
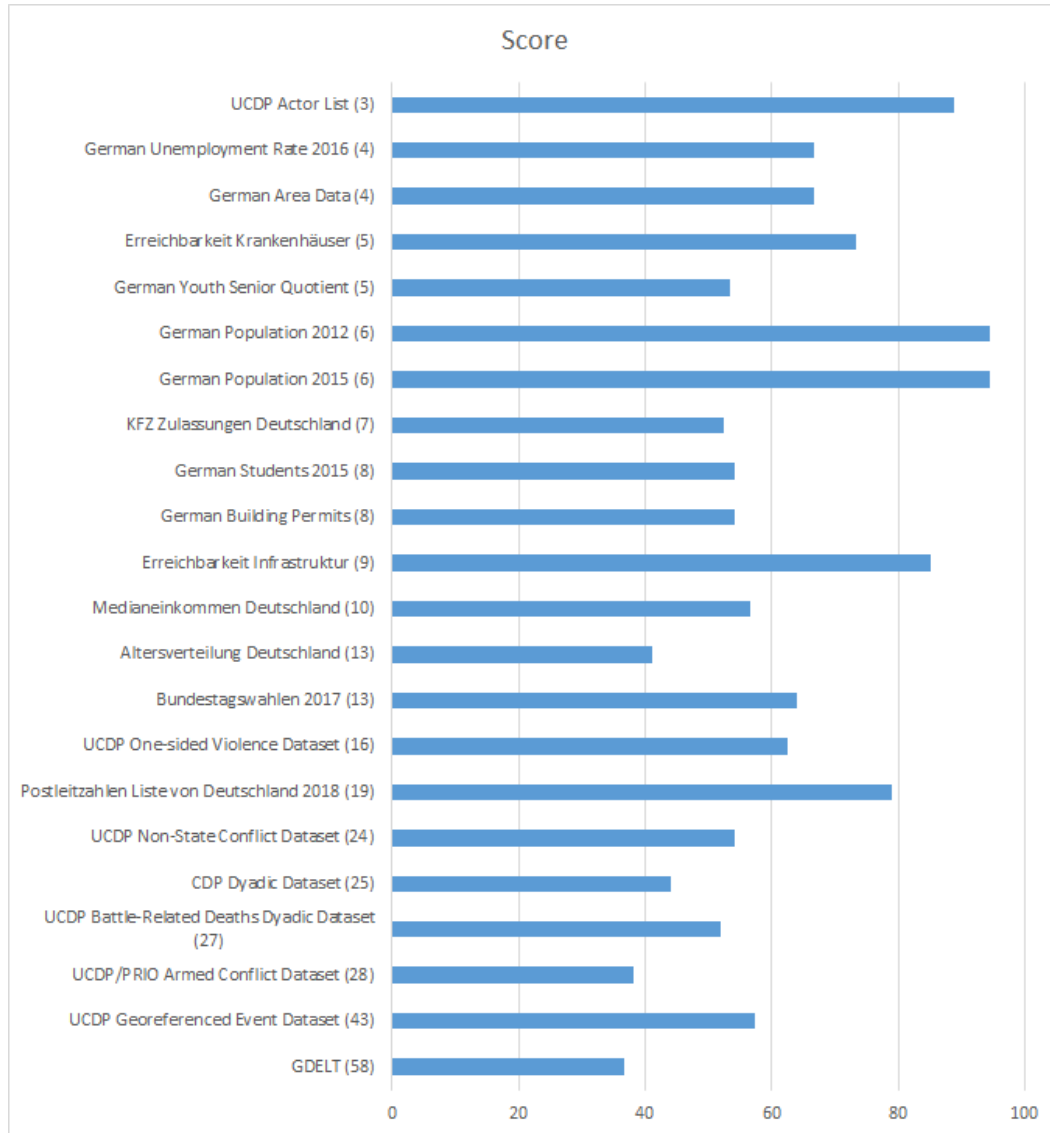
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} * 100$$

$$\begin{aligned} \min(x) &= 7 * 0 = 0 \\ \max(x) &= 7 * 3 = 21 \end{aligned}$$

$$x = \frac{15 - 0}{21 - 0} * 100$$

$$x' = 71$$

Results



Motivation

Research Methodology

Key Findings

Implemented Software

Evaluation

Summary & Outlook

Summary

- Every vendor models metadata based on their specific needs.
- Vendors did not identify semantic metadata as a key feature.
- Big players in industry recognized the need for model consolidation, and with it, metadata handling.
- It heavily depends on processes that specific vendor is using.
- Modularity is a way to go.
- Leverage global knowledge – it is becoming a trend.

Outlook

- Web application that could be used explicitly for metadata management.
- Matching datasets based on their metadata.
- Data integration automation.
- Support for other data formats.
- Keep agility and flexibility.



Thank You for Your
attention!

Questions?



B.Sc.

Branislav Vidojevic

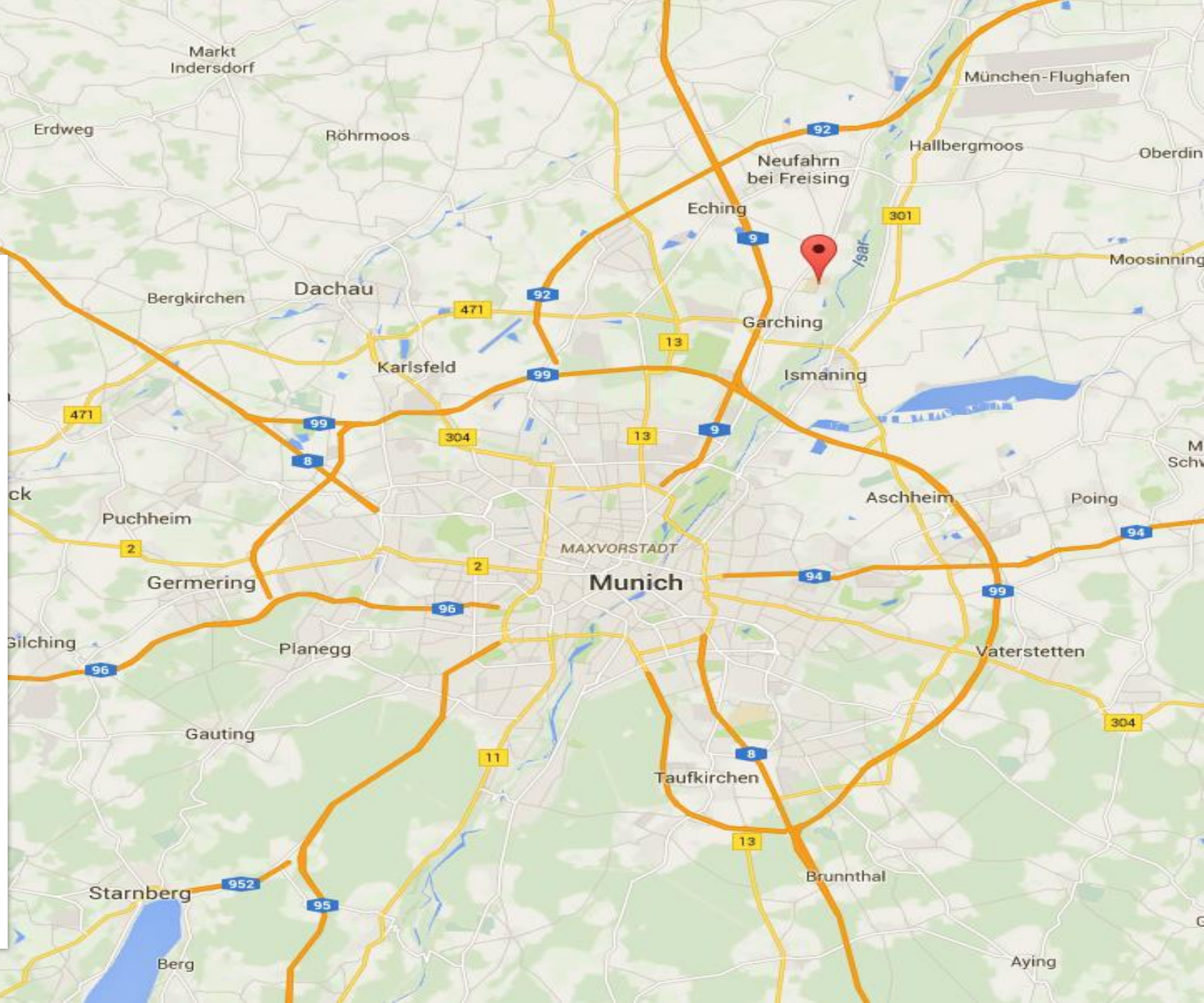
Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132

Fax +49.89.289.17136

matthes@in.tum.de
wwwmatthes.in.tum.de



- T. Berners-Lee, “Is your linked open data 5 star?” 2009, accessed 26-01-2019. [Online]. Available: <https://www.w3.org/DesignIssues/LinkedData.html>
- 5startdata, “5-star open data deployment scheme,” 2018, accessed 26-12-2018. [Online]. Available: <https://5stardata.info/en/>
- E. Wilde, C. Pautasso, and R. Alarcon, REST: Advanced Research Topics and Practical Applications. Springer, 01 2014.
- M. Rouse, “Metadata,” 2014, accessed 28-02-2019. [Online]. Available: <https://whatis.techtarget.com/definition/metadata>
- NISO, Understanding metadata. 4733 Bethesda Avenue, Suite 300, Bethesda, MD 20814 USA: NISO, 2004, iISBN: 1880124629. [Online]. Available: <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>
- W. Alpha, “Wdf (wolfram data framework),” 2018, accessed 20-12-2018. [Online]. Available: <https://reference.wolfram.com/language/guide/WDFWolframDataFramework.html>