

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

Branislav Vidojevic, 23.11.2018, Munich Kick-Off 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 Questions

Approach

Current State

Timeline

- **Data enrichment** is a general term that refers to processes used to enhance, refine or otherwise improve raw data [\[1\]](#).
- **MIDAS** [\[4\]](#) framework
 - there is no machine processing way to guess the semantics of the data
 - With increasing number of data columns and enrichers also increases the difficulty of managing the whole enrichment process
- There are good solutions for data enrichment
 - On a small and large scale
 - There are not many solutions that utilize semantic web
- We want to
 - Make the process of enrichment **easier** and **faster**
 - Make a job for data analysts easier
 - By shortening the time needed for getting to know the data
 - By automating the process to some extent
 - Contribute to Open Data Innovation movement

Motivation - Magic Quadrant for Data Integration Tools by Gartner, Inc [5]

A comprehensive list of:

- Vendor Strengths
- Vendor Cautions

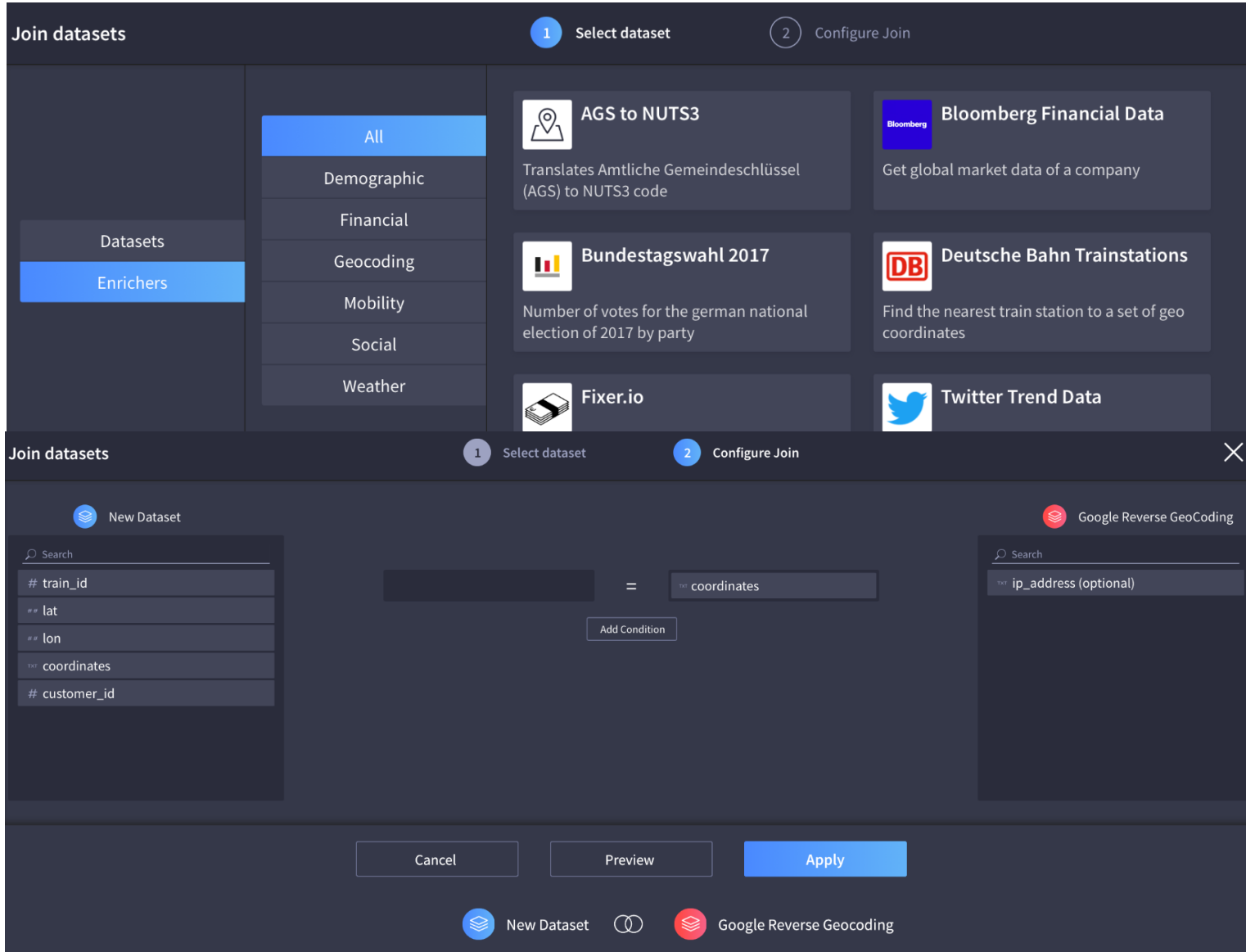
It feature solutions that are:

- Big players
- Expensive
- End-to-end
- Packed with features
- Takes time to be productive

“far beyond supporting extraction, transformation and loading (ETL) processes”



Motivation – how it works right now



The screenshot displays the 'Join datasets' interface in two states. The top state shows the 'Select dataset' step with a sidebar for 'Enrichers' (All, Demographic, Financial, Geocoding, Mobility, Social, Weather) and a grid of dataset cards including 'AGS to NUTS3', 'Bloomberg Financial Data', 'Bundestagswahl 2017', 'Deutsche Bahn Trainstations', 'Fixer.io', and 'Twitter Trend Data'. The bottom state shows the 'Configure Join' step, where a 'New Dataset' (with fields like train_id, lat, lon, coordinates, customer_id) is being joined with 'Google Reverse GeoCoding' (with field ip_address (optional)). The join is configured as 'train_id = coordinates'. Buttons for 'Cancel', 'Preview', and 'Apply' are at the bottom.

Official website:

<https://www.midas.science/>

- Join local and remote dataset
- Many REST based enrichers

Motivation – how it works right now

1. Customer imports its data (e.g. file, mongoDB etc.)
2. Customer selects one of the enrichers
3. Customers picks input parameters for specific enricher
4. Customer gets to explore enriched data

Pros

- Process is straightforward
- Results are quickly available
- It enriches the data

Cons

- Customer has to know which data type he has
- Customer has to know which data type he will get
- Customers knowledge about the data is never stored

1. Customer imports its data (e.g. file, mongoDB etc.)
- 2. Customer can explore potential enrichers (datasets)**
- 3. Customers picks one of the enrichers that he has input parameters for it**
4. Customer gets to explore enriched data

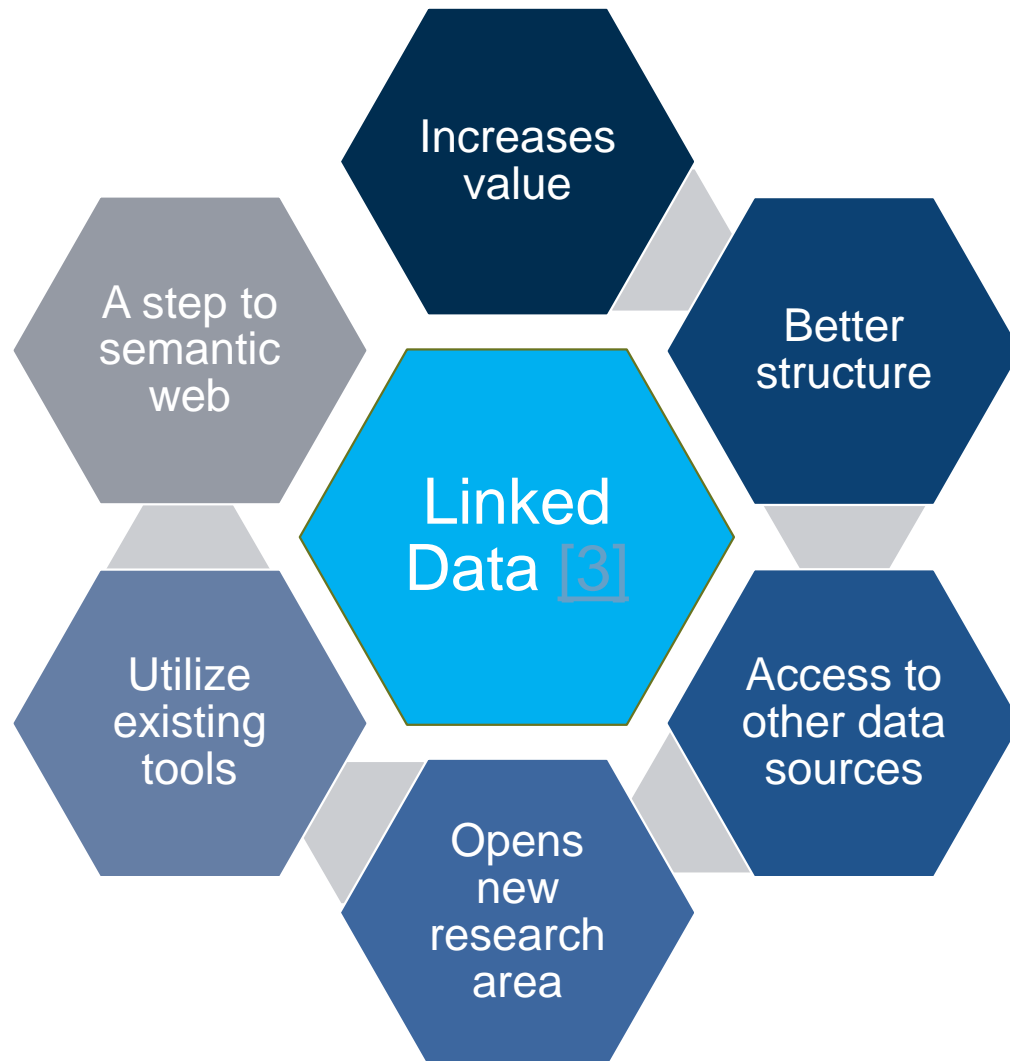
Pros

- Process is straightforward
- Results are quickly available
- It enriches the data
- **Customer can always check the type of data for each and every column**
- **Customer can export his data in JSON-LD format**
- **Customer can check which enrichers are available for usage (First step to automation)**

Cons

- More effort in the development of the enrichers (Only once per enricher)
- Existing data has to be described in the context of linked data and data type (Only initial data)

Why Linked Data?



- The Semantic Web (aka Web of data or Web 3.0)
- Link data from different sources
- Machines can understand links
- Schema.org [\[2\]](#) unifies common knowledge on many entities

Motivation

Research Questions

Approach

Current State

Timeline

RQ 1: How state of the art solutions for data integration handle metadata?



RQ 2: Can Linked Data be used to improve data integration process?



RQ 3: Which metadata to use to describe existing data?



RQ 4: How to attach metadata to existing dataset?

Motivation

Research Question

Approach

Current State

Timeline

RQ 1: How state of the art solutions for data integration handle metadata?

Analysis of tools and initiatives and their approach to metadata handling

- Wolfram Data Framework <https://www.wolfram.com/data-framework/>
- RapidMiner <https://rapidminer.com/>
- Segment Customer Data Integration (CDI) <https://segment.com/>
- Dremio <https://www.dremio.com/>
- Deeper <https://www.cs.sfu.ca/~jnwang/papers/TR-deeper.pdf>
- Google Dataset Search (Beta) <https://toolbox.google.com/datasetsearch>
<https://www.nature.com/articles/d41586-018-06201-x>
- Open Data Initiative <https://www.microsoft.com/en-us/open-data-initiative>
- Data Transfer Project <https://datatransferproject.dev/>

RQ 2: Can Linked Data be used to improve data integration process?

Literature overview

- Angela Lausch, Andreas Schmidt, and Lutz Tischendorf. **Data mining and linked open data - new perspectives for data analysis in environmental research.** Ecological Modelling, 295:5 – 17, 2015. Use of ecological indicators in models.
- Data Transfer Project. **Data transfer project overview and fundamentals**, July 2018.
- IBM Unified Governance I& Integration. **Data integration reaches inflection point: Survey results.**
- Markus Lanthaler and Christian Gutl. **On using json-ld to create evolvable restful services.** In Proceedings of the Third International Workshop on RESTful Design, WS-REST '12, pages 25–32, New York, NY, USA, 2012. ACM.
- Pei Wang, Yongjun He, Ryan Shea, Jiannan Wang, and Eugene Wu. **Deeper: A data enrichment system powered by deep web.** In Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18, pages 1801–1804, New York, NY, USA, 2018. ACM.
- S. Subhashree, Rajeev Irny, and P. Sreenivasa Kumar. **Review of approaches for linked data ontology enrichment.** In Atul Negi, Raj Bhatnagar, and Laxmi Parida, editors, Distributed Computing and Internet Technology, pages 27–49, Cham, 2018. Springer International Publishing.
- Sebastian Walter, Christina Unger, Philipp Cimiano, and Daniel Bar. **Evaluation of a layered approach to question answering over linked data.** In Proceedings of the 11th International Conference on The Semantic Web - Volume Part II, ISWC'12, pages 362–374, Berlin, Heidelberg, 2012. Springer-Verlag.
- Vanessa Lopez, Christina Unger, Philipp Cimiano, and Enrico Motta. **Evaluating question answering over linked data.** Journal of Web Semantics, 21:3 – 13, 2013. Special Issue on Evaluation of Semantic Technologies.
- Walter Renteria-Agualimpia, Francisco J. Lopez-Pellicer, Pedro R. Muro-Medrano, Javier Nogueras-Iso, and F. Javier Zarazaga-Soria. **Exploring the advances in semantic search engines.** In Andre Ponce de Leon F. de Carvalho, Sara Rodriguez-Gonzalez, Juan F. De Paz Santana, and Juan M. Corchado Rodriguez, editors, Distributed Computing and Artificial Intelligence, pages 613–620, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.
- William Tunstall-Pedoe. **True Knowledge: Open-domain question answering using structured knowledge and inference.** AI Magazine, 31(3):80–92, 2010.
- ...
- Many websites, blogs etc.

RQ 3: Which metadata to use to describe existing data?

Based on RQ1 and RQ2 find appropriate metadata to describe existing data

- Type
- Semantics
- Relations

RQ 4: How to attach metadata to existing dataset?

Implementation of the proof of concept in MIDAS project

- Library for converting flat data to JSON
- Enrichment of JSON data with context data – JSON-LD
- Demo of functionalities in MIDAS web application
- Documentation
- Pros and Cons analysis

Motivation

Research Question

Approach

Current State

Timeline

Phase 1 – tools and literature overview

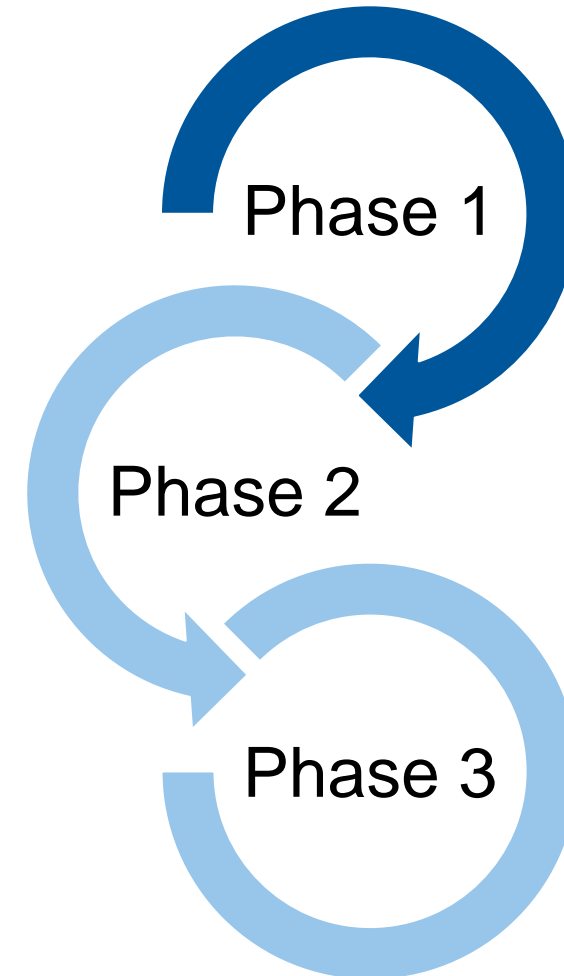
- **What is done**
 - Good literature overview
 - Good state of the art tools overview
 - Good MIDAS framework overview
- **To be done**
 - Writing everything down
 - Listing key takeaways

Phase 2 – Implementation and documentation

- Libraries Implementation
- Proof of Concept
- Documentation

Phase 3 – Analysis and Thesis writing

- Pros and Cons Analysis
- Conclusion



Motivation

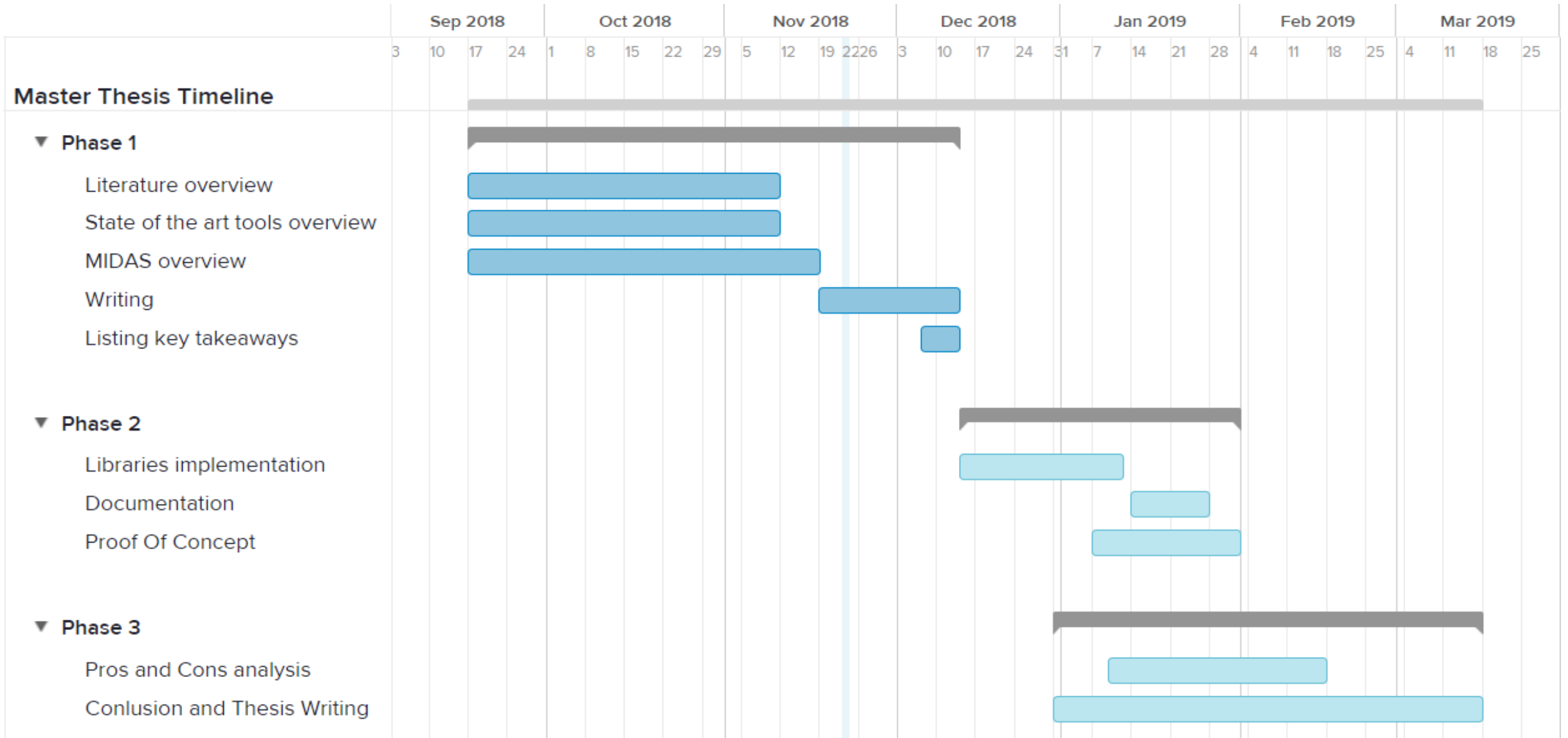
Research Question

Approach

Current State

Timeline

Proposed Timeline



Created with <https://www.teamgantt.com/>



Thank You for Your
attention!

Questions?

- [1] What is Data Enrichment? - Definition from Techopedia. (n.d.). Retrieved October 29, 2018, from <https://www.techopedia.com/definition/28037/data-enrichment>
- [2] Schema.org Vocabulary (n.d.). Retrieved November 10, 2018, from <https://schema.org/>
- [3] Linked Data - Connect Distributed Data across the Web (n.d.). Retrieved November 10, 2018, from <http://linkeddata.org/>
- [4] Midas – Get more out of your data (n.d.). Retrieved November 11, 2018, from <https://www.midas.science/>
- [5] Gartner - Magic Quadrant for Data Integration Tools (19 July 2018). Retrieved November 18, 2018, from <https://www.gartner.com/doc/reprints?id=1-5F1U3D0&ct=180907&st=sg>



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

