TU WIEN Informatics

Trends in Tunnel Information Modelling

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Masterstudium Business Informatics

eingereicht von

Nico Henglmüller, BSc.
Matrikelnummer 1129267

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Christian Huemer
Mitwirkung: Univ.Prof. Dipl.-Ing. Mag. Dr.techn. Alexandra Mazak-Huemer

Wien, 17. April 2021

Nico Henglmüller	Christian Huemer

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Trends in tunnel information modelling

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Master programme Business Informatics

by

Nico Henglmüller, BSc.
Registration Number 1129267

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Christian Huemer
Assistance: Univ.Prof. Dipl.-Ing. Mag. Dr.techn. Alexandra Mazak-Huemer

Vienna, 17th April, 2021

Nico Henglmüller	Christian Huemer

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Nico Henglmüller, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 17. April 2021

Nico Henglmüller

Danksagung

Ich möchte meinem Betreuer Ao.Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Christian Huemer für seine tatkräftige Unterstützung im gesamten Verlauf meiner Arbeit danken. Ein besonderer Dank gebührt außerdem Univ.Prof. Dipl.Ing.in Mag.arer.soc.oec. Dr.in techn. Alexandra Mazak-Huemer für die fachspezifische Beratung im Bereich Tunnelbau sowie Herrn Georg Heiler MSc. für die fachliche Unterstützung im Bereich Data Science. Zusätzlich will ich Dipl.Ing.in Ingrid Kriegl & Mag. Friedl Ebner für die Möglichkeit und die immerwährende Motivation zur Absolvierung meines Studiums danken. Abschließend gilt mein Dank meiner Lebensgefährtin und meiner Familie, die mich auf dem Weg des Studiums an der Technischen Universität Wien begleiteten und fortwährend unterstützten.

Acknowledgements

I would like to express my very great appreciation to my advisor Ao.Univ.Prof. Mag. rer.soc.oec. Dr. rer.soc.oec. Christian Huemer for his ongoing commitment and constructive suggestions during the work on my thesis. I would also like to thank Univ.Prof. Dipl.Ing.in Mag.a rer.soc.oec. Dr.in techn. Alexandra Mazak-Huemer, for her very valuable advice and assistance in the field of information models in infrastructure facilities. Also, I am thankful for the support and consultation received from Georg Heiler, MSc. in the area of data science. Moreover, I would like to offer my special thanks to Dipl.Ing.in Ingrid Kriegl and Mag. Friedl Ebner for enabling and motivating me to finish my master studies. Finally, I wish to thank my girlfriend and family for their support and encouragement throughout my study.

Kurzfassung

Spätestens durch die konsequente Anwendung von Informationssystemen über die Projektabschnitte Planen, Bauen und Betreiben hinweg ist die Digitalisierung zu einem zentralen Bestandteil im Infrastrukturbau geworden. Angetrieben von staatlichen Digitalisierungs- und Umweltstrategien ist das Forschungsinteresse im Bereich des *Building Information Modelling (BIM)* im Infrastrukturbau konstant angestiegen. In Anbetracht dieser Bestrebungen und dem nachhaltigen Nutzen von digitalen Bauwerksmodellen ermittelt die vorliegende Diplomarbeit die Relevanz des Themas *Informationsmodelle im Tunnelbau* innerhalb der internationalen Forschung. Um relevante wissentschaftliche Publikationen zu identifizieren und kategorisieren wird die wissenschaftliche Methodik *Design Science* angewandt, welche die Diplomarbeit in zwei Teile aufteilt. Der erste Abschnitt befasst sich mit der Durchführung einer *Systematic Mapping Study (SMS)*, welche nach den Richtlinien von Petersen et al. [PVK15] durchgeführt wird. Ziel der Klassifizierung relevanter Literatur im Forschungsbereich *Tunnel Information Modelling (TIM)* ist es (i) aktuelle und vergangene Forschungsaktivitäten zusammenzufassen, (ii) relevante Forschungstrends hervorzuheben und (iii) wenig beachtete Themen im Forschungsbereich aufzuzeigen. Der zweite Teil der vorliegenden Arbeit widmet sich der Automatisierung der Systematic Mapping Study mit Hilfe eines *Data Mining (DM)* Ansatzes. Dem Prozessmodells zur Datenanalyse *Cross Industry Standard Process for Data Mining (CRISP-DM)* folgend wird ein Softwareartefakt erstellt, welches durch die automatisierte Identifizierung relevanter, wissenschaftlicher Arbeiten jederzeit den aktuellen Forschungsstand widerspiegelt. Eine abschließende Evaluierung der Klassifizierungsergebnisse misst anhand von statistischen Metriken in welchem Ausmaß das *Data Mining* Artefakt die Klassifizierung der Mapping Study reproduzieren kann. Die Erkenntnisse dieser Diplomarbeit können wie folgt zusammengefasst werden (i) die Anzahl der publizierten wissenschaftlichen Arbeiten im Forschungsbereich TIM stiegen im Betrachtungszeitrum zwischen 2011 und 2019 konstant, (ii) die größten Treiber dieses Trends sind staatliche Digitalisierungs- und Umweltstrategien, (iii) die Mehrzahl relevanter, wissenschaftlicher Publikationen beschäftigt sich mit der Anwendung von TIM im Zuge der Planungsphase, (iv) die Anzahl der Publikationen mit Bezug zum mechanischen Vortrieb übertreffen jene des konventionellen Vortriebs, (v) die Kombination des Rankingalgorithmus Okapi BM25 mit Supervised Learning Modellen zeigt durchwegs positive Ergebnisse bei der Identifizierung relevanter Publikationen im Forschungsbereich TIM und (vi) die Klassifizierung von Publikationen in Bezug auf Vortriebsmethode und Projektphase auf Basis von Schlagworten erweist sich

als ineffizent. Die Resultate der vorliegenden Arbeit können Forscher dabei unterstützen (i) den aktuellen Forschungsstand und zukünftige Forschungstrends auszumachen, (ii) offene Forschungsfelder im Bereich TIM zu erkennen und (iii) weiterführende Forschung im Bereich der Textklassifizierung durchzuführen.

Abstract

Using Information Systems (IS) throughout the life cycle of constructed facilities paved the way of computer-aided engineering in the field of automation in construction. Especially national digitization strategies and governmental mandates accelerated the application of *Building Information Modelling (BIM)* throughout the life cycle of civil infrastructure facilities. In the view of potential benefits of information models for tunnel life cycle, this study aims to identify the relevance of *Tunnel Information Modelling (TIM)* in the scientific community. Therefore, a design science methodology is applied in order to identify and classify relevant research work. The first part of the design science approach is a *Systematic Mapping Study (SMS)* following the guidelines by Petersen et al. [PVK15]. We conduct the literature classification in the field of TIM (i) to summarize research activities, (ii) to identify relevant publication trends and (iii) to uncover possible blind spots of the scientific work. In the second part of the design science methodology we develop a data mining artifact to automatize the SMS by adopting the data analysis process model *Cross Industry Standard Process for Data Mining (CRISP-DM)*. In order to evaluate the results of the software artifact we use well-established classification metrics to measure how well the *Data Mining (DM)* approach may be used to automatize the literature identification and classification and, therefore, provide a set of relevant, state-of-the-art scientific studies. The main findings of this thesis are (i) the research effort in the area of TIM constantly increased between 2011 to 2019, (ii) governmental digitization and environmental strategies are major drivers of the increasing number of TIM related studies, (iii) the majority of studies concentrate on the application of information models during the design phase of a tunnel project, (iv) the usage of information models for continuous excavation receives more attention of the scientific community than its conventional counterpart, (v) a combination of Okapi BM25 and supervised text classification models yield positive performance measures by identifying relevant studies in the domain of TIM, and (vi) our term-based identification approach is not able to classify studies in regards to tunnel life cycle phases and excavation methods. The results of this thesis can assist researchers (i) to identify trends and state-of-the-art of the research domain, (ii) to identify open research issues and (iii) to propose new studies in the field of text classification.

Contents

CHAPTER 1

Introduction

1.1 Background

The application of Information Systems theories and concepts throughout the life cycle of constructed facilities paved the way of computer-aided engineering in the field of automation in construction [Els20a]. Starting in 2011 several countries introduced strategies and governmental mandates to apply BIM in public infrastructure projects to accelerate the introduction of digitization in the Architecture, Engineering, Construction, Owner and Operator (AECOO) industry [UK 11] [Cab16] [Bun15].

Additionally to the governmental strategies Bradley et al. [BLLD16] identify the potential increase of efficiency and environmental objectives as main drivers to apply BIM in the AECOO industry. Especially the infrastructure construction industry depends on 2D or 3D based design and static documentation in a large volume. This introduced several problems in large infrastructure projects as the exchange of digital assets such as data models, financial aspects, planning or logistics were practical non existent. Therefore, BIM introduces a shared, digital representation of the infrastructure facility which enables reliable information exchange between all parties during the whole asset life cycle [Joh16].

In the view of potential benefits of information models for the tunnel life cycle, this study aims to identify the relevance of TIM in scientific communities. Therefore, a SMS is carried out by gathering and analyzing abstracts of published literature in the interdisciplinary domain of IS and engineering focusing on TIM. This SMS is one of two parts of the applied design science research methodology. The second part of the design science approach is the realization of a data science artifact which automates the identification of relevant literature in the TIM domain.

1.2 Problem Statement

TIM is used to create shareable digital representations of physical facilities and their functional characteristics in the domain of subsurface engineering [VSS14]. During the last two decades the research works in the field of information models in the tunnelling domain continuously increased. In parallel, several governments developed digitization strategies for public funded infrastructure facilities such as bridges, roads, railways, and tunnels to meet environmental objectives. Therefore, the main objective is to reduce emissions and increase the efficiency during the whole life cycle of an infrastructure facility [BLLD16]. As a result numerous studies have been published in the interdisciplinary field between IS and subsurface engineering. The publications discuss different aspects of TIM and have been published in different venues by different research communities. However, so far no studies have been published summarizing the research activities with the topic TIM. In contrast scientific and industry stakeholders such as scientists, tunnel designers and construction engineers demand a state-of-the-art set of relevant scientific publications.

1.3 Methodology

We apply an IS design science research methodology as introduced by [HRM+04] which resides in the interdisciplinary environment of engineering and IS. The design science method consists of two parts:

- SMS: following the guidelines by Petersen et al. [PVK15] we conduct a literature classification in the field of TIM to provide an overview of publication forums and trends.
- Software Artifact: we apply a well established data mining process to implement the data science software artifact [Rau19] [GUP14].

In the design cycle we conduct technical experiments to continuously assess the results of the SMS to the results of the software artifact [PRTV12]. Based on the evaluation results we refine the software artifact and apply the design cycle according to the design science research methodology by Hevner [Hev07]. In chapter 3 we present the research methods in detail and give a thorough description how we apply those.

1.4 Research Scope

The aim of this study is to provide an overview of the research in the field of TIM and present an artifact in order to automatize the mapping process. Therefore, the study focuses on peer reviewed publications and conference proceedings written in German or English language. We explicitly exclude research works in other languages than German or English as well as grey literature. Moreover, it is not intended to include publications

in the general area of BIM or research works in the broader field of information model usage in infrastructure projects.

1.5 Research Objectives

In this thesis we intend to provide an overview and classify research works regarding information models in the life cycle phases design, construction and operation of a tunnel facility. Therefore, the SMS is fundamental for our applied design science methodology. With the evaluation of a software artifact we show how well a Data Mining (DM) approach may be used to automatize literature identification and classification to support the mapping process [SNM+18].

1.6 Thesis Organization

First, we have framed the study and described the topic information models in general in the introduction chapter. Then, we investigate related work in the domain of data analytics process models and BIM with focus on literature reviews in the research field. In the third chapter we go into detail about the theory and how we applied certain research methodologies. We present the results of both parts, the SMS and the software artifact, in the fourth chapter. In the fifth chapter we summarize, conclude and provide recommendations for future research.

CHAPTER 2

Related work

In this chapter we first take a closer look at Systematic Literature Review (SLR) and major differences to Systematic Mapping Studies. Furthermore, we analyze a sample of literature reviews in the area of BIM which has been provided by domain experts. This analysis focuses on the review methodology and process in order to identify major differences to the guidelines of Petersen et al. [PVK15]. After the analysis we outline well established data analytics process models based on recommendations by Rauber [Rau19].

2.1 Systematic Literature Reviews and Systematic Mapping Studies

The interdisciplinary character of the topics TIM and BIM enables us to apply well established research methodologies from the IS and Software Engineering (SE) disciplines. When we take a look at the SE research domain we observe that well established methodologies such as SLRs are applied for a large majority of literature reviews [KPBB+09]. Since 2004, Kitchenham published multiple guidelines for conducting reviews with the focus on an evidence based, auditable and repeatable methodology [Kit04]. These guidelines are grounded in Evidence Based Software Engineering (EBSE) principles which Kitchenham derived from evidence based research methods originated from the medicine discipline. The aim of a SLR is to provide an evidence-based evaluation as well as an interpretation of all available literature relevant to a topic area or research question [KC07].

In this thesis we want to structure the research progress and trends in the area of information models in the tunnelling domain by counting and classifying publications. Therefore, we use the systematic mapping methodology proposed by Petersen et al. [PVK15] which provides guidelines for research question specification, search strategies,

References	Search strategy	Review sources	Literature types	Review Methodology
[VSS14]	keyword based database search	ASCE Online Library EBSCO Host Elsevier Emerald Insight ScienceDirect SpringerLink Tayler & Francis Online Wiley Online Library	Conference Proceedings Journals	No references
[BLLD16]	keyword based database search	Engineering Village ScienceDirect Scopus Web of Science	Conference Proceedings Industry standards Journals	No references
[WPL19]	keyword based database search	Web of Science	Conference Proceedings Journals	No references

Table 2.1: Applied review process and search strategy in related literature reviews.

quality evaluation, and classification. One major difference when comparing mapping studies to systematic reviews is that research questions in systematic mappings are more general in order to discover trends [PVK15]. Research questions in systematic reviews on the other hand are more specific as they aim to aggregate evidence and therefore detailed objectives are required. In their guidelines Petersen et al. [PVK15] argue that articles with no empirical evidence would not be respected in SLRs, but for systematic maps those are important in order to spot trends of topics being worked on. This is of major interest when we discuss the classification by literature type facets following Wieringa et al. [WMMR06].

2.2 Review Methodologies applied to BIM

The following literature reviews have been proposed by domain experts as examples for literature reviews in the field of Building Information Models. We look at the proposed literature reviews in the field of BIM in order to find commonly applied methodologies by focusing on the manner in which the literature reviews were conducted. Especially the used search strategies, exclusion and inclusion criteria are of interest to us. Before we conduct the systematic mapping it is important to identify if there are well established review methodologies in the field of research as well as used sources for literature reviews. Such publication forums are abstract citation databases or online libraries where journals, book series or conference proceedings in the field of research can be found. Table 2.1 summarizes our analysis of the literature reviews in regards to applied search strategies, publication sources, relevant literature types and a reference to the used review methodology.

Volk et al. [VSS14] give an overview about recent research activities in the field of BIM

for existing buildings by conducting a literature review. The methodology section outlines the used process, but does not reference any well established guidelines in the field of interest. In order to find relevant literature Volk et al. [VSS14] conduct a keyword based search in 8 different publication databases or online libraries. Then, Volk et al. [VSS14] define inclusion and exclusion criteria to identify journal articles, conference proceedings and books which are relevant in regards to their research questions.

Bradley et al. [BLLD16] conduct a systematic literature research in the field of BIM in infrastructure to provide an overall review of research works and industry standards. The article applies a keyword based database search and rates the relevance of found literature. In order to rate the literature Bradley et al. [BLLD16] assign points based on the relevance of the article in regards to BIM in the infrastructure. This rating system represents the inclusion and exclusion criteria as at least 3 of 5 points are required for a journal article, conference proceeding and industry standards to be considered relevant. Literature with a rating with 2 points or less is excluded due to a lack of relevancy. Though, Bradley et al. [BLLD16] do not reference a methodology they apply to conduct the systematic literature review.

Wang et al. [WPL19] review literature regarding the topic BIM & Geographical Information System (GIS) and their integration in a sustainable built environment. In the methodology section the study does not reference a well established guideline which the article follows in order to review literature in the research area. Moreover, Wang et al. [WJMB11] do use a keyword based search strategy and apply inclusion & exclusion criteria on journal articles and conference proceedings. The article does barely describe the inclusion and exclusion criteria as well as the review process itself. When we compare the number of literature sources used by Wang et al. [WPL19] to Volk et al. [VSS14] and Bradley et al. [BLLD16] we observe that this article is only using the Web of Science database to gather literature. This stands in a strong contrast to suggestions from Kitchenham et al. [Kit04] and Petersen et al. [PVK15]. Petersen et al. [PVK15] suggest that a good sample of the population of articles for a targeted topic is better to represent the population. Therefore, one of their suggestions is to include different publication communities in the area of interest which are unlikely to cite each other on a regular basis.

Our analysis of related work and applied review methods is based on a sample of literature reviews provided by domain experts. None of the articles reference a well established method or guidelines for conducting literature reviews, shown in Table 2.1. It is also noteworthy that all reviews apply a keyword based database search strategy. Hence, keyword based database searches may be a common strategy in this research field in order to gather literature. We find major differences in the documentation of search strategies when we compare the guidelines from Petersen et al. [PVK15] to the reviews in our sample. In order to enable reproducibility of a study it is common practice in systematic mappings to provide the used keyword combinations including **AND** and **OR** operators. Some systematic mapping studies do also provide the search queries for each online library or literature database to enable reproducibility of the study [PVK15].

None article in our sample provides the used search query per database, but both Bradley et al. [BLLD16] and Wang et al. [WPL19] provide the keyword combination used to find relevant articles. The study by Volk et al. [VSS14] on the other hand specifies relevant keywords but it lacks a definition how the keywords are combined with the aforementioned boolean operators.

Other common search strategies identified by Petersen et al. [PVK15], such as manual search and snowball sampling based of citation relationships, are not used by any review in our sample. Moreover, we see that there are no commonly used or well established online libraries or publication databases for searching literature regarding Building Information Models. Only Bradley et al. [BLLD16] give an overview of the distribution of found literature per database or online library. Therefore, we cannot derive the relevance of a particular database or online library as two of the reviews do not provide the amount of relevant publications per database or online library.

2.3 Data Analytics Process Models

As described before the result of the mapping process is a set of classified studies identified as relevant by following a well established and thoroughly documented process. This qualifies the mapping process for automation in order to continuously deliver a state-of-the-art summary of relevant literature in the tunnelling domain. Therefore, we apply a design science research method to create a software artifact and evaluate the results of the artifact with the results of the conducted mapping study. Then we refine the artifact based on the evaluation outcome and repeat this process. Similar to the mapping study a well established process model for the design, implementation, and evaluation of the artifact is required to enable reproducibility and verifiability. In this section we present well established process models for data analytic artifacts in order to automate the mapping process.

Data analytics process models describe formal procedures, documents and structured tasks in order to enable reproducibility and verifiability of data mining artifacts in the domain of data science. Our selection of process models is based on the presentation given by the domain expert Rauber [Rau19] where he encouraged the application of a well established process model in the field of data mining. The following process models are generic references and have to be adopted to a specific business environment.

2.3.1 Knowledge Discovery in Database

In 1996, Fayyad et al. [FPS96] proposed a process model they named *KDD Process*, illustrated in Figure 2.1. The process model is based on nine generic tasks and a feedback loop from the evaluation task to the steps before. In their study they describe the steps of the process as:

Domain Understanding Understanding the application domain, the goals & problems

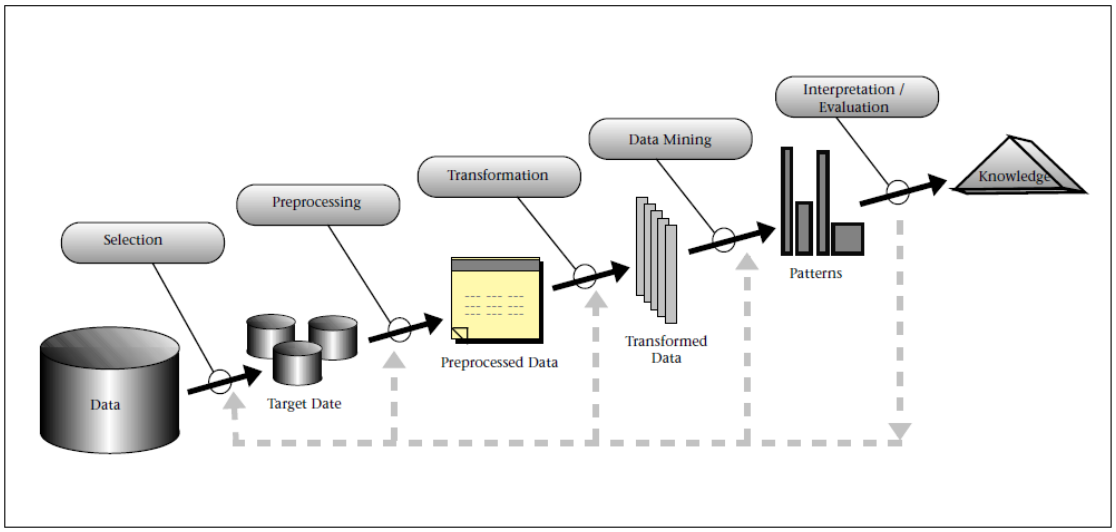


Figure 2.1: Knowledge Discovery in Database (KDD) Process by Fayyad et al. [FPS96]

and identifying required domain specific knowledge in the area where the phenomena resides.

Selection Create a target data set by selecting a subset of interest which is used for further discovery

Preprocessing Cleaning the data from noise and define how to handle missing values

Data reduction & projection Find useful features to present, reduce data dimensionality or transform the cleaned data in order to reduce the effective number of variables

Exploratory data analysis Discover and explore the data, create hypothesis and choose appropriate data mining methods to recognize patterns

Data Mining Searching for patterns by applying clustering, regression models or rule based classifications

Interpretation & Evaluation Interpretation and visualization of the surfaced data patterns which may feed into any of the previous steps

Acting Taking measures or develop new strategies based on the discovered knowledge or document the results

2.3.2 Cross Industry Standard Process for Data Mining

Four years later the article by Chapman et al. [CCK+00] describes a reference process model for data mining named Cross Industry Standard Process for Data Mining (CRISP-DM) which became a cross-industry standard process for data mining. At first the

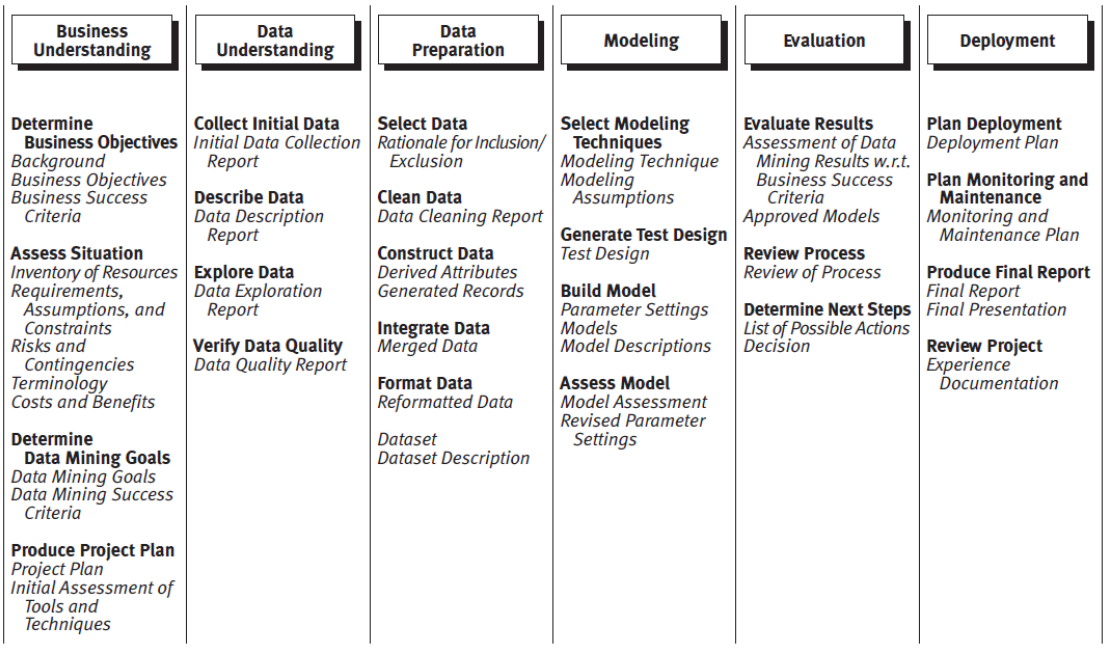


Figure 2.2: tasks (**bold**) and outputs (*italic*) of the CRISP-DM reference model by Chapman et al. [CCK+00]

CRISP-DM process tasks seem very similar compared to the presented KDD model, see Figure 2.2.

When we compare CRISP-DM to the KDD process model it becomes clear that Chapman et al. [CCK+00] focus on the formal definition of the data mining process. Documents, tasks and their outputs are described in detail in order to enable reproducibility, validation and evaluation of the processes. Another major difference between those models is that CRISP-DM specifies a deployment task which may be a specific application of the *Acting* task described by Fayyad et al. [FPS96].

In detail the reference process model consists of the following tasks:

1. Business understanding, the first process task has multiple outputs which gather business objectives, requirements, constraints and success criteria used at a later stage for model evaluation.

2. Multiple reports describe the data collection process, the data quality and the data itself as part of the data understanding process task.

3. The first tasks of the data preparation step is the data selection. Then the data is cleaned and formatted resulting in a dataset as well as a dataset description document.

4. The task of model creation starts by selecting a modeling technique, generate tests, build the model and assess the model by iterative assessment and adaption of the models parameters.

5. The results are then evaluated against the business objectives and success criteria defined at the beginning of the process. Moreover, if multiple data mining models are available then the different results are compared with each other in order to find the best model.

6. If the model has been approved, a plan for deployment of the data mining model as well as monitoring and maintenance procedures will be defined. The process ends with a final report and reflects on the process by conducting a review project.

CHAPTER 3

Methodological approach

In this chapter we look at the used research methods design science and systematic mapping study in detail. First, we outline the grounding design science theory and its application in the thesis. Followed by the definition of the mapping process and activities starting from the search strategy, data selection and filtering of papers, over to the classification and finally mapping the papers. Moreover, we describe how we adopt the previously introduced data analysis process reference model CRISP-DM to design and evaluate the design science artifact.

3.1 Design Science

Based on the guidelines by Hevner et al. [HRM+04] we apply a design science research method which is shown in Figure 3.1. The IS Research Framework is separated into the domains environment, IS research and knowledge base.

The environment or problem domain where the phenomena resides consists of the technological requirements, the processes or structure of organizations and the roles or characteristics of people experiencing the phenomena. This environment frames the problem where our phenomena of interest resides. The thesis is placed in the interdisciplinary field between IS and subsurface engineering. Therefore, our study addresses the research gap, that there is no study summarizing the research works in the field of Tunnel Information Modelling (TIM). People experiencing this gap are research scientists, experts in the field of tunnel constructions, geologists and engineers but also students in the AECOO domain and subsurface engineering. Hence, organizations experiencing this lack of research are construction companies in the field of tunnel construction and academic institutions like universities or research centers.

The technological requirements are defined by the retrieval, persistence and visualization of information required to enable knowledge generation and data insights. Searching

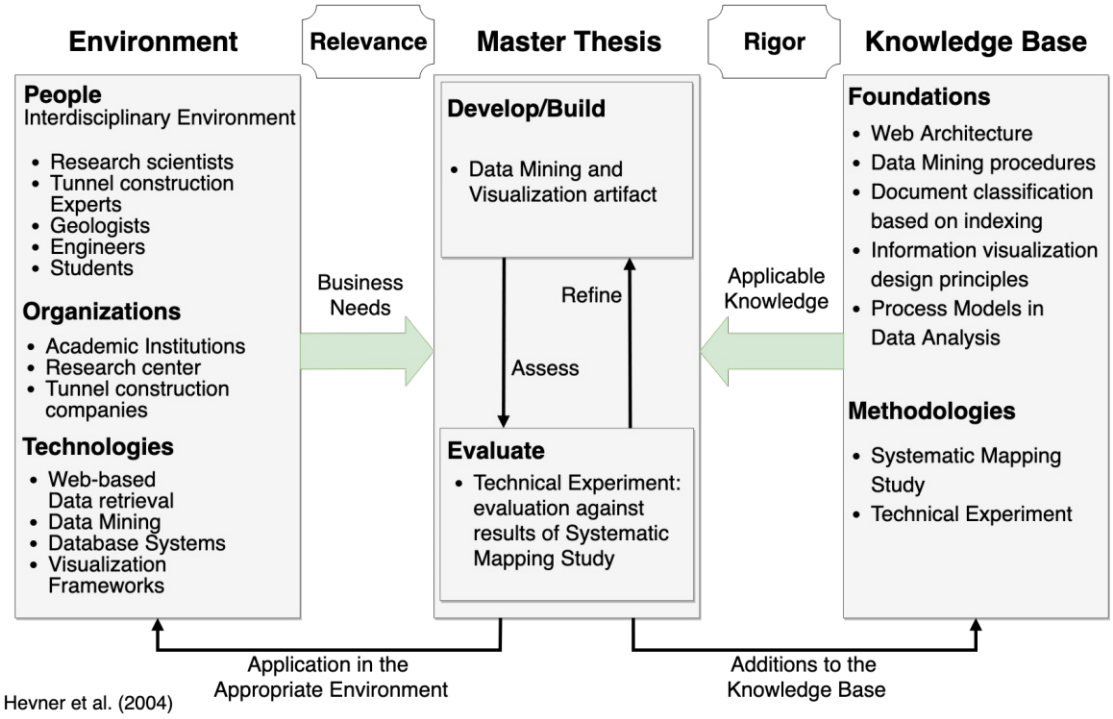


Figure 3.1: Adopted Information Systems Research Framework [HRM+04]

literature and meta-data collection makes heavy use of web based Application Programming Interface (API) provided by abstract citation databases, Digital Object Identifier (DOI) Registration Agencies or online libraries. Moreover, we require a database system in order to persist, clean and finally integrate the collected data in a structured way. With the data set we are able to build, test and assess data mining models in order to classify the previously gathered literature meta-data.

The knowledge base consists of foundations, methodologies and the communication of the IS research. Foundations are concepts, models, processes, instantiations and methods which we use to develop the artifact. Methodologies are guidelines, validation criteria or formal definitions which we apply to evaluate the artifact. In order to design and implement the artifact we require knowledge about web application architecture, data mining models and procedures, foundations in datawarehousing, and information visualization.

In the center of Figure 3.1 is the IS Research which in this case is the master thesis. The thesis consists of the development of the software artifact and the iterative evaluation against the results of a conducted systematic mapping study. This evaluation is based on the evaluation process task defined in the CRISP-DM process model by Chapman et al. [CCK+00] which focuses on data mining artifacts.

Hevner [Hev07] describes the interaction between those domains as design science research

cycles. The relevance cycle between environment and the master thesis consists of the requirement definition or business needs for the IS research. Moreover, by applying the artifact in the appropriate environment we induce a change of the environment. The rigor cycle resides between the knowledge base and the master thesis and defines grounding theories, models or methods which are applicable knowledge and enables us the realization of the IS research. Communicating the results of the research adds knowledge to the existing knowledge base and closes the rigor cycle. The iterative process of assessment, evaluation and refinement between artifact and evaluation criteria defines the design cycle [HRM+04], [Hev07].

3.2 Systematic Mapping Study

The process we use in this thesis to conduct the previously introduced SMS is based on the guidelines by Petersen et al. [PVK15] and the process used by Wolny et al. [WMC+20], see Figure 3.2. According to Petersen et al. [PVK15] we first define the Research Questions (RQ) . Then, we create, assess and refine literature searches based on keywords and keyword combinations. Those keyword combinations are reviewed by domain experts and adapted during the iterative process of search result assessment and refinement. Based on the keyword combinations a literature search is conducted which delivers all publications related to the previously defined research questions. All publications are filtered and screened based on inclusion and exclusion criteria. The output of this task are relevant publications in regards to the previously defined keyword combinations in title or abstracts, the subject area, the document type and the language. Then all abstracts of the publications are used to identify relevant literature and classify the studies. Similarly to the adoptions of the activities by Wolny et al. [WMC+20] we apply multiple classifications based on abstracts. Additionally to the research type facets recommended by Petersen et al. [PVK15] we also classify relevant literature based on the applicability of the TIM research during the three life cycle phases of a tunnel and the excavation type if it is defined in the abstract. As output, we get the classified abstracts of relevant publications which is used by the mapping activity. The mapping task results in a systemic map of the research field which enables us to extract main findings related to our research questions.

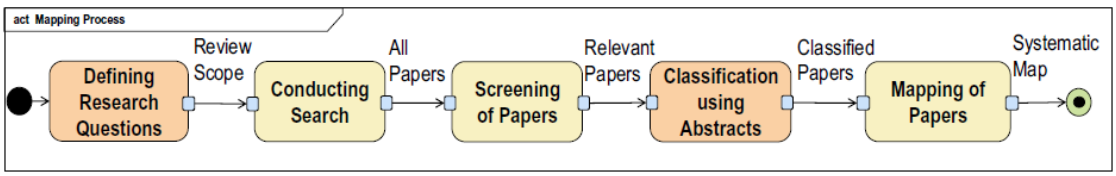


Figure 3.2: Activity diagram of the mapping process by Wolny et al. [WMC+20]

3.2.1 Research questions

Our first activity in the process of a mapping study is to define the research questions derived from the problem statement. We already specified the objective of this mapping study to provide an overview by classifying relevant literature in the area of tunnel information models. This leads to the following research questions:

- **RQ1:** *What are the bibliometric key facts of TIM publications?*
  This research questions aims to identify (i) the main publication types (e.g. journal article, conference proceeding), (ii) the number of studies which have been published between 2002 and 2019 and (iii) the main publication forums (conferences, journals) where the studies have been published.
- **RQ2:** *Which excavation types and life cycle stages of a tunnel facility are relevant?*
  The intention of this research question is to categorize publications based on their primary application in one or all tunnel life cycle phases: plan, construct & operate. Moreover, we are interested in research trends of information models primarily targeting specific tunnel excavation methods.
- **RQ3:** *Which relevant research type facets do the identified publications address?*
  Additionally to the tunnelling specific classification we intent to categorize the literature based on a well-established scheme [WMMR06] & [PVK15]. Hence, we use the classification scheme introduced by Wieringa et al. [WMMR06] and used in the guidelines by Petersen et al. [PVK15] to classify literature in regards to formal criteria.
- **RQ4:** *What are relevant search terms?*
  This research question intends to identify major keywords which characterizes literature in the domain of Tunnel Information Modelling (TIM)

This information is then used to provide an overview and show publication trends in the research area of tunnel information models.

3.2.2 Conducting search

Based on the research questions, we develop a search strategy. In our first task we identified relevant keywords based on our research questions and encountered the issue that the term *"Tunnel Information Modelling (TIM)"* is not well-established and rarely used in the field of research. Therefore, we start with a descriptive set of keywords and iteratively assess and refine the sets of keyword in cooperation with domain experts. Resulting in two major keyword sets (see Fig. 3.3), including their synonyms and German translations:

- Set 1: Scoping the search for the usage of BIM in the tunnelling domain, i.e. "BIM" **AND** "Tunnel"

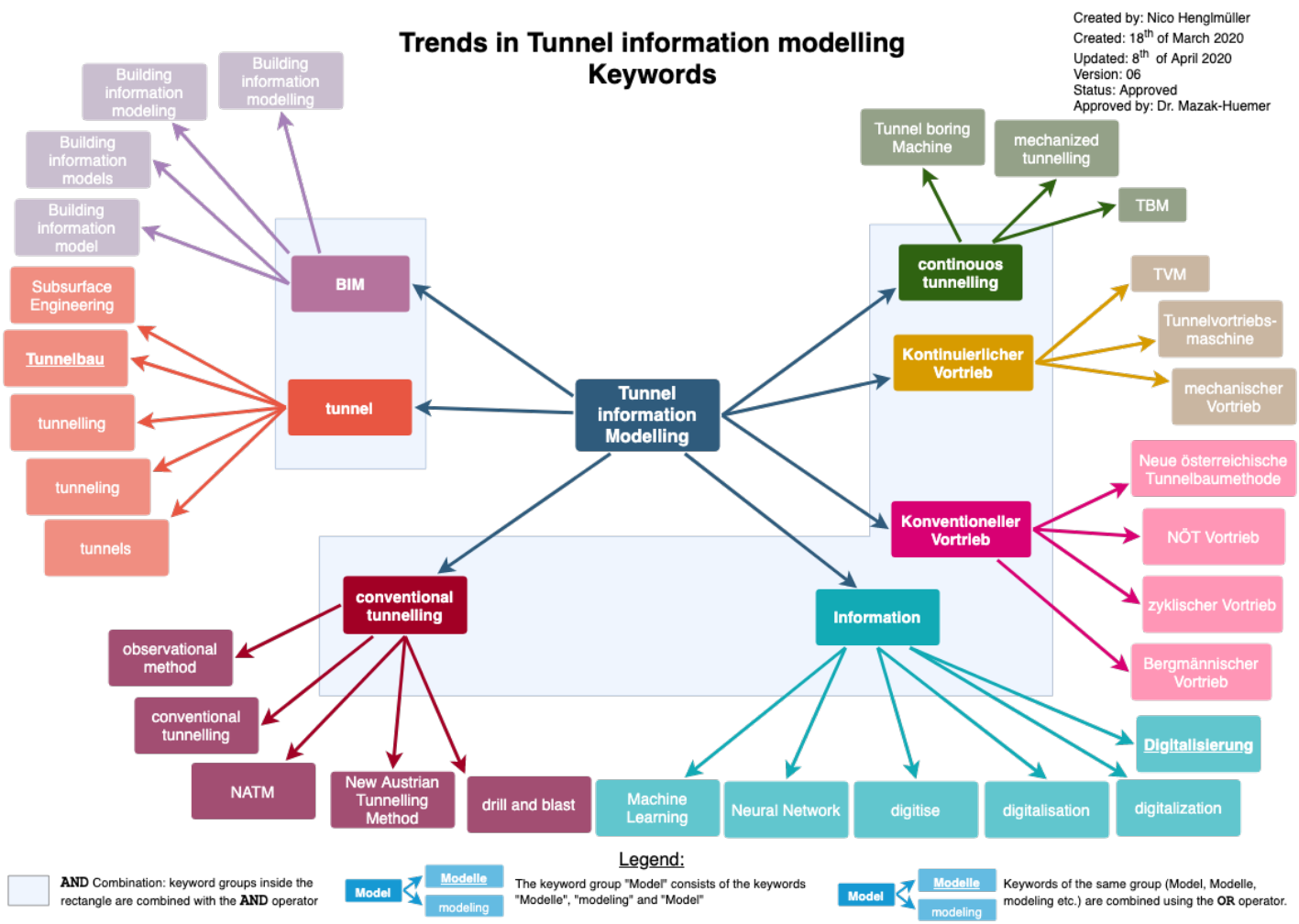


Figure 3.3: Keyword Groups and combinations

Table 3.1: Number of search results per database

Database	Search result
Scopus	3163
SpringerLink	6269
Wiley	84

- Set 2: Search terms directly related to excavation methods and digitization, e.g. "NATM" **AND** "Neural Network"

With the resulting combinations of keyword sets and boolean operators we performed searches on the databases of Scopus, SpringerLink and Wiley Online Library. The abstract citation databases have been selected based on major publication forums (e.g journals, conferences) recommended by domain experts. The search strings for each database we used are provided in Table 3.2.

This study was conducted during 2020, therefore we limited the time period starting from 2002 until 2019. The number of results per database is shown in Table 3.1.

3.2.3 Screening of Papers

Based on the search result set we included and excluded publications using its title and abstracts which is sufficient according to the guidelines of Petersen et al. [PFMM08]. When it was unclear whether to include or exclude a study, the author conducted a full-text reading of the publication. We excluded the study, if the full-text of the study was not available. If we were still inconclusive we discussed these publications with a domain expert and decided whether to include or exclude a study. It is worth noting, that the publications have been reviewed by a single author (i.e. the author of this thesis) which poses a threat to the reliability of the mapping study. The validity evaluation of the mapping process is discussed in Chapter 4.1.6.

The screening process stages and the corresponding number of results are shown in Figure 3.4. By searching the citation databases we received more than 9500 results which we were able to reproduce in multiple search attempts. Therefore, we decided to apply inclusion criteria in order to obtain more precise results of publications in the field of tunnel information models.

We applied the following inclusion criteria:

- *Studies are in the subject area of Computer Science or Engineering.*

  In order to reduce noise we focus our search on the subject areas of the interdisciplinary field of information models used in the life cycle of tunnel facilities. The vast majority of studies in the original result set is related to other disciplines, e.g.

Table 3.2: Searches in databases

Database	Search
Scopus	((TITLE-ABS-KEY("continouos tunnelling" OR "tunnel boring machine" OR "TBM" OR "mechanized tunnelling" OR "kontinuierlicher vortrieb" OR "tunnelvortriebsmaschine" OR "TVM" OR "mechanischer Vortrieb" OR "conventenional tunnelling" OR "observational method" OR "conventional tunnelling" OR "NATM" OR "new austrian tunnelling method" OR "drill and blast" OR "konventioneller vortrieb" OR "neue österreichische tunnelbaumethode" OR "NÖT Vortrieb" OR "zyklischer Vortrieb" OR "bergmännischer Vortrieb") AND TITLE-ABS-KEY("information" OR "digitalization" OR "digitalisation" OR "digitise" OR "machine learning" OR "neural network" OR "digitalisierung")) OR (TITLE-ABS-KEY("tunnel" OR "subsurface engineering" OR "tunnelling" OR "tunnelling" OR "tunnels" OR "tunnelbau") AND TITLE-ABS-KEY ("BIM" OR "Building information modelling" OR "Building information modeling" OR "Building information models" OR "Building information model")))
SpringerLink	(("tunnel" OR "subsurface engineering" OR "tunnelling" OR "tunnelling" OR "tunnels" OR "tunnelbau") AND ("BIM" OR "Building information modelling" OR "Building information modeling" OR "Building information models" OR "Building information model")) OR (("continouos tunnelling" OR "tunnel boring machine" OR "TBM" OR "mechanized tunnelling" OR "mechanischer Vortrieb" OR "kontinuierlicher vortrieb" OR "tunnelvortriebsmaschine" OR "TVM" OR "conventenional tunnelling" OR "observational method" OR "conventional tunnelling" OR "NATM" OR "new austrian tunnelling method" OR "drill and blast" OR "konventioneller vortrieb" OR "neue österreichische tunnelbaumethode" OR "NÖT Vortrieb" OR "zyklischer Vortrieb" OR "bergmännischer Vortrieb") AND ("information" OR "digitalization" OR "digitalisation" OR "digitise" OR "digitalisierung" OR "neural network" OR "machine learning"))
Wiley	(("continouos tunnelling" OR "tunnel boring machine" OR "TBM" OR "mechanized tunnelling" OR "kontinuierlicher vortrieb" OR "tunnelvortriebsmaschine" OR "TVM" OR "mechanischer Vortrieb" OR "conventenional tunnelling" OR "observational method" OR "conventional tunnelling" OR "NATM" OR "new austrian tunnelling method" OR "drill and blast" OR "konventioneller vortrieb" OR "neue österreichische tunnelbaumethode" OR "NÖT Vortrieb" OR "zyklischer Vortrieb" OR "bergmännischer Vortrieb") AND ("information" OR "digitalization" OR "digitalisation" OR "digitise" OR "machine learning" OR "neural network" OR "digitalisierung")) OR (("tunnel" OR "subsurface engineering" OR "tunnelling" OR "tunnelling" OR "tunnels" OR "tunnelbau") AND ("BIM" OR "Building information modelling" OR "Building information modeling" OR "Building information models" OR "Building information model"))

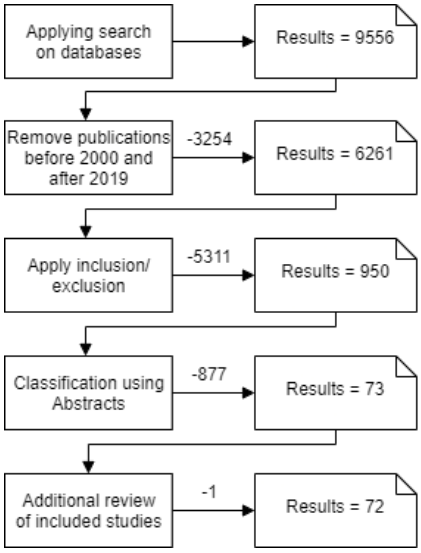


Figure 3.4: Study selection process: number of included articles [PVK15]

medicine, biology and naval sciences. The reason for this noise are the abbreviations *Tunnel Boring Machine (TBM)* and *TVM* we used in our search keywords to identify literature in the field of mechanized tunnelling.

- *Studies with an online published date in the period from 2002 to 2019.*

  As this survey was conducted in mid/2020 we decided to include the year 2019, but exclude the ongoing year 2020 as mitigation strategy to enable reproducibility of our results. The reason is, that there might still be relevant literature published until the end of 2020. Hence, studies published in the second half of 2020 would not be part of the survey leading to different result sets depending on the search date. In our pilot searches we found first publications in the field of computer-aided prediction models published in 2002 [EH02]. Thus, we use this year as starting point for our mapping study.

- *Studies presenting research which makes use of information models during at least one of the three life cycle phases of an infrastructure tunnel facility.*

  This inclusion criteria is especially important for studies focusing on very specific problem areas in the tunnelling domain. Publications about predicting the geology in front of the TBM are not included, because they do not primarily make use of BIM and its principles. Moreover, this study focuses on the tunnel information models for infrastructure tunnels, e.g. rail or road tunnels. Hence, special tunnel facilities such as utility tunnels are not part of the mapping study.

Moreover, we introduce exclusion criteria to define a minimum set of quality constraints which publication must fulfill in order to be part of the result set. Therefore, we applied

the following exclusion criteria on results:

- *Duplicates*
- *Books*
- *Theses*
- *Papers:*
  - without available English or German abstracts
  - without any relation to BIM or TIM
  - with identical abstracts or titles
    Some publications are part of conference proceedings and were identically published in journals. We decided to keep one of the publications and removed the other one from our result set.

Our process to derive keywords from the research questions in order to identify relevant studies is based on Petersen et al. [PVK15]. Therefore, we developed the keyword groups and combinations in an iterative refinement process in close cooperation with domain experts. Together with the co-advisor we agreed on the resulting keyword group combinations to identify publications as relevant for the mapping study. The keyword groups and combinations are shown in Figure 3.3. In order to evaluate the study identification process the co-advisor and domain experts provided a test-set of nine papers shown in Table 3.3. The table shows several papers which do not match any approved keyword group combinations. A detailed analysis of studies shows, that those papers do not fit the definition of Tunnel Information Modelling (TIM) according to the agreed keyword group combinations. Hence, papers which are not identified by the keyword group combinations are not relevant for this study as the keywords were derived from the research questions and approved by domain experts and the co-advisor during the activity *Conducting search*. The remaining papers of the test-set have been identified as relevant by the activity *Screening of Papers*. Therefore, the result set contains all relevant papers in regards to the research questions. A full list of the journal articles and conference papers which we identified as relevant can be found in Chapter 5.

3.2.4 Classification using abstracts

We adapt the guidelines of Petersen et al. [PVK15] to classify the publications based on the abstracts. Similar to Wolny et al. [WMC+20] we decided to classify the literature already in this activity and therefore deviate from the original process proposed by Petersen et al. We classify publications based on the well established research type facet classification scheme proposed by Wieringa et al. [WMMR06], see Table 3.4. This classification scheme enables us to get a deeper insight about the research context and the publication behavior in this scientific domain.

Additionally to the classification in research type facets, we examine each abstract in regards to the life cycle of a tunnel facility and the primary used excavation method, if it

Table 3.3: Test-set provided by domain experts and matched keyword groups

Publication	Matched Keyword Group Combination	Identified as relevant?
[DiEL16]	`BIM AND Tunnel`	Yes
[DFSS14]	None	No
[BLLD16]	None	No
[VB17]	None	No
[BKD+15]	`BIM AND Tunnel`	Yes
[FFG+19]	`BIM AND Tunnel`	Yes
[TK19]	None	No
[KRNS17]	`BIM AND Tunnel`	Yes
[GWZ18]	`BIM AND Tunnel`	Yes

Table 3.4: Research Type Facet [WMMR06]

Category	Description
Validation research	The author investigates the properties of a not yet implemented solution proposal in practice. Therefore, the author uses sound research methods.
Evaluation research	These papers apply sound scientific methods in order to investigate problems in practice or implementations of a technique in practice.
Proposal of solution	The author proposes a novel technique or solution and argues for its relevance without conducting a rigor validation. Often proof-of-concepts are used to show case the solution or argument.
Philosophical papers	These papers present a new conceptual framework or a new way of looking at things.
Opinion papers	The author describes personal opinions about what is good or bad or how we should do something.
Personal experience papers	In these papers, the author describes his or her personal experience about a project by listing the lessons learned. The main focus of this type is about the *what* and not about the *why*.

is defined. This results in a more fine grained categorization compared to the guidelines by Petersen et al. [PFMM08] and the proposed mapping process. Those categorizations are:

Excavation methods: Studies in the field of tunnelling may be classified based on the excavation type. Therefore, we use the following classifications:

- *Mechanized tunnelling:* These studies focus on information models and their application in any life cycle phase of bored tunnels.
- *Conventional tunnelling:* These papers describe the application of information models in regards to the New Austrian tunnelling Method (NATM) or drill and blast excavation methods.

- *Unspecified:* The authors of these publications do not specify a specific excavation method.

Tunnel life cycle phases: We examine the primary application of the research based on the tunnel life cycle stages defined by Stascheit et al. [SNM+18]. Therefore, we apply the following categorization scheme:

- *Planning & Design:* These studies focus on information models used during the planning and design process of a tunnel, e.g. optimization of the tunnel alignment or applied tunnelling process.
- *Construction:* These papers present research about information models used primarily during the excavation and construction phase of a tunnel, e.g. unify multiple heterogeneous data sources in order to streamline the construction process.
- *Operation:* These studies focus on maintaining an operative tunnel by creating information models to support maintenance procedures, e.g. creating as-built information models or detecting cracks or leakages based on three-dimensional laser scanning.
- *All phases:* The authors do not specify a specific life cycle phase, but present research about information models applied in all three stages of the tunnel life cycle.

We classified the literature based on the research type facets, tunnel life cycle stages, and excavation methods. The result of our classification serves as input for the last activity *mapping of papers* and is described in section 4.1.

3.3 Design Science Artifact

This chapter elaborates on the design process of the software artifact following the proposed data analytics process model CRISP-DM presented in Chapter 2.3.2. First, we derive the research question from the problem domain and environment in order to frame the objectives of the artifact. Then, we describe DM requirements regarding data structure and data quality as well as technical requirements used to realize the artifact. We continue with a detailed elaboration on the data retrieval process and a data structure description. Based on the data structure description we evaluate, if the data set meets the DM requirements. Then an exploratory data analysis and a data quality analysis provide a detailed discussion of the data set. This data exploration process is applied for each publication database in order to develop an information retrieval architecture. Based on the initial data set descriptions we elaborate on data selection criteria, study identification and data cleaning measures.

3.3.1 Research question

In Chapter 1.2 we identified the problem, that stakeholders in the research domain TIM require a state-of-the-art set of relevant scientific publications. Hence, we derived the following research question to engage the problem from an IS perspective:

- **RQ5:** *How well does a software artifact classify tunnel information model studies compared to a systematic mapping study?*
  This research questions aims to evaluate result sets of studies identified as relevant for the research domain TIM based on the design science research methodology. Hence, we propose a data science research methodology in order to automatize the process of study classification. At first we conduct an SMS in order to identify and classify relevant studies as well as gather insight about the problem domain. We apply the develop cycle to design, evaluate and refine the data science artifact based on the mapping studies result set. The main objective of the data science artifact is to (i) identify studies which are relevant in the field of Tunnel Information Modelling and (ii) classify those studies based on the excavation types and tunnel life cycle phases.

We describe the mapping process and derived activities in detail in Chapter 3.2. Following the research question the main objective is to assess the resulting studies of a data mining artifact based the mapping studies result set. Therefore, we evaluate the success of a proposed model based on well-established statistic metrics. We describe the modeling technique selection, test design and model building in Chapter 3.3.5.

3.3.2 Data Mining Requirements

We propose technical requirements as well as data mining requirements in order to enable a reproducible data analysis process. Data specific requirements aim to describe required attributes and quality measurements to answer the research question. Therefore, the data set specific requirements apply the inclusion and exclusion criteria defined in Chapter 3.2.3.

The identification of relevant scientific publications requires at least one attribute which thematically positions the publication and describes the focus of the research. Therefore, potential candidate attributes for study identification and classification are (i) the studies abstract, (ii) the study title and (iii) author assigned keywords. Author assigned keywords are five to twelve descriptive keywords assigned by the study authors to specify the major topics of their study. A classification based on author keywords seems natural. Nevertheless, we prefer the abstracts over author keywords in order to identify and classify publications. The major reasons are that, (i) author keywords may contain abbreviations (e.g. TBM) in a context unrelated to subsurface engineering which leads to a relatively high number of false positives and (ii) the data quality and availability of the author keywords attribute is dependent on the citation database. We reject the publication title

Table 3.5: Required technical resources to deliver the design science artifact

Requirement	Resource	Used instantiations
Artifact implementation	Programming language	Python 3.7
Data Analysis and Modeling	Data Science Frameworks	scikit-learn 0.24.1 [PVG+11] pandas 1.2.0 [RMj+21]
Natural Language Processing	NLP Framework	spaCy 2.1 [HMVLB20] rank_bm25 0.2.1 [Bro20]
Deployment	Virtual Machines, Build Tools	CentOS8, PyPi and RPM
Access to Citation databases	Login for unrestricted access	TU Wien VPN

as a potential candidate attribute due to the complexity of a title based identification and classification. A major reason for our decision is that the study titles are often too short and specific compared to abstracts. Consequently, the identification and classification based on the title of scientific publications is out of scope of this study.

In some cases citation databases do not provide abstracts or author keywords as part of the result set. For those data sets we require a Digital Object Identifier to be included in the attribute list. A DOI is used to identify digital assets in digital networks and distributed systems. Therefore, an independent DOI registration agency assigns an identifier to a digital asset and persists metadata about the asset. As a consequence a unique DOI is assigned to each publication. This unique DOI is used by libraries, publishers or citation databases to identify a scientific study. We use the identifier to query APIs of DOI registration agencies in order to receive metadata including the abstract of the study. Additionally to the DOI we require the study title for duplicate identification, as the DOI is not provided for each study. Therefore, we exclude tuples in case the attributes, abstract and title, have no valid value set.

Our analysis of technical requirements is based on (i) the data analysis process described in Chapter 2.3.2, (ii) the search strategy of the mapping study outlined in Chapter 3.2.2 and (iii) data preservation strategies for long-term preservation and data sharing. Hence, a major requirement is to extract the data from the publication databases SpringerLink, Scopus and Wiley Online library in an automatized way. Moreover, we persist received data in a structured way which is optimized for data analysis. Based on the cleaned and tokenized data set a data science model identifies relevant studies and classifies them. In order to deploy the resulting artifact we require a build environment to implement packages for deployment. We use the built packages to deploy the software artifacts on a standardized environment, the deployment target. Table 3.5 gives an overview of the identified technical requirement and resources used to implement the design science artifact. We choose Python for the artifact implementation as it is a well-established programming language in the domain of data science and analytics. In regards to data preservation we provide the virtual environments as part of the thesis in order to ensure the research results and processes are reproducible. Moreover, the build artifacts are distributed via PyPi which requires the plain source code in order to build the package.

Therefore, the deployment process consist of the installation of Python and the software artifact via PyPi.

As deployment target we choose the open source operating system CentOS due to its popularity in the scientific community. Based on the decisions about the programming language and the operating system of the deployment target, the RPM build tools and PyPi are set. In order to query citation and abstract databases, publisher databases and DOI registry APIs an unrestricted database access is required. Without an unrestricted access a query is sent to a citation database and either is rejected or the response contains a set of open access studies. Therefore, we require an unrestricted access to the online citation databases and use the Virtual Private Network (VPN) of the Technische Universität (TU) Wien for this purpose. In regards to the APIs of SpringerNature [Spr21e] the VPN is not sufficient as SpringerNature requires an API token in order to query the API. Therefore, we have requested an API token at the SpringerNature API Portal which allows us to query the SpringerNature literature database [Spr21e].

3.3.3 Data understanding

This section elaborates on the application of the CRISP-DM sub-process *Data Understanding* and describes the processes for each data source. First, we document how the initial data sets are retrieved from abstract and citation databases as well as Digital Object Identifier registration agencies. Additionally, we discuss issues to access publication metadata from available APIs or other interfaces provided by the publication databases. The second task of the process is to describe the initial data sets in regards to its structure, quantity and relevant features. Therefore, relevant attributes are selected with the objective to answer the research question of the design science artifact. We also assess and evaluate if the data sets satisfy the data mining requirements defined in Chapter 3.3.2. In case a data set does not satisfy the requirements, we describe measures to fulfill the requirements or reject a subset. In the third step we explore the data for key attributes as well as relationships between attributes, discuss findings and hypothesis accompanied by plots. The discussion of the exploratory data analysis deviates from the process model as used data sets are cleaned from duplicates beforehand. Finally, the data quality of each received data set is outlined in regards to missing values, errors and duplicates.

Scopus

Scopus [Els20d] is an abstract and citation database operated by Elsevier and comprises numerous journals, conference proceedings and books from different publishers. Hence, the vast majority of publications can be found by querying Scopus as it is not limited to scientific work published by Elsevier associated journals or conferences. As a consequence Scopus has a large repository of publications and metadata such as authors, affiliations, references, et cetera. Scopus provides a web-based graphical user interface for manual search tasks which are commonly used during literature reviews. Additionally, Elsevier [Els20b] provides several web-based APIs to search Scopus repositories for affiliations,

Table 3.6: Scopus Query with additional filter constraints

Scopus Query
((TITLE-ABS-KEY("continouos tunnelling" OR "tunnel boring machine" OR "TBM" OR "mechanized tunnelling" OR "kontinuierlicher vortrieb" OR "tunnelvortriebsmaschine" OR "TVM" OR "mechanischer Vortrieb" OR "conventenional tunnelling" OR "observational method" OR "conventional tunnelling" OR "NATM" OR "new austrian tunnelling method" OR "drill and blast" OR "konventioneller vortrieb" OR "neue österreichische tunnelbaumethode" OR "NÖT Vortrieb" OR "zyklischer Vortrieb" OR "bergmännischer Vortrieb") AND TITLE-ABS-KEY("information" OR "digitalization" OR "digitalisation" OR "digitise" OR "machine learning" OR "neural network" OR "digitalisierung")) OR (TITLE-ABS-KEY("tunnel" OR "subsurface engineering" OR "tunnelling" OR "tunneling" OR "tunnels" OR "tunnelbau") AND TITLE-ABS-KEY("BIM" OR "Building information modelling" OR "Building information modeling" OR "Building information models" OR "Building information model"))) AND (SUBJAREA(engi) OR SUBJAREA(comp)) AND (PUBYEAR > 1999 AND PUBYEAR < 2020) AND (LANGUAGE("English") OR LANGUAGE("German")) AND (PUBSTAGE("final")) AND (DOCTYPE ("ar") OR DOCTYPE("cp") OR DOCTYPE("ch"))

authors, citations and abstracts of publications. Therefore, Rose and Kitchin [RK19] proposed a software artifact in order to access and communicate the Scopus APIs in a transparent and reproducible way. The primary use cases of this artifact in our implementation are:

1. authenticate against the Scopus API using an API token received from [Els20b],
2. query the Scopus API and
3. receive the result set.

Moreover, during upcoming tasks we are required to identify the used language of studies in order to explore relationships and use appropriate models. Therefore, we use a software artifact developed by Danilák [Dan20] which itself is a Python port of the software artifact by Nakatani [Nak10]. The software artifacts are based on the research work by Cavnar and Trenkle [CT94] describing an n-gram method for text classification.

The Scopus query presented in Table 3.2 serves as starting point for the data retrieval task. Moreover, we apply the inclusion and exclusion criteria specified in Chapter 3.2.3 by adding filters to the original query. The result is shown in Table 3.6. In more detail, we added constraints for (i) a range of publication years, (ii) languages, (iii) venue types such as journal and conference proceeding articles, (iv) publication stages, and (v) the subject areas of interest.

We query the API by using the software artifact from Rose and Kitchin [RK19] and receive an initial data set with a total number of 806 tuples. The API documentation [Els20c] and the data set allow us to derive entities and their relations. Figure 3.5 shows the resulting ER-Diagram of a selected subset of attributes. We select this subset of

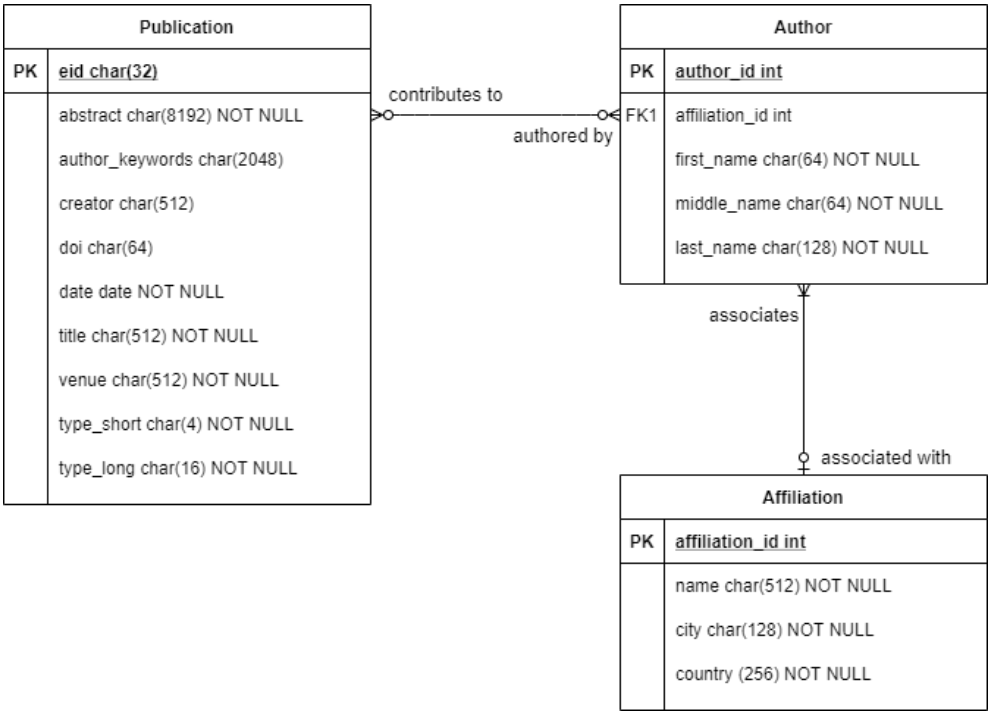


Figure 3.5: ER Diagram derived from relevant Scopus API attributes [Els20c]

attributes based on the data mining requirements, defined in Chapter 3.3.2, in order to answer the research question. For each attribute we derive constraints based on the analysis of missing values. Therefore, we assign a *NOT NULL* constraint to those attributes which do not hold missing or empty values. We discuss missing and empty values in more detail during the data quality analysis of the result set later in this section.

The data structure of the received data set specifies an *Author* entity which is identified by an *author_id*. This entity holds name attributes as well as a foreign key for the associated *Affiliation*. The *Affiliation* entity holds attributes which may allow us to assign the publication to a specific institution and, therefore, country. According to the received data set an author is associated with at most one or none affiliation. An *Affiliation* on the other hand is associated with at least one or many authors. We would expect the *Affiliation* to be the bridge entity of a many to many relation between `Author` and an entity such as *Institution*. But neither the result set or the API documentation explicitly describe such a many to many relation.

The most prominent entity is the *Publication* itself which is identified by the attribute `eid` or electronic identifier. This primary key is a Scopus exclusive identifier assigned by Scopus to any electronic asset. Our analysis shows that the `eid` is unique for all electronic artifacts even if the asset is a duplicate. As a consequence the *eid* does not qualify for duplicate analysis.

The many to many relation between the entities *Publication* and *Author* has a modality of zero on both sites of the relation. At first, this seems odd as a publication naturally requires at least one author. The reason for the modality is that the result set contains two publication tuples without any assigned authors. Those tuples may be identified by their titles:

- *"Bilbao Metro Line Two Gets Support." Tunnels and Tunnelling International, no. SEP (2012): 31–34* [Bil12]
- *"Blast Design Using Measurement While Drilling Parameters." Coal International Mining and Quarry World 250, no. 2 (2002): 82–85.* [Bla02]

One of the most important attributes for our data analysis is the abstract attribute as the software artifact uses it to identify and classify the publications. As mentioned aboveScopus indexes publications collected from numerous publishers and online libraries. Therefore, 43% of all abstracts contain a copyright notice of the originating publisher. Additionally to the abstract the *Publication* entity holds attributes for the study `title` as well as *author_keywords*. The author assigned keywords are a character sequence which values are concatenated with a pipe symbol, e.g. *Conceptual design | HCCR TBM | Preliminary design*. In total 70% or 563 out of 806 publication tuples have keywords assigned. It is important to note that a publication has no language attribute. This indicates that the language of a publication has to be derived from other attributes such as the abstract.

Each publication may also have a creator attribute. This feature is derived by the attributes of the referenced `Author` and is the name of the primary author. In more detail, the attribute is a concatenation of the primary authors last name together with the initials of the first an middle names, e.g. *Yang W.W.*. Hence, the attribute is not unique and very coarse compared to the fine grained structure of the *Author* entity. Also, the creator attribute is empty for both of the two edge cases of publications without author references, see above.

The next attribute of the `Publication` entity is the standardized DOI. The official documentation of the Digital Object Identifier is published by the International DOI Foundation in [Int15]. Moreover, our analysis shows that 80% or 651 of 806 publications have a DOI assigned. Additionally to the DOI each publication has a publication date assigned, e.g. `2019-10-01`. This date is based on the online publication date of the volume or inproceeding. Therefore, the venue attribute represents the name of the journal or conference which contains the article. The type attributes, *type_short* and *type_long*, provide information about the venue type, f.e. conference proceeding or journal article. In more detail the *type_short* is a two character description of the venue type, e.g. `ar` for journal article and *cp* for conference proceeding. Consequently, the *type_long* attributes contains the full type description, e.g. *journal article*.

As we finish the description of the structure and attributes of the initial set we continue with the data set evaluation task. Therefore, we are able to assess the data set received

from the Scopus API in regards to our data mining requirements outlined in Chapter 3.3.2. The first requirement is the presence of an abstract which is satisfied as all received tuples hold a valid abstract. Moreover, we require a study title in order to analyze duplicates. This requirement is met as well as the data set has a title assigned to each publication. Hence, the received data set satisfies the defined requirements.

Our next task according to the CRISP-DM data analysis process described by Chapman et al. [CCK+00] is to explore the data set. Therefore, we want to discuss and visualize relationships of key attributes which contribute in answering the research question. Based on our data mining requirements we identified that the abstracts and author keywords are essential for the identification and classification of relevant publications. The upcoming analysis uses a data set without duplicates and therefore contains 803 instead of the initially 806 publications. We discuss the duplication identification and how we handle them later in this section.

The main objectives of the upcoming analysis is to (i) show trends by the amount of publications relating to the domain of Tunnel Information Modelling, (ii) provide data in order to develop hypothesis used to identify relevant publications, (iii) identify and adjust too broad search terms, as well as (iv) to surface relations between different keywords groups. We use the keyword groups presented in Chapter 3.2.2, Figure 3.3 to analyze their occurrences in abstracts, titles and author keywords. Therefore, we apply a common data preparation technique called stemming in order to identify word variants with the same stem [Lov68]. The stem occurrence analysis may indicate a tendency about the relevance of certain keywords or their stems as well as the amount of relevant publications in the Scopus result set. If the abstract contains the keywords or a derived word of the same stam at least once then this publication is considered exactly one match. We use a regular expression to ignore any symbols before or after the keyword or stem to prevent exact search term matches.

Figure 3.6 shows the relative number of matched publications per keyword or derived word stem grouped by abstracts, titles and author keywords. We use the relative distribution, because the number of publications with titles and abstracts (n=803) differs from the amount of publications with author keywords (n=563). When we compare the shapes of the bar chart we can observe that the abstract has the highest amount of search terms. This is to be expected as the length, in terms of word count, of abstracts is much higher compared to title and author keywords. Therefore, we argue to use the abstract for further analysis and the development of the text classification model. In regards to the topics we see that the majority of matches are related to the word stem *tunnel*. In detail 26.9% of all titles, 44.7% of all abstracts and 25.6% of all author keywords contain the word tunnel. A major reason may be that tunnel is not an exclusive keyword group, but is used by multiple subdomains in the tunnelling domain. Examples for this observations are the keyword groups *continuous* and *conventional tunnelling*. Another important indication is that 55.8% of publications in the Scopus result set do not contain the word tunnel in the abstract. As a consequence those tuples may be unrelated to the research area Tunnel Information Modelling. Additionally, we expect irrelevant matches in the

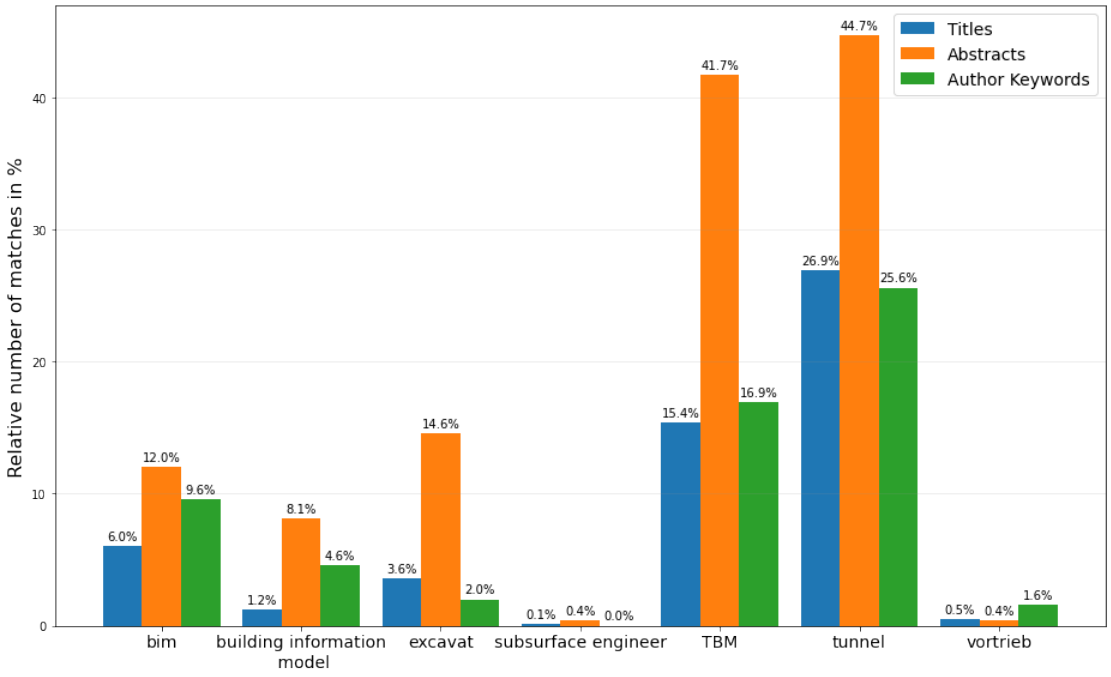


Figure 3.6: Relative number of keywords used in abstracts (n=803), titles (n=803) and author keywords (n=563)

tunnel and *excavat* groups. One major reason is, that studies may be located in different tunnelling domains such as utility tunnelling or mining. Another reason is that papers are located in the tunnelling domain, but may not be associated with TIM. An example for this are studies about prediction models of Tunnel Boring Machine disk cutterhead wear off. The TBM search term also shows the second highest amount of matches. Similar to the generic *tunnel* search term we expect a high number of studies without relation to the tunnelling domain in this set of publications. One major reason is, that *TBM* is used as an abbreviation by multiple unrelated research domains in the subject areas engineering and computer science. Prominent examples are the *Transfer Believe Model* described by Smets [Sme90] or *Time-based Maintenance* discussed by Mann et al. [MSK95]. Hence, we suggest the analysis of possible relations of matches in the *TBM* set to the search terms *information* and *BIM*. Moreover, based on the number of matches of the *tunnel* search term we may consider to exclude publications without a match. The main argument is, that the research question is specifically positioned in the research domain of tunnelling and TIM. Therefore, we require an analysis of the matches in the *tunnel* group compared to other keywords to make a decision in this matter. Especially the relation between *tunnel* and keywords from the groups *information* and *BIM* are of interest as this is the intersection of topics where TIM is positioned.

When we look at search terms with the lowest amount of relative matches we can observe, that the keyword *subsurface engineer* occurs in less 1% of the attributes. Moreover, we

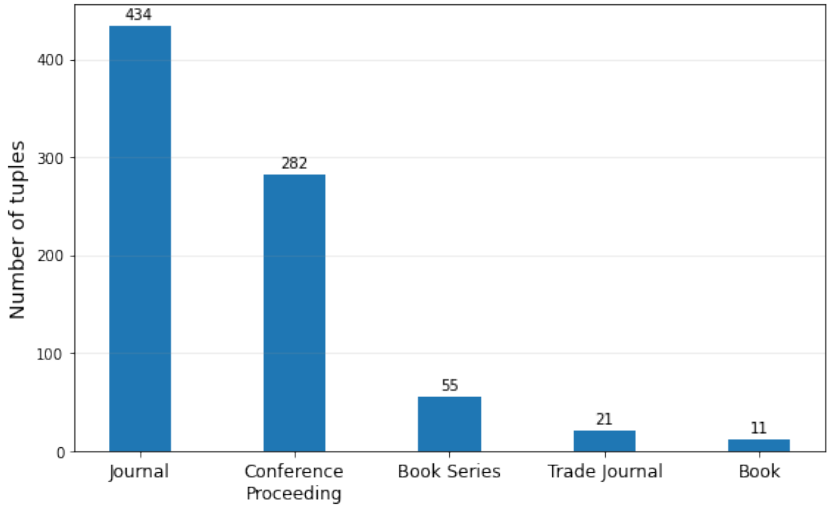


Figure 3.7: Number of tuples (n=803) per venue type in the deduplicated Scopus data set

can see a majority of English publications containing the stem *excavat* when compared to the relative amount of matches of the German equivalent noun *vortrieb*. In order to detect the language of each publication we analyze its abstract by using the aforementioned software artifact developed by [Dan20]. The results show that three publications are detected as German where about 800 publications are detected to be written in English language. Moreover, our exploratory analysis suggests that most of the publications written in German have an abstract and title in both English and German language. Hence, the language detection result may be off by a minor number of studies. Overall we conclude that the vast majority of abstracts is available in English language.

Building Information Model and its abbreviation show similar distributions of search term matches. Noticeable is the different percentage of matches between the search terms *BIM* and *Building Information Model*. A reason is simply that the abbreviation is more often used than the full version in abstracts, titles and author keywords. Moreover, we would expect a minor number of irrelevant matches of the search term *BIM* due to the usage of the abbreviation in geology for *block-in-matrix* structures described by Adam et al. [AMB14]. A possible strategy to minimize these can be a combination with keywords from the information keyword group such as *information*, *digit* or *Industry Foundation Classes (IFC)*.

Additionally, we require information about the distribution of publications per venue type for the application of inclusion and exclusion criteria during the data preparation subprocess of the CRISP-DM. Therefore, Figure 3.7 shows the absolute amount of studies grouped by venue types. The visualization is based on the deduplicated Scopus data set with a total of 803 tuples. Regarding the data we observe a total number of 434 publications published in *Journals*, followed by 282 publications in *proceedings*.

According to Scopus metadata only 7% or 55 publications of the 803 publications have been published as *Book Series*. Moreover, the data set contains 21 publications with the venue type *Trade Journal* and 11 *Books*. This was not expected as the query we use contains a filter to exclude books, book series and chapters, see Table 3.6.

Following the CRISP-DM process we will continue with the data quality analysis. Therefore, we follow the data mining requirements described in Chapter 3.3.2. First, we analyze the title and DOI to identify potential duplicate studies received by Scopus. In detail the DOIs based duplication identification for the data set received from the Scopus API yielded two potential duplicates:

- *Modern Tunneling Science and Technology: Volume 1. Routledge, 2017.* [TA17]: This is a book with several chapters where each chapter has the same DOI assigned. As those chapters are not the same publication we reject this candidate.
- *Sheet, et al. "Location Information Verification Using Transferable Belief Model for Geographic Routing in Vehicular Ad Hoc Networks." IET Intelligent Transport Systems 11, no. 2 (2017): 53–60.* [SKA+17]: Complete duplicate in regards to title, authors and venue.

The duplication analysis based on the study title yielded four potential duplicates:

- *Rafie, "Interpretation of EPB TBM Graphical Data," North American Tunneling Conference, 1:111–20, 2018.* [Raf18]: Complete duplicate in regards to the study title as it has been published by two different conferences under the same name.
- *Dhote et al., "Quantification of Projection Angle in Fragment Generator Warhead." Defence Technology 10, no. 2 (2014): 177–83.* [DMRS14]: This study has been published in a journal and in a conference proceedings with the identical title and is therefore a duplicate.
- *Cho et al., "Automatic Data Processing System for Integrated Cost and Schedule Control of Excavation Works in NATM Tunnels." Journal of Civil Engineering and Management 20, no. 1 (2014): 132–41.* [CCK14]: Complete duplicate in regards to title, publication date, authors and published venue.
- *Sheet et al., "Location Information Verification Using Transferable Belief Model for Geographic Routing in Vehicular Ad Hoc Networks." IET Intelligent Transport Systems 11, no. 2 (2017): 53–60.* [SKA+17]: this study has already been identified by the DOI duplicate identification and is a duplicate based on title, authors and venue.

In regards to data quality we want to explore the different values of each attribute of the publication entity. We focus specifically on Not Available values as those values

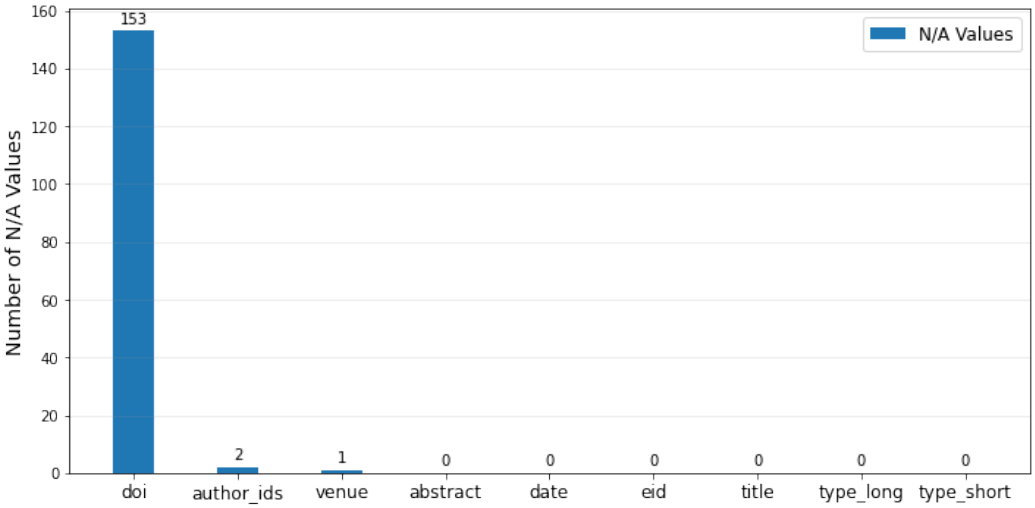


Figure 3.8: Amount of Not Available (N/A) values per publication attribute (included number of publications: 803)

are not provided by the Scopus API. A high amount of N/A values of an attribute is a first indication of poor data quality. Depending on the attribute and the usage during the further data analytics process we may take measures to improve the data quality of specific attributes. Figure 3.8 shows the distribution of N/A values per attribute of the publication entity. Duplicates have been removed from the underlying data set resulting in 803 tuples. We observe that the DOI attribute has the highest amount of N/A values. One reason for this may be that Scopus uses different sources and third party databases which may not provide a DOI for each publication. In total 650 DOI have a non-N/A value. This is a major reason to decline the DOI as identification for a publication as 20% of all tuples do not have a valid DOI assigned. Moreover, we observe that only two publication tuples do not have any author assigned. Those two tuples are the reason for the optionality of the many to many relation between *Author* and *Publication* shown in the ER-Diagram in Figure 3.5. Additionally we see that only a single venue name has a N/A value.

SpringerLink

SpringerLink [Spr21d] is an online literature database of the publisher Springer and provides access to scientific journals, books and reference work. Therefore, the data accessible via SpringerLink is limited to publications published by Springer in contrast to Scopus where publication from multiple publishers are indexed. Moreover, SpringerLink as well as Springer itself are part of the SpringerNature publisher group.

In order to search for studies SpringerLink offers a web-based graphical user interface [Spr21a]. In regards to an interface for data retrieval there is no Application Programming Interface available for automatized communication with SpringerLink. As a result we are

Table 3.7: SpringerLink Query with additional filter constraints

SpringerLink Query
{'query': '((("tunnel" OR "subsurface engineering" OR "tunnelling" OR "tunneling" OR "tunnels" OR "tunnelbau") AND ("BIM" OR "Building information modelling" OR "Building information modeling" OR "Building information models" OR "Building information model")) OR (("continouos tunnelling" OR "tunnel boring machine" OR "TBM" OR "mechanized tunnelling" OR "mechanischer Vortrieb" OR "kontinuierlicher vortrieb" OR "tunnelvortriebsmaschine" OR "TVM" OR "conventenional tunnelling" OR "observational method" OR "conventional tunnelling" OR "NATM" OR "new austrian tunnelling method" OR "drill and blast" OR "konventioneller vortrieb" OR "neue österreichische tunnelbaumethode" OR "NÖT Vortrieb" OR "zyklischer Vortrieb" OR "bergmännischer Vortrieb") AND ("information" OR "digitalization" OR "digitalisation" OR "digitise" OR "digitalisierung" OR "neural network" OR "machine learning")))', 'facet-sub-discipline': 'Civil Engineering', 'date-facet-mode': 'between', 'facet-start-year': '2000', 'facet-end-year': '2019', 'showAll': 'false', 'facet-discipline': 'Engineering', 'sortOrder': 'newestFirst'}

required to find creative solutions to receive the same data set as it has been used for the Systematic Mapping Study. Therefore, we analyze the advanced search and export functionality of SpringerLink and found a solution to receive a Comma-separated values (CSV) based result set. The received data set is identical to the data set we use for the mapping study.

The original query used during the mapping study is shown in Chapter 3.2.2, Table 3.2. Based on this query we apply the exclusion and inclusion criteria specified in the same chapter. Table 3.7 shows the resulting query we use in order to receive the initial data set. Therefore, we add the query as payload to the Hypertext Transfer Protocol (HTTP) `GET` request header and send it to the CSV export HTTP endpoint of SpringerLink [Spr21b]. The HTTP response holds a CSV-formatted data set with 130 publications. For the HTTP-based communication we use the software artifact *requests* created by K. Reitz and maintained by the Python Software Foundation [RF21].

The structure of the result set is just a single table which is caused by the nature of the CSV export. Figure 3.9 shows the columns, identities, derived constraints, and datatypes of the CSV export. The entities name *Item* is derived from the attribute names used for a study representation in the data set. Moreover, we are able to identify the Digital Object Identifier as identity attribute in the received data set. The characteristics of the DOI system architecture enable us to identify a study independent of a specific online library or citation database. This enables us to query DOI registry agencies by providing a DOI in order to receive structured metadata of a study.

The next attributes are the *item title* or title of the study and the *publication title* which is equivalent to the publishing venue. We cannot describe the attribute *book series*

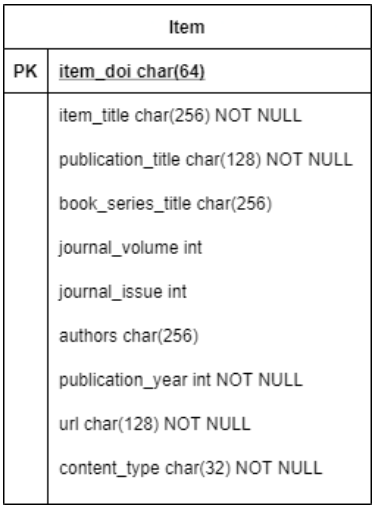


Figure 3.9: ER Diagram based on the attributes received via SpringerLink CSV export

title as the result set does not provide any tuples of this column. *Journal volume* and *issue* are positive integers to identify periodical literature publications such as conference proceedings or journals. The *authors* attribute is a single character sequence which holds the first and last names of all participating authors of the study. The first and last names of an author are inconsistently separated by blank characters. Moreover, the names of different authors are concatenated, e.g. `Mario GalliMarkusThewes`. The next attributes are the *publication year* represented as integer and the web address, *url*, to the study hosted on the SpringerLink web portal. Additionally, each tuple holds a *content type* attribute which is similar to the venue type we describe in the Scopus section, e.g. `Article` or `Chapter`.

When we evaluate the data set received from SpringerLink we observe the lack of abstracts and author keywords. As a consequence the data mining requirements described in Chapter 3.3.2 is not satisfied by this data set. Due to the identity attribute DOI we are able to query third party databases in order to receive metadata for each study. Therefore, we may use the DOIs to query APIs offered by metadata providers such as SpringerNature [Spr21e] to gather additional metadata for each paper in the SpringerLink data set.

In regards to the structure of the Scopus and SpringerLink data set we can observe that the SpringerLink data set has no author, institution or affiliation entity. Moreover, the data structure does not specify any relations and is based on a single entity.

This seems natural as the possibilities of the CSV format are limited compared to Extensible Markup Language (XML) or JavaScript Object Notation (JSON)-based API. Another difference between the data sets is, that SpringerLink uses the Digital Object Identifier as identifier instead of a surrogate identity attribute. This is an important fact as it enables us to query third party data sources for metadata about a publication by

using its DOI.

According to the data analysis process the next task is the identification of relationships. Therefore, the first analysis gathers insight about the categorizations of data tuples based on the content or publication type such as *article* and *book chapter*.

This analysis is important for the exclusion of specific publication types later in the data analysis process. Therefore, Figure 3.10 presents the number of studies in the deduplicated SpringerLink data set grouped by venue types. We observe, that 90 of the 124 studies are *Articles*, 29 items are *Book Chapters* and 5 papers are *Reference Work Entries*. Due to the limitations of the SpringerLink filter features in combination with the CSV export we are not able to exclude books and chapters by adding a filter constraint to the query. Another difference when we compare the attribute values to the Scopus data set is that the SpringerLink data set is not as specific. An example is the separation of *Trade Journals* and *Journals* in the Scopus data set. We discuss our solution to the problem of missing abstracts later in this chapter.

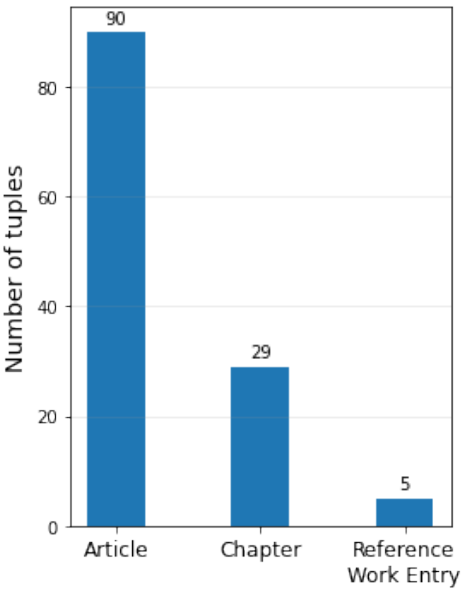


Figure 3.10: Number of tuples (n=124) per venue type in the deduplicated SpringerLink data set

Due to the lack of abstract and author keyword attributes we are unable to analyze relationships and hypothesis for these two attributes. As a consequence we use the title attribute in order to create an initial visualization of found search terms and word stems. Applying a keyword-based analysis on titles may not be as reliable compared to the analysis of a combined analysis of titles and abstracts. Therefore, the main objective of this analysis is to get a first glance on the amount of relevant studies in the context of TIM. In our second analysis we apply the same keyword-based search and visualization as described in the last Section 3.3.3. In detail, we use TIM relevant keywords specified together with domain experts in Chapter 3.2.2 and apply word stemming [Lov68]. Based on the resulting search terms we search the titles of the SpringerLink data set by applying a regular expression. Each study is counted only once for each search term the title contains.

Figure 3.11 shows the relative amount of matched keywords in the title of 124 studies held in the SpringerLink data set without duplicates. We observe a relative similar pattern compared to the keyword analysis of the Scopus data set. 33% of the titles contain the word stem *tunnel* which may indicate a relative high number of unrelated studies. Additionally, 2.4% of titles contain the abbreviation *Tunnel Boring Machine (TBM)* which is very low when compared to the amount of matches of the same search

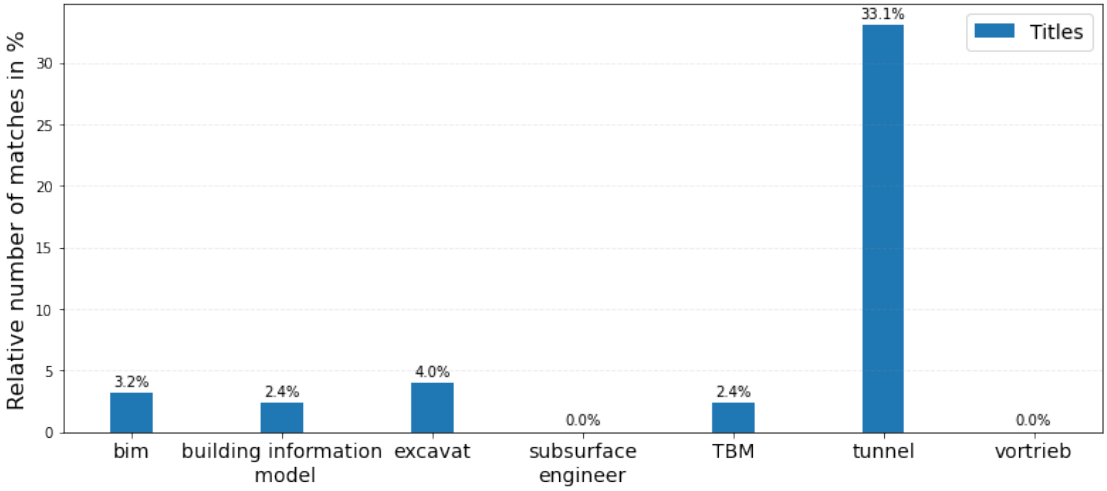


Figure 3.11: Relative number of keywords used in titles in the deduplicated SpringerLink data set (n=124)

term in of the Scopus data set. Another similarity between the SpringerLink and the Scopus data set is the low number of matches for the search term *subsurface engineer* and the German equivalent *vortrieb*. In case of SpringerLink no title contains the two search terms. Moreover, we observe that the English word stem *excavat* occurs in 4% of all titles. In regards to the information model related keywords, 3.2% of titles contain *BIM* and 2.4% contain *Building Information Model* which may be an indicator for a low amount of publications in the field of BIM. As already mentioned during the requirement specification, in comparison to abstracts, study titles are too short in order to make reliable statements about the data set.

The final task of the data understanding process for the SpringerLink data set is the data quality verification. Therefore, we first analyze duplicates based on DOI and titles and continue in a second step by analyzing N/A values in the attributes. As already outlined, we observe data quality issues with the unstructured authors attribute due to its inconsistent use of separators. For further studies we suggest to find alternative solutions in order to receive a structured representation of authors.

The DOI based duplicate analysis yields no results. This is to be expected due to the fact that the DOI attribute is an identity or primary key attribute and therefore unique. The duplicate analysis of the title attribute yields several candidates including several book chapters of the same book and one duplicate reference work:

- *Zilch, et al. "Allgemeine Grundlagen." In Grundlagen des Bauingenieurwesens, 1–378. Berlin, Heidelberg: Springer, 2013.* [ZDKB13]: This is not a duplicate, but the title is too broad and, therefore, matches another book chapter with the same title.

- *Brenguier, Florent. "Noise-Based Seismic Imaging and Monitoring of Volcanoes." In Encyclopedia of Earthquake Engineering, 1561–66. Berlin, Heidelberg: Springer, 2015.* [Bre15]: Full duplicate based on title, authors and venue.
- *Girmscheid, Gerhard. "Bauhof- und Bauinventarmanagement." In Strategisches Bauunternehmensmanagement: Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft, 849–939. VDI-Buch. Berlin, Heidelberg: Springer, 2010.* [Gir10a]: This book chapter is a duplicate as the data set contains the book twice.
- *Girmscheid, Gerhard. "Industrielle Bauprozesse." In Strategisches Bauunternehmensmanagement: Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft, 525–51. VDI-Buch. Berlin, Heidelberg: Springer, 2010.* [Gir10b]: This book chapter is a duplicate as the book occurs twice in the data set.
- *Girmscheid, Gerhard. "Kooperations- und Outsourcingstrategien." In Strategisches Bauunternehmensmanagement: Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft, 287–362. VDI-Buch. Berlin, Heidelberg: Springer, 2010.* [Gir10c]: This book chapter is a duplicate as the data set contains the book twice.
- *Girmscheid, Gerhard. "Organisation von Bauunternehmen." In Strategisches Bauunternehmensmanagement: Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft, 363–424. VDI-Buch. Berlin, Heidelberg: Springer, 2010.* [Gir10d]: This book chapter is a duplicate as the book occurs twice in the data set.
- *Girmscheid, Gerhard. "Risikomanagement in Bauprojekten und Bauunternehmen." In Strategisches Bauunternehmensmanagement: Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft, 697–806. VDI-Buch. Berlin, Heidelberg: Springer, 2010.* [Gir10e]: This book chapter is a duplicate as the data set contains the book twice.

The second part of the data quality assessment is the analysis of Not Available (N/A) values in each attribute of the data set. Therefore, we use the SpringerLink data set where duplicates have been excluded, resulting in 124 tuples.

Figure 3.12 shows the absolute amount of N/A values of each attribute in the data set. We observe a very limited amount of N/A values in the SpringerLink data set. However, the major reason for this observation is, that SpringerLink is not indexing studies from third party sources. As mentioned before, the *Book Series Title* has not a single value assigned and, therefore, has a total number of 100% Not Available values. This is to be expected as, according to our venue type analysis, the data set does not contain any book series. The graph also shows a total of 34 unset values for *Journal Volume* and *Journal Issue*. When we compare this number with the results of the venue type analysis, we can

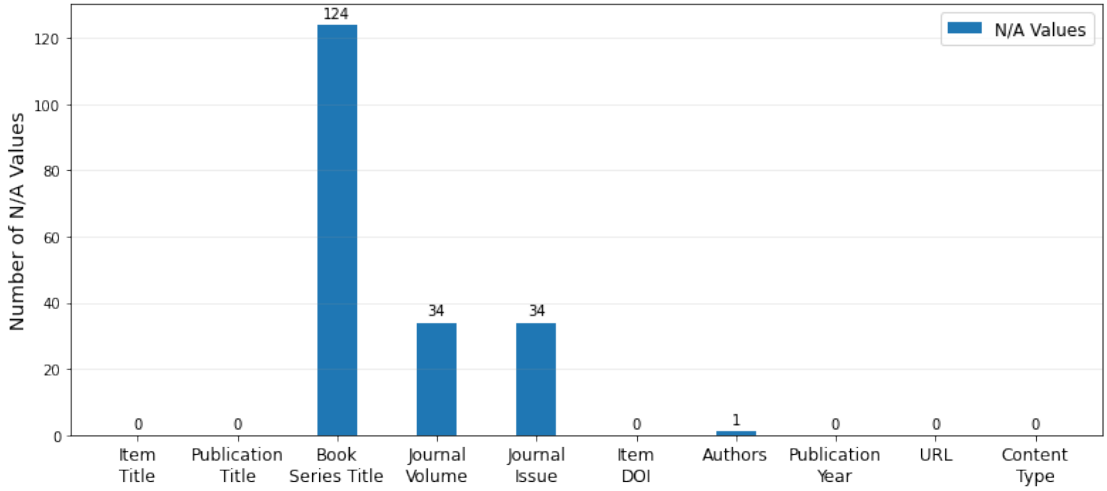


Figure 3.12: Amount of N/A values per Item (n=124) in the deduplicated data set

observe that the data set contains 29 book chapters and 5 reference work. Consequently, those study types are not journals and, therefore, do not have a volume or issue number assigned. As a result the associated attributes are set to N/A. Additionally, we observe a single tuple without author in the data set:

- *Girmscheid, Gerhard, ed. "Ausführungsvorbereitung." In Angebots- und Ausführungsmanagement — Leitfaden für Bauunternehmen: Erfolgsorientierte Unternehmensführung vom Angebot bis zur Ausführung, 89–206. Berlin, Heidelberg: Springer, 2005.* [gir05]

SpringerNature

The dataset received from SpringerLink does not contain an abstract and, therefore, we query the SpringerNature API to receive the abstract for each tuple. SpringerNature is the holding company which unites a group of publishers from different research areas, i.e. Springer. Moreover, SpringerNature offers an Application Programming Interface [Spr21e] to query metadata of studies. The major difference to the API provided by Scopus is, that the SpringerNature API uses DOIs to search for specific studies. This is convenient as the primary key attribute of data set received from SpringerLink is in fact the Digital Object Identifier. Therefore, we send requests to the SpringerNature API containing a set of DOIs we extract from the CSV response described in the last section.

Following the CRISP-DM process, the first task of the data understanding process is the description of data retrieval. The precondition for the data retrieval process of SpringerNature is the data set received from SpringerLink as we use the primary key to retrieve metadata from SpringerNature. In addition to the DOI-based search another characteristic of the SpringerNature API is that requests have to use a pagination

mechanism. This means, that the API splits the result set of the query into multiple pages. A request has to specify the requested page of the data set and the number of results per page in order to retrieve a certain page. The API enforces a specific maximum number of results per page. Therefore, we implemented a simple pagination algorithm in order to receive the full data set.

Alg. 3.1 illustrates the preconditions as well as the algorithm itself. For each page, (i) it evaluates the index of the first and the last entry on the page, (ii) creates the query by concatenating the DOIs for the respective page, (iii) unifies the metadata set with the data set received by querying the API, and (iv) increases the page index by one.

The input data consists of the set of DOIs D extracted from the SpringerLink data set as well as a scalar s, which represents the number of tuples per page. The algorithm starts by initializing the data set and the cursor for the current page (line 2 & 3). Then, for each page the start (line 5) and end index (line 6) of the current pages entries are calculated by using the number of entries per page s. If the current page is the last page according to the number of DOIs n, then the end index is set to n (line 7). The start and end index are used to create the query by concatenating the DOIs with starting index k to end index u (line 9). Then the $SendQuery(q)$ function is called by providing the concatenated DOIs as parameter. The function $SendQuery(q)$ uses the *requests* library [RF21] to send a HTTP-based `GET` request to the SpringerNature API [AG21] and returns the received data set. Next, the received data set is added to the data set (line 10). Finally the page cursor is moved to the next page (line 11) and for the case that there are no pages left, the data set is returned (line 13).

Algorithm 3.1: SpringerNature data retrieval using DOIs

```
   Data: Set of DOIs D = {d_1, d_2, ..., d_n} received from SpringerLink
   Data: Number of tuples per Page s
   Result: Data set A = {a_1, a_2, ..., a_n}
 1 begin
 2     A ← ∅ ;
 3     i ← 0 ;
 4     while (i * s) < n do
 5         k ← i * s;
 6         u ← ((i + 1) * s);
 7         if ((i + 1) * s) > n then
 8             u ← n;
 9         q = (d_k d_{k+1} ... d_u);
10         A ← A ∪ SendQuery(q);
11         i ← i + 1;
12     end
13     return A;
14 end
```

After the elaboration of the data retrieval process we continue with the data structure analysis of the data set received from the SpringerNature API. Similar to the description of the Scopus data set we select relevant attributes based on the data mining requirements described in Chapter 3.3.2. The reason for this decision is, that many attributes are not relevant regarding our data mining goal. A full list of attributes including the JSON representation of a tuple is provided by the SpringerNature API documentation [Spr21c].

Based on the JSON objects in the APIs response we are able to identify relations and entities of the 124 tuples in the data set. As shown in Figure 3.13 the structure of the data set is based on two major entities, the *Item* and the *Creator*. An *Item* is equivalent to a research work published by a publisher in a publication or venue such as a journal, a book or conference proceedings. Each *Item* is authored by one or many *Creators* where each creator is identified by the *identifier* attribute of the *Item*. This seems unintuitive, but the data set does not specify a primary key or identity for the *creator* entity. The data structure does only specify a list of creators with first and last names as a single, character sequence typed attribute of the *creator* entity. Moreover, first and last names are separated by a comma, but the order of the first and last name is inconsistent.

An *Item* has the identity attribute *identifier* which in fact is a prefixed Digital Object Identifier. Similar to the SpringerLink data set, the *contentType* is the type of the publication such as *Article*, *Chapter*, et cetera. Moreover, the *publicationName* represents the name of the venue and the *publicationType* attribute holds the venue type, i.e. *Journal*. Each item also has a *title*, an *abstract*, a *doi* and a *publicationDate* assigned. Additionally, each *Item* tuple has a *keyword* attribute. The *keyword* attribute holds N/As values and therefore does not have a `NOT NULL` constraint.

We continue with the data set evaluation task according to the CRISP-DM data analysis process. Therefore, we use the data mining requirements described in Chapter 3.3.2 to assess the data set received from the SpringerNature API. Each tuple has a title and an

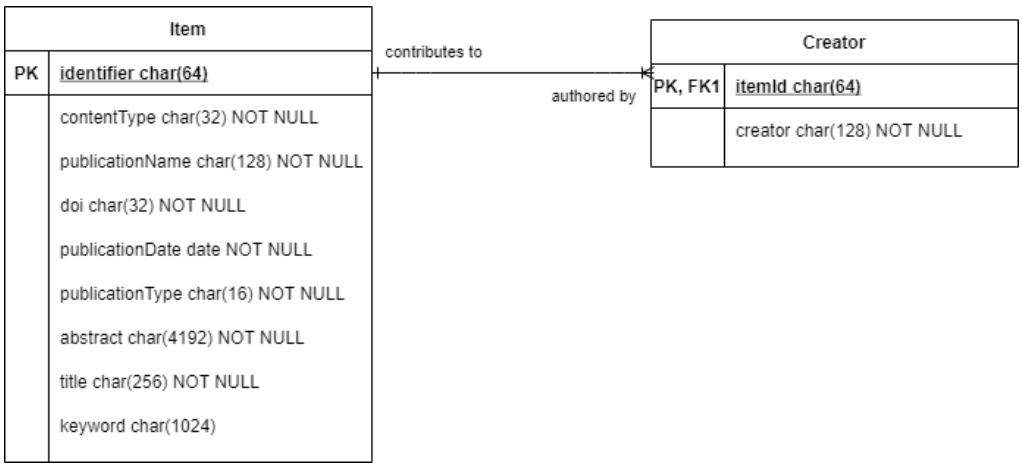


Figure 3.13: ER Diagram based on selected attributes received via SpringerNature

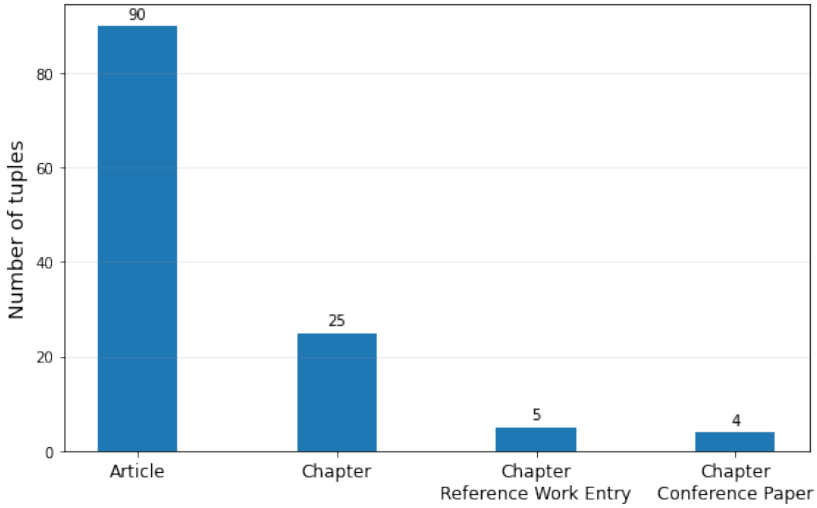


Figure 3.14: Number of tuples (n=124) per venue type in the SpringerNature data set

abstract assigned and, therefore, satisfies the requirements in regards to the duplicate and data mining analysis.

According to Chapman et al. [CCK+00] the next task in the data analysis process is the analysis of relationships and hypothesis. We start this task with the analysis of the number of publications per publication type, e.g. as *Article* and *Book Chapter*. The distribution is relevant in order to apply the exclusion criteria according to Chapter 3.2.3 and feeds directly into the selection task described in Chapter 3.3.4. The second analysis focus on the amount of relevant and irrelvant tuples in the received data set. Based on these observation we propose candidate keywords to identify relevant studies, discussed in Chapter 3.3.4.

Figure 3.14 shows a bar chart based on the amount of studies for the respective venue type. The total amount of tuples in the data set is 124 and is equal to the number of tuples in the SpringerLink data set after removing duplicates. The categorization of studies is based on the values of the *contentType* attribute. It is noticeable, that the publication types *Conference Papers* and *Reference Work Entry* have a *Chapter* prefix. The underlying reason for the prefix is unknown and an exclusive feature of the SpringerNature data set. Moreover, the publication type of SpringerNature is more specific about the different types and has the *Chapter Conference Paper* as a fourth publication type. When we compare this distribution to the publication type related chart of SpringerLink, Figure 3.10, we observe, that the amount of conference papers are added to the *Chapter* category in the SpringerLink data set. Moreover, the most prominent publication type in terms of the amount of tuples is the *Article* with 90 studies. A total of 25 tuples are book chapters and a minor number of 5 studies are *Reference Work Entries*. The least prominent publication type is the *Conference Paper* with a total of 4 tuples.

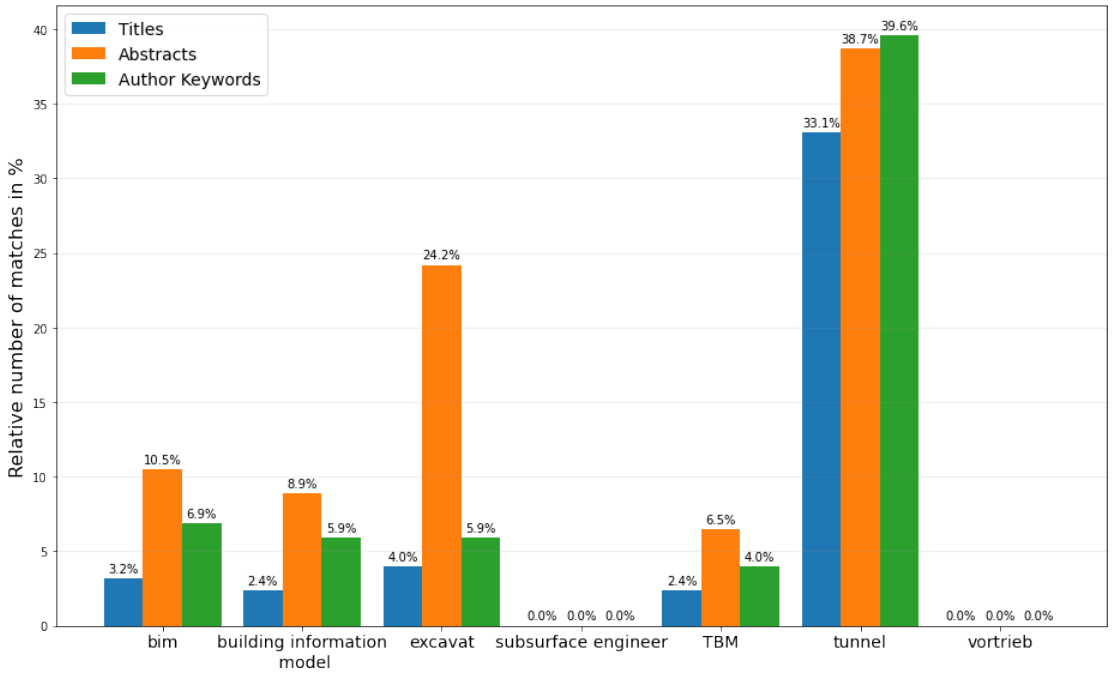


Figure 3.15: Relative number of keywords used in abstracts (n=124), titles (n=124) and author keywords (n=101) of the SpringerNature data set

In order to identify the amount of studies relating to the domain of Tunnel Information Modelling we apply a keyword-based search strategy on the attributes *title*, *abstracts* and author assigned *keywords*. Therefore, the keywords defined together with domain experts, see Chapter 3.2.2 Fig. 3.3, are used and each publication with at least one occurrence is counted as a match of the respective search term. We apply well the established data preparation practice known as stemming in order to match different variations of the same term or keyword [Lov68]. The resulting word stem or keyword is added to a regular expression which ignores any symbols before and after the search term as well as any capitalization. A detailed explanation of the motivation and major objectives of the analysis is described in the relationships and hypothesis task in Chapter 3.3.3.

Figure 3.15 shows the relative amount of matches per keyword or word stem categorized by title, abstracts and author assigned keywords. The plot visualizes the *relative* amounts due to different number of tuples of the underlying data set. Especially, the author assigned keywords attribute has only 101 tuples with non-N/A values in contrast to the 124 tuples of titles and abstracts. In terms of relevance it becomes clear that about 35-45% of all studies are positioned around the topic *tunnel*. We already observed this trend in the keyword analysis of the Scopus data. A total of 33.1 % of all titles, 38.7% of all abstracts and 39.6% author assigned keywords contain the word *tunnel*. Therefore, we propose hypothesis, that a high amount of matches in the *tunnel* group is unrelated on the topic TIM.

The plot shows that *TBM* related studies are a minority in contrast to the results of the Scopus data set. Consequently, the amount of unrelated matches due to a different meaning and usages of the abbreviation *TBM* may be high. Which has two possible reasons, either (i) the abbreviation is used for other research topics than continuous tunnelling or (ii) the result set contains a low amount of studies in the engineering field of Tunnel Boring Machine. The first reason is more likely as indicated by the relative high amount of word stem matches of *excavat* in the abstracts. Therefore, 24.2% of the studies in the SpringerNature data set may be positioned around a topic with a relationship to excavation. Moreover, the SpringerLink data set analysis already indicated a low amount of matches of the word stem *subsurface engineer* and *vortrieb*. This pattern is continued in the SpringerNature data set as none of the search terms occurs in any abstract, title or author keywords.

When we compare the Scopus and SpringerNature analysis in terms of the matches in the *BIM* and *building information model* group we can observe a very similar distribution in the abstract attribute. This may be an indicator that the Scopus data set contains a subset of the SpringerNature data set in the domain of BIM. Therefore, the data preparation process has to be aware of duplicates in the different data sets. Moreover, the proposed hypothesis and required analysis according to the Scopus analysis remains as-is. Due to non unique character of abbreviations as well as the different amounts of matches in the abstract attribute for *BIM* and *building information model* we suspect a minor number of irrelevant studies in the data set. Therefore, a further analysis of study identification using the keyword groups *BIM* and *information* is required.

According to the CRISP-DM data analysis process the final task for the SpringerNature data set is the evaluation of the data quality as well as the duplicate analysis. The first part, the duplicate analysis, has already been conducted for the underlying data set. As a consequence the duplicate analysis of the SpringerNature data set does yield a duplicate candidate, caused by a too broad title [ZDKB13]. The duplicate analysis and duplicate identification of the underlying data set is discussed in Chapter 3.3.3. Therefore, we continue with the analysis of N/A values in the attributes of the SpringerNature data set.

Figure 3.16 shows the amount of N/A values per attribute of the SpringerNature data set. When we compare this plot to the N/A analysis in the SpringerLink Chapter, Figure 3.12, we see several differences. The one N/A value in the *Authors* attribute for a single tuple does not occur in the SpringerNature data set. Therefore, SpringerNature has authors assigned to this study. Moreover, it is noticeable that 23 tuples have no author assigned keywords. The majority of these tuples are book chapters with a total of 22 N/A values in the *keyword* attribute. Only a single *ReferenceWorkEntry* [Bre15] has no keywords assigned.

Wiley Online Library

Wiley online library [Joh21b] is a research work database service containing journals, reference works and books. Wiley offers different methods to retrieve metadata or studies

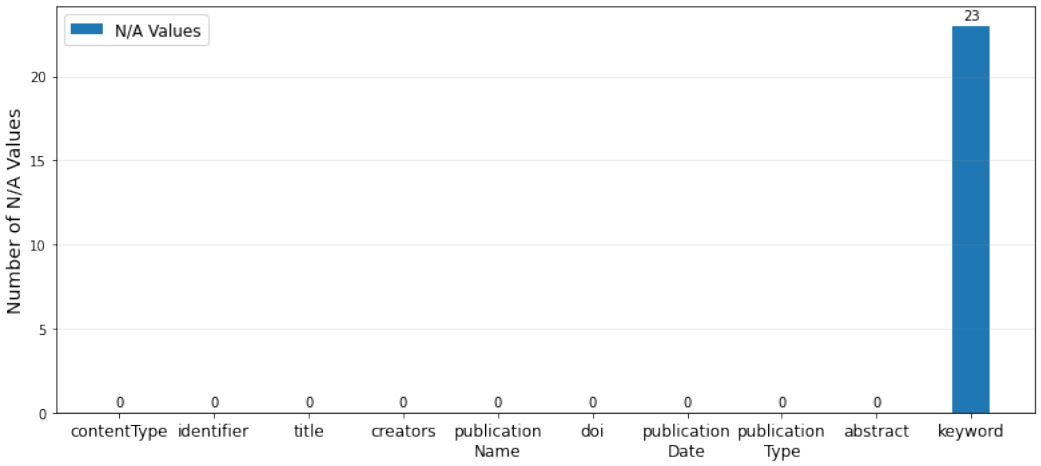


Figure 3.16: Amount of N/A values per attribute of the SpringerNature data set (included number of tuples: 124)

published by Wiley. The preferred solution to retrieve data for text and data mining is the standardized Search/Retrieval via URL (SRU) protocol which is an HTTP-based protocol to query library databases [Joh21a]. This API allows to define search terms, optionally combined with boolean operators, in order to find matching studies. It is important to mention, that the data on the web-portal of the Wiley online library [Joh21b] differs from the data received by the API in terms of metadata. This is especially important for abstracts and other study metadata where issues occurred. For example some abstracts are available in the online portal, but are not included in the received data set, e.g. [SDvB+99]. Another study occurs twice in the data set with two different authors, but the web-portal shows only a single study with both authors. Moreover, the web-portal provides metadata and metrics for each publication which are not part of the data set received from the API [Joh21a]. This metadata includes author assigned keywords, copyright information, information on citations, references, et cetera. More creative attempts and efforts to receive this missing metadata were unsatisfying.

Following the inclusion and exclusion criteria defined in Chapter 3.2.3, we added filter constraints in regards to the publication year and the content or publication type to the original Wiley query. Table 3.8 shows the resulting query we use to receive the XML formatted data set from the SRU API. The search API is limited to specific attributes and does not apply the search term on multiple attributes such as title, abstract and author assigned keywords. This is a major difference when compared to the APIs from SpringerLink or Scopus as both apply the search terms on multiple attributes. Moreover, we encountered issues with the combination of different search criteria, especially the limitation of the research subject lead to empty result sets. As a consequence the retrieved data set may contain publications from different subjects such as nursing or ecology.

We send the SRU request containing the query as HTTP `GET` parameter to the API

Table 3.8: Wiley query with additional filter constraints

Wiley Query
(((dc.description="tunnel" OR dc.description="subsurface engineering" OR dc.description="tunnelling" OR dc.description="tunneling" OR dc.description="tunnels" OR dc.description="tunnelbau") AND (dc.description="BIM" OR dc.description="Building information modelling" OR dc.description="Building information modeling" OR dc.description="Building information models" OR dc.description="Building information model")) OR (dc.description="continouos tunnelling" OR dc.description="tunnel boring machine" OR dc.description="TBM" OR dc.description="kontinuierlicher vortrieb" OR dc.description="tunnelvortriebsmaschine" OR dc.description="TVM" OR dc.description="conventenional tunnelling" OR dc.description="observational method" OR dc.description="conventional tunnelling" OR dc.description="NATM" OR dc.description="new austrian tunnelling method" OR dc.description="drill and blast" OR dc.description="konventioneller vortrieb" OR dc.description="neue österreichische tunnelbaumethode" OR dc.description="NÖT Vortrieb" OR dc.description="zyklischer Vortrieb" OR dc.description="bergmännischer Vortrieb") AND (dc.description="information" OR dc.description="digitalization" OR dc.description="digitalisation" OR dc.description="digitise" OR dc.description="digitalisierung")) AND dc.date>1999 AND dc.date<2020 AND dc.type="article"

using the *requests* software artifact [RF21]. The received data set contains 63 tuples and specifies multiple attributes which we use to derive entities and relations. Figure 3.17 shows the resulting ER-Diagram based on the `record` elements of the XML formatted data set.

The data set does not specify a unique identifier for each publication and, therefore, a combination of two attributes is required as primary key. Naturally, the DOI is a unique identifier, but the data set contains duplicate tuples where different contributors are the only distinction. Therefore, a combination of the DOI and the *contributor* attribute form the primary key of the entity. A *Record* may contain one or many contributors formatted as a list or a character sequence with first and last name separated by a space character. As the *contributor* attribute is not an entity we decided to concatenate lists of contributors to a single character sequence joined by a separator. Similar to SpringerNature and SpringerLink, the author information may not be used anyway due to its unstructured character. For further studies we recommend to query a DOI registry agency for the studies metadata in order to increase the data quality of the author and affiliation attributes.

Each record has a study *title*, a publication *date* and the name of the *venue* where the study has been published. The `NOT NULL` constraint indicates, that this attribute does not contain empty values. Nevertheless, the data set has tuples with missing values in the *abstract* attribute. Moreover, the received data set contains tuples having a list of

Record	
PK	contributor char(1024)
PK	doi char(64)
	title char(256) NOT NULL
	date date NOT NULL
	venue char(128) NOT NULL
	abstract char(4192)
	volume int
	issue char(8)
	issued int

Figure 3.17: ER Diagram derived from the data set received from Wiley API [Joh21a]

abstracts assigned. An in-depth analysis of tuples without abstracts is discussed at the end of this section. Additionally, a record may have a *volume* number and *issue* identifier for periodic literature such as journals. The *issued* attribute represents the year, when the journal has been issued. Those three attributes have a valid value, if the record is an article in any other case those attributes have an *N/A* value.

We continue with the data exploration task after we identified the entity and its attributes as well as the absence of any relationships. Therefore, we analyze dependencies in the data and propose hypotheses related to the research question for further analysis. In the last sections we started this analysis with a discussion regarding the distribution of studies per venue type accompanied by a bar chart. But this analysis cannot be executed for this data set as it does not contain an attribute or indicator for the venue type of a study. Based on the missing values in related attributes we suspect that the majority of tuples are journal articles. Therefore, we discuss the occurrence of missing values in the *volume* or *issue* attribute during the data quality assessment.

The main objective of the upcoming analysis is to identify the amount of relevant studies in the Wiley data set and to identify important features as well as relationships. Therefore, we apply a regular expression matching keywords and word stems to abstracts and titles. It is not possible to apply this procedure on author assigned keywords as the Wiley data set does not contain such an attribute. We select the search terms based on keywords describing Tunnel Information Modelling, which we specified together with domain experts during the mapping study process, see Figure 3.3. Then, we apply stemming [Lov68] to reduce keywords and keyword combinations to a common word stem. This generalization enables the matching algorithm to identify variations of the same word. For example, the word stem *excavat* is derived from the noun *excavation*, but represents

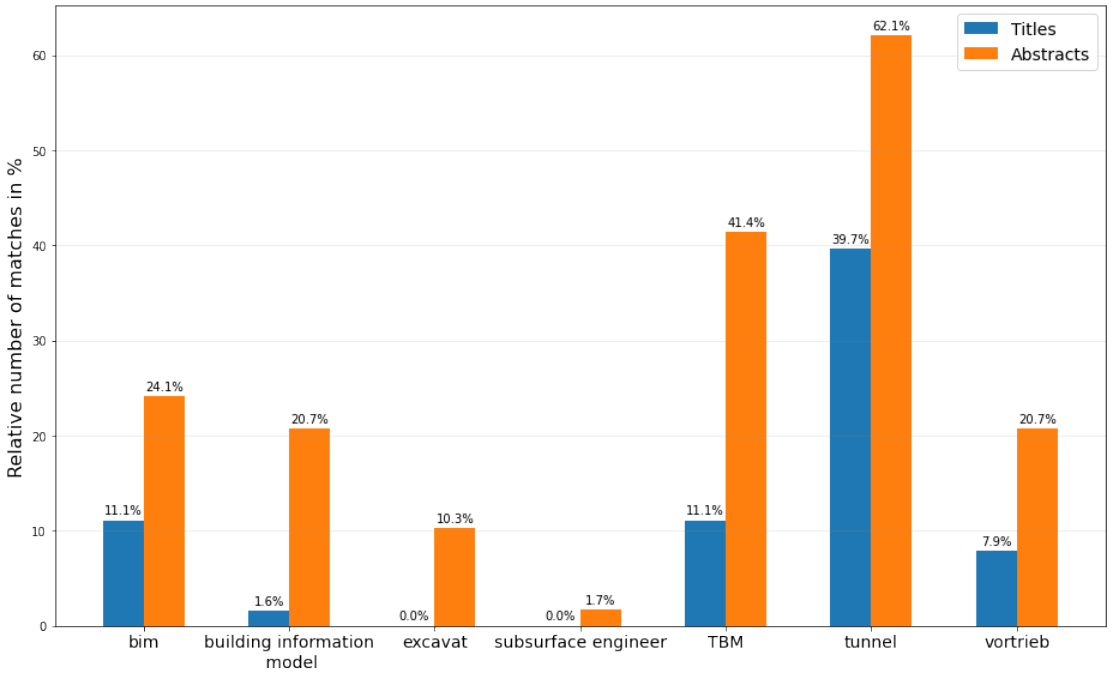


Figure 3.18: Relative number of keywords used in abstracts (n=58) and titles (n=63) of the Wiley data set

variations with common endings such as *ing* or *ion*. The matching algorithm of the regular expression ignores lower and upper case usage. Moreover, only the first occurrence of a search term in an abstract or title is considered a match for this search term and the respective attribute.

Figure 3.18 shows the relative amount of matched keywords in abstracts and titles of the Wiley data set. The two major reasons for the usage of relative numbers are that (i) the underlying set of abstracts and titles have different sizes and (ii) to enable a comparison of the amount of matches with the analysis from other data sets. *Tunnel* is the most prominent keyword with a relative amount of 62.1% with at least one occurrence in abstracts and 39.7% in titles. This indicates, that a majority of studies in the data set is related to the topic tunnelling and, therefore, may be related to TIM. Also, the relative high amount of occurrences in the titles indicates that the *tunnel* topic is well represented in the data set. As a consequence the amount of unrelated studies may be lower than in the Scopus and SpringerNature data set, due to the high percentage of studies containing the *tunnel* search term. Therefore, we suggest to analyze a relationship of abstracts containing the *tunnel* keyword in regards to matches of keywords from the *Information & BIM* group. A major reason is, that TIM is the application of Building Information Modelling principles in the tunnelling domain.

The abbreviation *TBM* has the second most relative number of matches with 41.4% in

abstracts. In contrast, only 11.1% of the titles contain the keyword *TBM*. This may suggest, that studies are not primarily focusing on the topic *TBM*, but describe topics with relations to Tunnel Boring Machines. We observed similar patterns in the Scopus and SpringerNature data set. Our proposal is to analyze potential relations between the keywords *TBM*, *BIM* and *Information* with respect to the amount of matches per group. This may indicate a potential relationship between the occurrence of keywords such as *information model* in *TBM* related studies and, therefore, show research work of TIM in the field of mechanized tunnelling. Due to the high number of matches of the *tunnel* keyword it seems natural, that only a low amount of *TBM* matches are unrelated to tunnelling.

24.1% of all abstracts contain the keyword *BIM*, where only 11.1% of the titles contain the information model abbreviation. Nevertheless, this is twice the number of matches in abstracts compared to the Scopus (12%) and SpringerNature (10.5%) data sets. Moreover, this is a strong indicator for an increased number of publications in the field of TIM, as 62% of abstracts contain the keyword *tunnel*. The keyword *Building Information Model* shows a similar pattern, where 20.7% of abstracts and only 1.6% of the titles contain the keyword combination. The low amount of matches in the title attribute is expected as this phenomena is observable in the data sets of SpringerNature and Scopus as well. One major reason is the length of the study title, which is rather short compared to the length of an abstract. Therefore, the usage of an abbreviation is more favorable for titles.

The surprise of the Wiley data set is the relative high amount of matches of the German keyword *vortrieb*. This was not expected, especially when comparing the numbers to the results of the Scopus and SpringerNature data sets. In detail 20.7% of abstracts and 7.9% of the titles contain this word stem. On the other hand the keyword *excavat* is only matched in 10.3% of the abstracts, but never in title. This indicates that at least a fifth of all abstracts in the Wiley data set are written in German language and are related to excavation. Therefore, an analysis of the abstract attribute in regards to the used language is required and may influence tasks of the data preparation sub-process. Finally, only one abstract contains the keyword *subsurface engineer* which indicates that the keyword may not be relevant in this data set.

The next task according to the CRISP-DM data analysis process is the assessment of data quality and, therefore, analysis of duplicates and N/A values. As outlined in the analysis of the data set structure we encounter one potential duplicate with different authors in the received data set:

- *Goger, Gerald, and Tobias Bisenberger. "Tunnelling 4.0 – Construction-Related Future Trends." Geomechanics and Tunnelling 11, no. 6 (2018): 710–21.* [GB18]: The data set contains two identical tuples for each author of the paper.

The derived data structure is identified by a combination of the *doi* and the *contributor* attribute as the data set contains two identical tuples with different authors. The

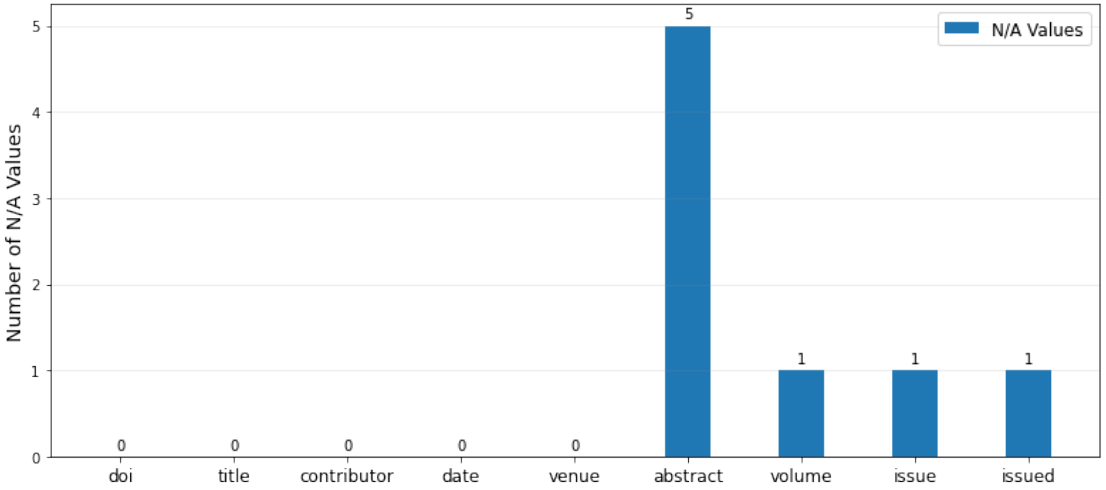


Figure 3.19: Amount of N/A values per attribute of the Wiley data set (included number of tuples: 63)

handling of this duplicate depends on the set of selected attributes as one duplicate may be excluded, if the author attribute is not selected. Another option is to merge this duplicate by adding the author to one tuple and dropping the remaining tuple. Both strategies enable the use of the unique *doi* as the identifying attribute of the entity.

In regards to the analysis of N/A values in each attribute we use the identified entity and its attributes according to the ER-Diagram of the Wiley data set, see 3.17. Therefore, Figure 3.19 shows the number of N/A values per attribute. The bar chart shows, that five tuples [SDvB+99], [QYHW16, VCP+13], [PBC+19] & [SSJ+17] have no abstract assigned. Therefore, we use the DOI attribute to search for the studies on Wiley's web-portal [Joh21b]. According to the Web-portal results all studies have an abstract assigned. Nevertheless, none of these publications is related to the research area of tunnelling and, therefore, may be rejected due to a missing abstract. Chapter 3.3.4 contains a detailed discussion on data selection. Moreover, the bar chart shows one missing value in the attributes *volume*, *issue* and *issued*. In fact it is only a single study [PBR+16] which has no values assigned to the aforementioned attributes. According to the data mining requirements specified in Chapter 3.3.2, the attributes *volume*, *issue* and *issued* are not relevant to answer the research question, hence there is no reason to reject this tuple.

We continue with the data set evaluation task after finishing the analysis of the data quality and the discussion of missing data. Based on the data mining requirements as well as the analysis of missing values we propose to reject tuples with empty abstracts. The abstract of a study is required in order to answer the research question of the design science artifact. Therefore, the Wiley data set tuples with non-N/A abstracts fulfill the data mining requirements defined in Chapter 3.3.2. Additionally, the requirement of a

title attribute without N/A values is satisfied by the data set as well.

3.3.4 Data preparation

Following the CRISP-DM data analysis process this section discusses data exclusion, preparation and cleaning measures. First, we describe exclusion criteria for data selection such as attributes and attribute values. Then, we adapt the CRISP-DM process by adding an relationship analysis of the merged data set. In the last chapter, we proposed relevant search term combinations in order to identify relevant studies. The reason for the adaptation is, that the analysis of relationship is reasonable after the data set has been merged. At the end of the section we describe the data cleaning procedure.

Data selection

For the data selection task we use the inclusion and exclusion criteria described in Chapter 3.2.3. The major reason is that the data mining goal is to identify and classify relevant TIM studies in order to provide a state-of-the-art set of relevant studies. Therefore, we apply the following exclusion criteria on the data set:

- *Duplicates*: identified based on DOI, title attributes
- *Books*: identified based on venue classes
- *Papers:*
  - without available English or German abstracts
  - that do not contain the word stem tunnel

We suspect numerous duplicates in the unified data set as discussed in the section before and, therefore, require duplicate identification and handling. The duplicate analysis in the past section show, that title- and DOI-based duplication measures are applicable. Hence, we eliminate duplicates based on the study titles and, if available, DOI attributes. The deduplication procedure is applied after the data sets are unified. We are able to exclude books and theses, based on attributes which classify the venue type of a study. This criteria applies to the data set received from Scopus and SpringerNature and excludes studies with the following venue types:

- *Book*
- *Book Series*
- *Chapter*
- *Chapter ReferenceWorkEntry*

We apply this exclusion criteria to the SpringerNature attribute *content_type* and the Scopus attribute *type_long*. As described in the last section, we added filter constraints to the queries as part of the data retrieval process to exclude studies in regards to

study types (e.g. Theses), online publication date and subject areas. Therefore, it is not required to apply additional exclusion criteria in regards to research area and publication date. Moreover, the exploratory data analysis shows that there are several studies in the Wiley data set without abstracts. Those tuples are excluded as an abstract is mandatory for the data analysis process. The last criteria helps to identify studies which are thematically positioned in or around the tunnelling domain. If an abstract of a study does not contain the word tunnel or a derivation of the word stem tunnel we assume that the research is not relevant. As a consequence, we exclude studies unrelated to the tunnelling domain from further analysis.

During the iterative design cycle of the data science methodology, we identified three tuples which have at least two abstracts assigned as an array. This is not surprising as some journals are publishes in multiple languages, e.g. English and German or English and Chinese. Therefore, we define the following preferences and requirements for the selection of abstracts and arrays of abstracts:

1. The abstract has more than 250 characters
2. An English or German abstract is available
3. English abstracts are preferred over German abstracts

The first requirement is derived by observation, as we observe that some elements of the abstract arrays are in fact the English title of studies written in German. To overcome this issue, we define a minimum number of characters of 250 to exclude potential study titles in the abstract attribute. This value is based on the shortest valid abstract in the data set, which has 258 characters. Moreover, if the abstract is not written in English or German the study is excluded. In cases where an English and German version of the abstract is available, we prefer the English abstract. Our preference is based on the results of the exploratory data analysis conducted in the last section, which indicate that the majority of abstracts is available in English language.

Study	
PK	title char(512)
	abstract char(4192) NOT NULL
	language char(2)
	doi char(64)

Figure 3.20: Selected attributes and resulting *Study* entity

Figure 3.20 shows the attributes of the merged data set. We select those attributes from the Scopus, SpringerNature and Wiley data sets according to the data mining requirements defined in Chapter 3.3.2.

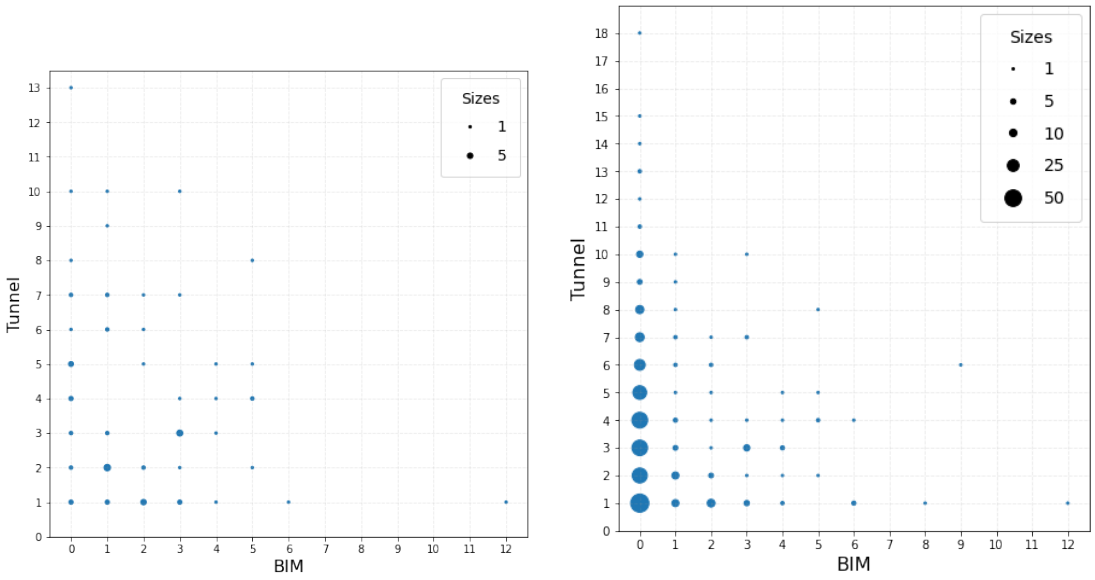
After the data set is merged we eliminate duplicates using the title and DOI attributes. Therefore, the title attribute becomes the identity attribute. The DOI attribute does not qualify as identity as several tuples contain N/A values in this attribute. Additionally, we determine the language of each study by applying the software artifact from Nakatani [Nak10] for each abstract. The result is an ISO 639-1 compliant language code which is assigned to each tuple [ISO16]. The distinction between English and German studies is important for the development of potential text classification models. The resulting data

set has 381 tuples where 72 are labeled as TIM-relevant according to the result set of the mapping study.

Study identification

After the data selection and unification of the different data sets we want to find search terms which identify relevant studies. Therefore, we apply a regular expression based search on the abstracts of the data set received by the data selection task using the keywords proposed in the exploratory data analysis conducted in the last section. The search is applied on the abstract attribute of relevant studies according to the Systematic Mapping Study as well as on the data set received from the data selection task of the design science artifact. Hence, we are able to compare the amount of studies containing a keyword and to analyze the resulting distribution of a search term. The upcoming scatter plots show the amount of keyword matches of a study for each search term on the x- and y-axis.

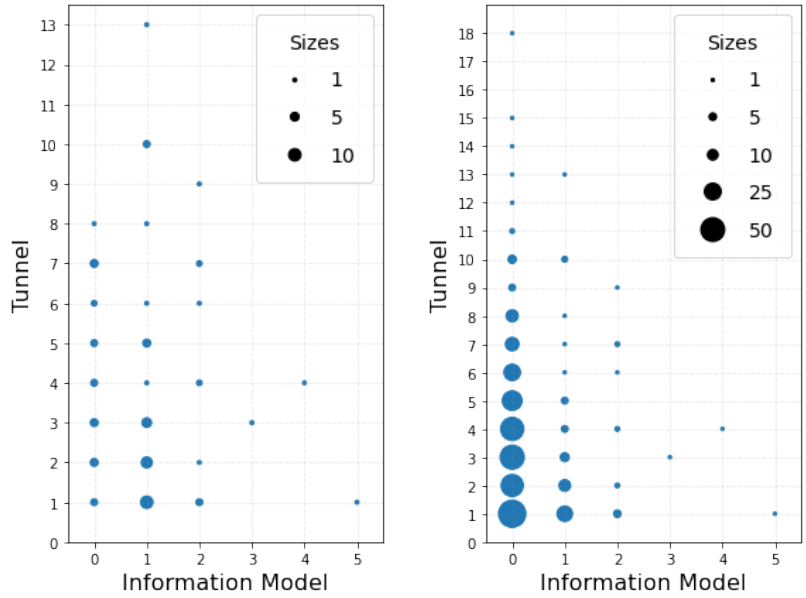
Figure 3.21 relates the number of matches for the keywords *BIM* and *tunnel* using different data sets. Plot 3.21a visualizes this search term matches for the data set retrieved by conducting the Systematic Mapping Study (SMS) with 72 tuples. Whereas the scatter plot 3.21b shows the amount of matches received by applying the search on the abstracts of the Data Mining (DM) artifact data set with a total number of 381 tuples. The first observation is, that only 72 or 19% of the 381 tuples of the DM artifact data set are relevant according to the results of the mapping study. Therefore, the data set is



(a) Data set: mapping study, n=72

(b) Data set: DM artifact, n=381

Figure 3.21: Number of Matches of search terms *Tunnel* & *BIM* in abstracts



(a) Data set: mapping study, n=72 (b) Data set: DM artifact, n=381

Figure 3.22: Number of Matches of search terms *Tunnel* & *Information Model* in abstracts

unbalanced in regards to the labels of the binary classification which is important for the model creation and evaluation.

Moreover, the plots show that all studies contain the word stem *tunnel*. This seems natural as the exclusion criteria of the mapping study and the data selection task of the DM artifact define measures to exclude studies which are unrelated to the tunnelling domain. Based on the size of the circles we observe that the amount of studies which do not contain the search term *BIM* differs in both data sets. In detail, 91 of the 381 (24%) DM artifact abstracts contain at least one occurrence of the search term *BIM*. Whereas 52 out of 72 (72%) abstracts in the SMS data set contain the search term *BIM*.

Additionally, we observe that abstracts containing the search term *BIM* are not necessarily relevant according to the results of the mapping study. In detail 33 abstracts in the DM artifact data set contain the keyword *BIM* once compared to 18 abstracts of the SMS data set with one occurrence of the search term. Therefore, we conclude that BIM is a candidate keyword in order to identify relevant studies, but in order to raise the precision we recommend a combination with other keywords.

Hence, we continue with the analysis of the matches of the keywords *Information Model* and *Tunnel*. Figure 3.22 shows the amount of matches for those keywords in the abstract attribute of the SMS and DM data sets.

We observe, that the amount of abstracts which do not contain the search term *Information Model* shows a different distribution based on the patterns of both data sets. In detail,

71 of the 381 (18%) abstracts in the DM artifact data set contain the search term *Information Model*. The SMS data set on the other hand contains 48 abstracts (66%) with at least one occurrence of the keyword in the abstract attribute.

The amount of abstracts with at least one occurrence of the search term *Information Model* show a very similar pattern in both data sets. This is an indicator for a low amount of false-positives and consequently an increased precision. Therefore, we conclude that the keyword *Information Model* qualifies as search term candidate to identify relevant studies and may be combined with other search terms.

Other word stems from the information keyword group such as *artificial intel* or *machine learn* have no matches in any abstract of the mapping study data set. Moreover, the analysis of keywords derived from the word stem *digit* show a relative high amount of false-positives which indicates a bad identification of relevant studies. As a consequence, these keywords and word stems may not be used by the design science artifact to identify relevant studies.

Data cleaning

The data cleaning task focuses on the *abstract* attribute since potential classification models use the abstract as input data. Therefore, we apply the procedure shown in Algorithm 3.2 to clean and prepare the abstracts for the modeling sub-process. In order to clean the data we use the software artifact spaCy [HMVLB20] in version 2.1 for the identifications of word classes, tokenization and lemmatisation.

The input data for the procedure is a set of abstracts D containing TIM-related studies. The set W contains all words from all abstracts where each word is available in lowercase, also known as bag of words representation. We define an index set I for our bag of words, where each abstract d is an ordered set of words w_i.

First, the matrix Q and the set of tokens T are initialized. Second, we remove copyright notices from the abstract d as those add unnecessary noise which we want to minimize. Then, we remove any punctuation and line breaks as those are irrelevant for further analysis. Moreover, words containing numbers and stop words, such as *the, is, at, etc.*, are removed from each abstract as well. The next data cleaning step preserves the abbreviation *BIM*, proper nouns, nouns, verbs, and adjectives and removes words of other word classes from the abstract. The result of this procedure are words describing the essence of the study based on its abstract also known as the corpus.

Each of these character sequences is split up into tokens. A for-loop iterates over all tokens to receive the lemma representation of each token. The result is an $n \times m$ matrix Q, where the rows represent abstracts and the columns are the lemmas for each abstract.

3.3.5 Modeling

This chapter describes the selection of modeling techniques for text classification problems using well-established procedures. Then, metrics to measure the performance of the

Algorithm 3.2: Data cleaning procedure

Data: Set of lowercased abstracts $D = \{d_1, d_2, \ldots, d_n\}$ $n \in \mathbb{N}$
Data: A set of words $W = \{w_1, w_2, \ldots, w_m\}$ $m \in \mathbb{N}$, where
$\forall d \in D \; \exists \; I \subseteq \{1, \ldots, m\} : d = (w_i)_{i \in I}$
Result: Array Q of lemmas for each abstract $d \in D$

```
1  begin
2      Q ← ∅;
3      T ← ∅;
4      forall d ∈ D do
5          Remove copyright notices from d;
6          Remove punctuation and line breaks from d;
7          Remove stop words and non-alphabetic words from d;
8          Filter unreleated word classes from d;
9          T ← Tokenize(d);
10         forall t ∈ T do
11             Q_{d,t} ← Lemmatise(t);
12         end
13     end
14     return Q;
15 end
```

classification of a model as well as the test design is described. Finally, used input parameters of the models are presented.

Modeling technique selection

Our selection of the modeling technique is based on well established methods in the domains of Information Retrieval (IR), pattern recognition and text classification. As part of IS the major task of IR is the design of retrieval models and the ranking of documents. According to the systematic literature analysis conducted by Mirończuk and Protasiewicz [MP18], neural networks and transferred learning have not yet taken the dominant role in text classification problems. Therefore, we apply a combination of ranking followed by a supervised learning method to classify the English bag-of-words representation of the abstracts. Especially the combination of the ranking followed by the classification using a Logistic Regression (LR) or Support Vector Machine (SVM) based models are well established according to Lin [Lin19] as well as Mirończuk and Protasiewicz [MP18]. In more detail we apply the Atire Best Match 25 (BM25) ranking algorithm discussed by Trotman et al. [TPB14] before training a text classification model. Then we train, test and evaluate the following three different model types for the binary classification task to identify relevant studies in the TIM domain:

- Gradient Boosting Tree, using the Python implementation of XGBoost v1.3.3

Table 3.9: Confusion Matrix [Rau17]

# of samples correct in A B True Positive (TP) N_{11}	# of samples correct in A, wrong in B False Positive (FP) N_{10}
# of samples wrong in A, correct in B False Negative (FN) N_{01}	# of samples wrong in A B True Negative (TN) N_{00}

[CG16]

- Support Vector Classifier, using scikit-learn v0.24.1 [PVG+11]
- Logistic Regression, using scikit-learn v0.24.1 [PVG+11]

The classification of TIM-relevant studies is a binary decision: either a study is relevant in the domain or it is not. Therefore, each abstract is assigned a binary label (0 or 1) to encode if it is relevant or not according to the tuples of the mapping study result set. The discussion of the multi-class classification of studies regarding excavation types and life cycle phases is presented in Chapter 4.2. Regarding the six German papers received by the DM artifact it is ineffective to train and test a separate model due to the low amount of available tuples.

Test design

As already introduced in Chapter 3.3.4 we use several metrics to measure the quality of a model. The most common measures for model quality assessment are precision (π), recall (ρ) and F_1-Score. We use the metric definition of García et al. [GPUA14]:

$$\pi = \frac{TP_i}{TP_i + FP_i}, \ \rho = \frac{TP_i}{TP_i + FN_i} \tag{3.1}$$

where TP_i are the number of true-positive and FP_i represent the number of false-positive matches of studies in the class c_i, Table 3.9 shows an exemplary confusion matrix. The F_β measure combines recall and precision in a single measurement where β defines the weight towards precision or recall $0 \leq \beta \leq \infty$. If β is lower than one, the emphasis is on precision, where a value above 1 weights recall higher.

$$F_\beta = \frac{(\beta^2 + 1) \cdot \rho \cdot \pi}{(\beta^2 \cdot \pi + \rho)} \tag{3.2}$$

The value 1 represents the balance between precision and recall and is often used. Therefore, the equation is transformed into:

$$F_1 = \frac{2 \cdot \rho \cdot \pi}{(\pi + \rho)} = \frac{2 \cdot TP_i}{2 \cdot TP_i + FP_i + FN_i} \tag{3.3}$$

In order to measure the averaged F_1 value a micro- or macro measurement is used instead of using the F_1 average for each category. The macro average focuses on uncommon categories and is defined by

$$F_1^{\mu} = \frac{\sum_{i=1}^{|C|} 2 \cdot TP_i}{\sum_{i=1}^{|C|} (2 \cdot TP_i \cdot FP_i \cdot FN_i)} \tag{3.4}$$

Whereas the micro average gives more emphasis to performance of more frequent classes and is defined by

$$F_1^{M} = \frac{\sum_{i=1}^{|C|} F_{1_i}}{|C|} \tag{3.5}$$

The accuracy measurement is not applicable to our classification problem, as the classes are unbalanced, see Chapter 3.3.4. Therefore, we use the Matthews Correlation Coefficient (MCC) introduced by Matthews [Mat75] to measure the quality of binary classifications. The coefficient is applicable to unbalanced classes and returns $-1 \leq MCC \leq 1$, where

$-1 \ldots$ represents a disagreement between prediction and observation,
$0 \ldots$ indicates that the prediction is not better than a random selection and
$1 \ldots$ is the prefect prediction.

We use the definition by Matthews [Mat75]:

$$MCC = \frac{(TP_i \cdot TN_i) - (FP_i \cdot FN_i)}{\sqrt{(TP_i + FP_i) \cdot (TP_i + FN_i) \cdot (TN_i + FP_i) \cdot (TN_i + FN_i)}} \tag{3.6}$$

The Matthews coefficient considers the proportion of each class of the confusion matrix. As a consequence the value of the MCC gets closer to 1, if the amount of true-negative and true-positive, predicted by the classifier, increases.

For model testing we apply a well established approach by splitting the labeled and cleaned data set into a train set with 70% of the tuples and a test set with 30% of the original tuples. The tuples are randomly selected and assigned to one of the sets and, therefore, the train and test set are independent. Then we apply a k-fold cross validation on the classifier, measuring the arithmetic mean as well as the standard deviation of the metrics shown above. Cross validation is a procedure known from statistics and is used to assess the performance of a model or learning algorithm. First, the training data set is shuffled and then partitioned into k subsets of equal size. Then, k different models are trained on $k-1$ partitions and each model is tested against a subset not used for training. Naturally, the selected test partitions are different for each model. Finally, the predictions of the model are evaluated against the test set resulting in the quality measurements recall, prediction, F_1 and MCC.

Table 3.10: Required technical resources to deliver the design science artifact

Function	Parameters
Okapi BM25	$k_1 = 1.5$; $b = 0.75$; $query =$ "*bim build building information model*"
Support Vector Classifier	$C = 1000$; $gamma = 10^{-3}$; $kernel = rbf$
Gradient Boosting Regressor	$booster = gbtree$; $max\_depth = 10$; $n\_estimators = 2$; $objective = binary\_hinge$
Logistic Regression	$C = 0.5$; $penalty = l2$; $class\_weight = \{0 : 0.24, 1 : 1\}$; $fit\_intercept = True$

Rauber [Rau17] recommends different hypothesis testing methods in order to find the model that performs best, given a selected test set. Therefore, the null hypothesis or $H0$ is defined as *both classifiers are statistically equal*. The amount of FP and FN of each classifier A and B serve as input for the hypothesis test. Consequently, the null hypothesis is defined by

$$N_{01} = N_{10} = \frac{(N_{01} - N_{10})}{2} \tag{3.7}$$

$H0$ is rejected, if the probability of the null hypothesis is less than 5% or $\chi^2 > 3.841$, for one degree of freedom, according to the tabulated χ^2 distribution. Therefore, we use the recommended discrete Edward's correction formula of the McNemar's test in order to find χ^x which Rauber [Rau17] defines as

$$\chi^2 = \frac{(|N_{01} - N_{10}| - 1)^2}{N_{01} + N_{10}} \tag{3.8}$$

Model building

In order to tune the hyperparameters of the gradient boosting tree, Logistic Regression (LR) and the Support Vector Machine (SVM) classifiers we apply an iterative grid search to find a local optimum of hyperparameter values. Therefore, the prediction returned by each parameter combination is tested using a 10-fold cross validation. The estimator with the maximum value in terms of the arithmetic mean MCC metric received by the 10-fold cross validation is selected. Table 3.10 shows the parameters yielding the highest MCC performance measurements in average for the English bag-of-words representation. The search terms for the query have been selected based on the results of the study identification analysis conducted in Chapter 3.3.4.

CHAPTER 4

Results

In this chapter we present results to answer the research questions of the design science research method and, therefore, the mapping study and the design science artifact. First, we analyze the mapped studies and interpret the results of mapping the papers accompanied by graphs to answer the research questions. Then, we analyze the results from the design science artifact and evaluate text classification models against the outcome of the mapping study.

4.1 Systematic Mapping Study

In this section we present results and analysis to answer the research questions (RQ1-RQ4) of the mapping study. According to Chapter 3.2, Figure 3.2 this section represents the last activity of our mapping process namely the *Mapping of Papers*. Hence, our results and analysis are based on a set of classified papers from the *Classification using Abstracts* activity. At the end of each subsection we summarize the main findings to answer the research question.

4.1.1 Overview and result presentation

We propose an interactive visualization of the study results to summarize the research activities with the topic Tunnel Information Modelling. This visualization enables people and organizations in the environment, such as scientists, students and geologists to apply the studies results. Therefore, the visualization aligns with the guidelines proposed by Hevner et al. [HRM+04] and is part of the design science research cycles presented in Fig. 3.1.

In detail the proposed visualization uses a graph based representation where nodes are connected by edges. The root node is represented by the main research topic of the study, TIM. Figure 4.1 shows the root node and the first level of nodes.

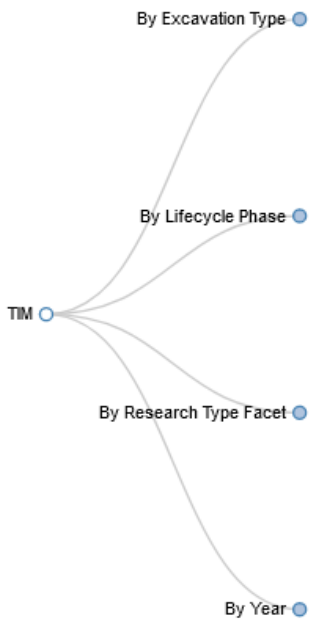


Figure 4.1: Interactive visual presentation of the mapping study results by category

Users may interactively select nodes to expand a new set of categories and classifications and, therefore, drill-down in order to find the studies of interest. The first level of nodes holds the classification categories we defined during the *Classification using Abstracts* activity, see Chapter 3.2.4. Additionally, the option to show studies by their publication year has been added as the distribution of published studies during the last decades is part of the first Research Questions. This enables interested scientists to select a categorization of the studies:

- by excavation type
- by lifecycle phase
- by research type facet
- by year

On the second level each category is connected to its respective classes, e.g. the excavation type node is connected to continuous, conventional excavation and unspecified for studies which do not specify an excavation method. The third level represents the studies for the expanded classes and is a leaf or terminal node. Figure 4.2 shows the fully expanded tree for the excavation category. On the third tree level the user is able to select a study in order to show key information, such as:

- the study title
- the abstract
- the assigned Digital Object Identifier (DOI)
- the source Uniform Resource Locator (URL) for further information
- the study authors
- the publication year
- the name of the publication or conference
- the volume and pages if available

This detailed information is shown in a separate panel located next to the graph to enable the user to quickly find study related information. Nodes are either filled with blue or white color to indicate whether the node has child nodes assigned and, therefore, is expandable. Moreover, the selected study or terminal node is highlighted to reflect the current study selection.

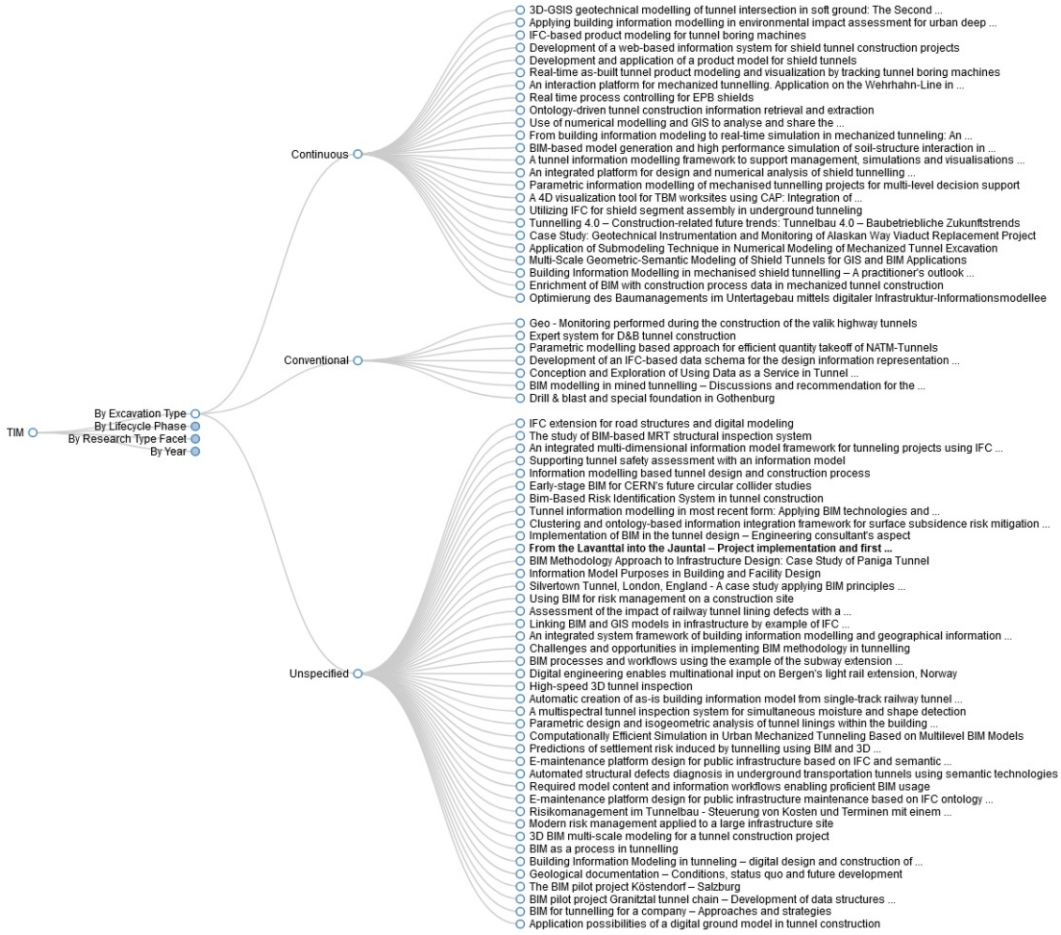


Figure 4.2: Tree visualization of the studies classified by excavation type

4.1.2 RQ1: bibliometric key facts of TIM publications

In Figure 4.3 the distribution of published studies for the years 2002-2019 are shown. In order to focus on relevant years in regards to publications we decided to exclude the years where no studies were published. The analysis of the number of publications in the field of information model application in the tunnelling domain shows a growth of interest in this topic, particularly since 2011. In 2018 the number of published studies reached a peak of 16 publications. The graph shows that the increase of published studies is interrupted in 2015. Therefore, we identified governmental digitization strategies in the design and construction of public infrastructure facilities as main incentives for a delayed increase of publications in the period from 2011 to 2019 [Bun15], [UK 11], [UK 12]. As a consequence studies are being published as part of large public infrastructure projects such as railway or highway tunnels. Overall this can be considered as an indicator of how information models in the domain of tunnelling gained importance during the recent years. For further analysis we relate the number of studies per year to the publication

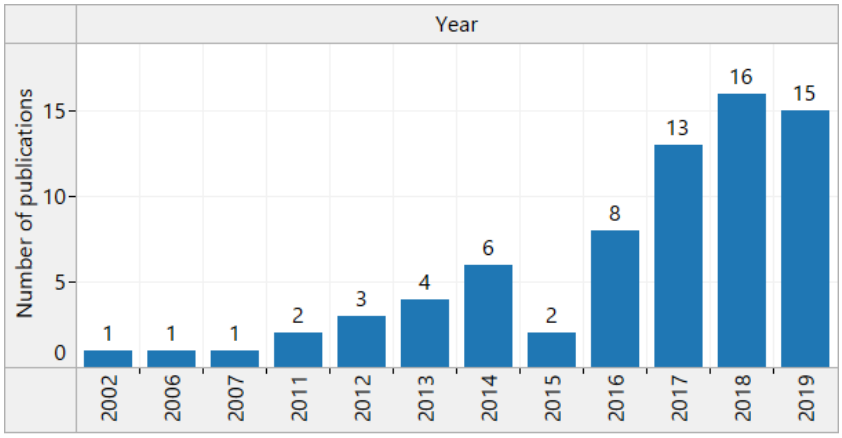


Figure 4.3: Number of publications per year for the years 2002-2019 (included number of studies: 72)

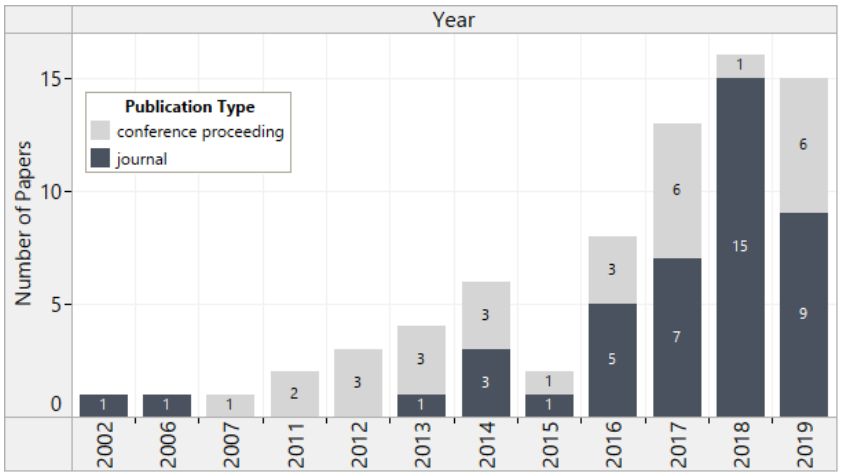


Figure 4.4: Number of publications per year regarding publication type (included number of studies: 72)

type in Figure 4.4. The bar chart is color coded where dark gray bars represent the number of published journal articles and light gray bars represent the number of published studies in conference proceedings during the respective year. The plot shows an overall growth of studies published in journals and a peak indicates that most journal papers were published in 2018. The number of studies published in conference proceedings, on the other hand, has a continuous growth phase in the period from 2011 to 2014, but is subject to fluctuation in the years 2015-2019. The number of published studies in conference proceedings reaches its peaks in 2017 and 2019. Overall 40% of the screened papers were published as part of conference proceedings and 60% as journal article.

Figure 4.5 relates the number of publications to the articles country of origin. The

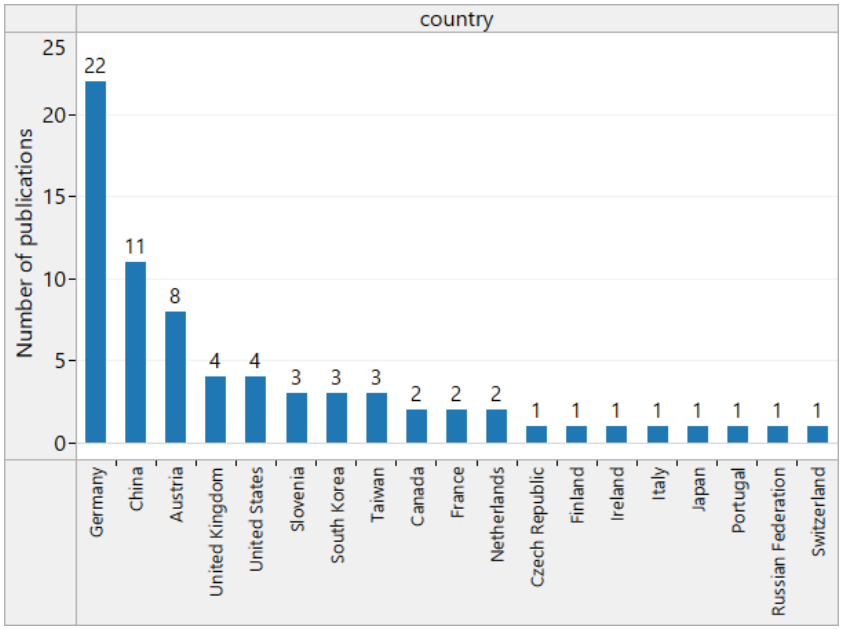


Figure 4.5: Number of publications per country (included number of studies: 72)

country of origin is determined by the affiliation of the studies primary author. The plot shows that a significant amount of 22 studies was written by authors with affiliations to German institutes. One reason for this observation may be the digitization strategies in the AECOO sector of the German government. Another driver of the increased research interest in the subsurface engineering domain may be the state-owned railway company Deutsche Bahn AG and DB Netz AG as major operator of railway infrastructure in Germany. The number of studies written by scientists with affiliations to institutes stationed in Germany continuously increased from a single study in 2012 to six publications in 2017 and 2019.
With eleven published studies China is the second most prominent country regarding the number of TIM publications. The amount of published studies increased from a single article in 2013 to four publications in the years 2018 and 2019. Similar to Germany, the main driver for Chinese scientists may be public funded subsurface infrastructure projects as well as the maintenance and operation of existing tunnel facilities.
According to the plot the country with the third most number of published studies is Austria. The first study with an Austrian affiliation was published in 2016 and the number of studies peaked to four in 2018. A total of eight TIM studies were published between 2016 and 2019. One reason for the high amount of publications with affiliation to Austria may be the numerous TIM pilot projects during the construction of railway and highway tunnels in Austria. Another reason may be the strong role of Austrian engineers in the domain of tunnelling, especially in area of the conventional tunnelling, e.g. NATM. According to Figure 4.5 all other countries of origin have at most 4 TIM related publications.

Table 4.1: Most Prominent Proceedings regarding the number of published studies with a minimum of 2 publications

Venue	Category	Number of Publications
ISARC (International Symposium on Automation and Robotics in Construction)	Automation	7
WTC (World Tunnel Congress)	Engineering	4
IOP Conference Series: Materials Science and Engineering	Engineering	2
Geotechnical Frontiers	Engineering	2
EG-ICE (International Workshop on Intelligent Computing in Engineering)	Automation	2

Moreover, we want to investigate the distribution of the published journal articles and studies published in conference proceedings related to the venue. Our motivation is to find out if there are a few venues with a high number of published studies, or if the number of publications distributes equally over the venues. Therefore, Table 4.1 lists workshops, conferences and congress with the total number of published TIM studies in proceedings. The underlying result has 17 different venues of conference proceedings with a total of 29 studies. For the sake of relevance and clarity, we only present conference proceedings with a minimum of 2 TIM related papers. The category column shows the main research area of the venues.

The *International Symposium on Automation and Robotics in Construction (ISARC)* is the most prominent conference with 7 submitted TIM publications from 2002 to 2019. One cause for ISARC being the most prominent conference may be the agenda of the hosting association, the *International Associtation for Automation and Robotics in Construction (IAARC)*. The main objective of the IAARC is to encourage and promote the technical development of Automation and Robotics in Construction. A primary topic of the conference are building information models for project collaboration and life cycle management where tunnel facilities are part of. Additionally, IAARC is the host of the interdisciplinary journal *Automation in Construction*.
Our statistic analysis shows an arithmetic mean value of 2 (1.7) publications per venue and a standard deviation of 1 (1.5). Hence, we identified a spread of 1 to 3 published studies per venue in this descriptive analysis. This qualifies the ISARC as statistical outlier compared to the average values of the other congresses, workshops and conferences. The second most prominent conference according to the number of published studies is the *World Tunnel Congress (WTC)* where 4 TIM papers were presented and submitted. The congress is, among others, one of the main conferences for presentations in the field

Table 4.2: Most Prominent Journals regarding the number of published studies with a minimum of 4 publications

Venue	Category	Number of Publications
Geomechanics and Tunnelling	Engineering	14
Tunnelling and underground space technology	Engineering	5
Bautechnik	Engineering	4
Automation in Construction	Automation	4

of tunnel projects and subsurface engineering. This world congress is hosted by the *International tunnelling and Underground Space Association (ITA)* which focuses on the use of the subsurface and promotes advances of all tunnel life cycle phases. The usage of information models in the domain of subsurface engineering aligns with the objectives of ITA as TIM has the potential to increase efficiency of all tunnel life cycle phases.
All other conference proceedings listed in Table 4.1 have at most two publications. The remaining conference proceedings *Geotechnical Frontiers*, *IOP Conference Series: Material Science and Engineering*, and *International Workshop on Intelligent Computing in Engineering (EG-ICE)* focus on different subjects. Though, all conferences do capture main topics of TIM such as the interaction of computing with engineering, tunnel modelling, tunnelling methods as well as in-situ investigations and characterizations.
After our analysis of the distribution of publications in conference proceedings we now identify the most prominent journals. Therefore, Table 4.2 shows the most prominent TIM journals according to the number of published studies and their main research categorization.

The most prominent venue with 14 published articles in the domain of TIM is *Geomechanics and Tunnelling*. The journal publishes German papers with English abstracts as well as bi-lingual papers where articles are written in both languages. This observation is not surprising due to correlation with the number of published studies with affiliations to the German speaking countries, shown in Figure 4.5. Geomechanics and Tunnelling is a journal rooted in Austria with a main emphasis on tunnel construction, geology engineering in practice as well as rock and soil mechanics.
The second most prominent journal is *Tunnelling and Underground Space Technology* with 5 articles published in the field TIM. Main topics of the journal are advances on methods and improvements of each phase in a tunnel life cycle. A reason for the number of publications may be that the journal is related to ITA and regularly features articles from ITA members.
The third and fourth most prominent venues are *Bautechnik* and *Automation in Construction* with 4 publications each in the time frame from 2002 to 2019. In this list the venue Automation in Construction is the only prominent journal focusing on automation

compared to the other prominent venues with an emphasis on subsurface and civil engineering. Bautechnik on the other hand is a special candidate as well. The venue publishes papers in German and in some cases with English translation. Moreover, the journal focuses on advances in the civil engineering sector. This may seem to be a contradiction, but the natural roots of building information models are located in the civil engineering sector and then started to influence the domain of infrastructure facility structures such as tunnels, railway, roads, and bridges.

RQ1—Main Findings: The scientific communities of the research subjects Engineering and Computer Science show an increased interest in information modelling in the tunnelling domain. This can be observed when analyzing the increasing number of publications in the observation period from 2002 to 2019 (see Fig. 4.3). Based on the most prominent venues it is noticeable that established journals and conference proceedings in the engineering domain have a higher number of publications in the domain of TIM. The influence of the criteria to exclude studies written in other languages than German or English and the usage of German and English keywords in the database search is noticeable, see Fig. 4.5. Therefore, the results concentrate on studies written by authors with affiliations to English or German speaking countries.

4.1.3 RQ2: excavation types & tunnel life cycle phases

After the analysis of bibliometric key facts this subsection aims to identify the most prominent excavation types and tunnel life cycle phases according to the number of TIM related publications. For the classification of excavation types and tunnel life cycle phases we used the criteria and classes described in chapter 3.2.4. We classified the studies in a two phase processes of initial classification followed by a review and an optional discussion with a domain expert to resolve conflicts. Figure 4.6 presents the result of this classification process. The plot shows that the planning & design phase of a tunnel project qualifies as the most prominent stage in the tunnel life cycle. With a total number of 41 published studies more than a half of all TIM publications focus on the planning phase. One reason for the increased research interest in the planning phase may be the potential of information models to increase the efficiency and influence all further stages of a tunnel life cycle. Hence, during this stage TIM principles may support the design & planning process by standardized information models to enable information exchange between the different stakeholders. The gathered and persisted information is fundamental for all subsequent tunnel life cycle phases. Moreover, more than a half (22) of all studies with an emphasis on the planning stage focus on tunnel design, simulations as well as the design and application of derived data models without specifying an excavation method. A possible reason for this may be that BIM principles are located on a higher level of abstraction. Whereas standardized data models, such as IFC, may go into detail on specific excavation types by providing appropriate data objects. Noteworthy is that 10 publications thematically focus on the operation phase, but do not present research specific to an excavation method. Moreover, no study has been identified discussing topics of the operation stage with relations to a specific

Excavation Method	Life Cycle Phase			
	plan	build	operate	all
continuous	16	7		1
conventional	3	3		1
unspecified	22	2	10	7

Figure 4.6: The number of published studies related to the project phase and the excavation type (included number of studies: 72)

excavation type. A reason for this observation may be that operational processes may be independent of the excavation method used during construction. When comparing the two excavation methods it becomes clear that the continuous excavation is dominant in terms of published studies. In total 16 of the 24 TBM related studies are thematically located in the planning phase whereas only 7 studies describe advances in the construction phase. A single paper presents research regarding all tunnel life cycle phases with focus on continuous tunnelling. The number of publications in the field of conventional tunnelling identified in this study is 7 where 3 of them primarily focus on topics relevant during the planning stage. Another 3 publications in the field of conventional excavation methods describe advances in the field of tunnel construction and a single paper presents research which is applicable in all tunnel life cycle stages. This indicates a stronger interest of the research community in TBM related topics in contrast to the research niche of conventional excavation.

RQ2—Main Findings: The identified research efforts of the scientific community show an increased interest in the planning stage when comparing all tunnel life cycle phases. This observation is based on an analysis of the distribution of published studies related to the life cycle phases of a tunnel shown in Figure 4.6. In regards to the different excavation methods the continuous excavation method and studies without emphasis on any specific excavation method are dominant in all tunnel life cycle phases. Especially the high number of publications which describe the adaption of BIM principles in specific areas of the tunnelling domain are standing out. This phenomena could be explained by the fact that BIM principles are located on a higher level of abstraction where fundamental digitization strategies for the facility life cycle phases are defined.

4.1.4 RQ3: research type facets

Based on the classification scheme of Wieringa et al. [WMMR06] we classified the selected studies in different type facets. For classification we applied the aforementioned two phased classification process with an optional discussion of conflicting classified studies with a domain expert. We assigned each study to a single class based on the results of

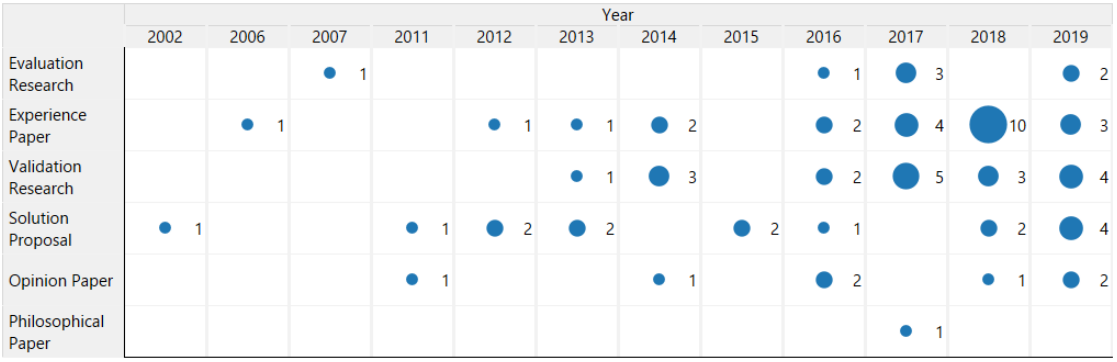


Figure 4.7: The number of published studies per year in relation to the Research Type Facets (included number of studies: 72)

the classification (see chapter 3.2.4) process and the optional discussion. This facet type classification allows us to analyze the research and publication practices of scientists in the field of TIM. Figure 4.7 shows the number of publications per year related to the type facet classes. The plot indicates an increased prominence of experience papers and validation research from 2012 onward. In total the result set comprises 1 philosophical paper, 7 opinion papers, 7 evaluation papers, and 15 solution proposals. Noteworthy is the single philosophical paper which presents a new framework to apply BIM in the design phase of a subsurface underground structure. Moreover, the 7 identified opinion papers provide recommendations how BIM should be applied to the subsurface domain or how a tunnel should be maintained and inspected. The seven identified evaluation studies evaluate simulation methods, BIM methodology use cases and IS solutions to specific problems in the tunnelling design, construction and operation. The majority of classified publications are validation research (18) and experience papers (24). This observations suggests that the majority of papers value personal experience, opinions and proof of concepts more than the use of sound research methods and scientific rigor.

We relate excavation methods and tunnel life cycle phases in order to show the interaction between all classes as result of the last activity of the systematic mapping process shown in Fig. 3.2. Figure 4.8 presents this systematic map of TIM related publications and the classified result set. It is worth noting that the most prominent facet type, experience papers, has just a single publication in the operation life cycle phase. A reason may be that this field is too young in order to gather experience of TIM application in the phase of tunnel operation and to present any lessons learned. This may be a fact as currently operated tunnels may not have a thorough digital representation and tunnel facilities which do apply BIM principles are currently planned or constructed.

Figure 4.8: The number of published studies in relation to Tunnel Life Cycle Phases, Excavation Method and Research Type Facet (included number of studies: 72)

RQ3—Main Findings: The analysis of the type facet classification indicates that scientists in the field of TIM tend to publish solution proposals, experience and opinion papers. Moreover, the low amount of experience papers in the operation phase of a tunnel could be explained by the fact that tunnels which apply BIM principles are currently planned or in construction. This observation also indicates that tunnel facilities in the operation phase adopt information modelling principles very slow if at all.

4.1.5 RQ4: search terms

We use the abstracts of publications in the result set in order to identify the most prominent words to characterize literature in the domain of tunnel information modelling. Therefore, Figure 4.9 shows a tag cloud of the most prominent used words in the selected studies abstracts. In order to create the tag cloud each word occurrence in the selected abstract is counted. Moreover, we ignored keywords like "tunnel", "building", "information", "modelling", and "bim" as these are part of the activity *Conducting Search* defined in chapter 3.2, Fig. 3.4. The visualization provides an overview of the 45 most used words with at least 23 occurrences in the abstracts. The most used keywords are *data*, *construction*, *project*, *system*, and *design*. The frequency of the words may indicate that TIM is most frequently used in design and construction phases of a tunnel project. This observation supports the results from RQ 2 in chapter 4.1.3. Moreover, the keywords data and system may be interpreted as foundation of every digitization process.

Figure 4.9: Tag cloud of the most prominent terms (created with http://tagcrowd.com, included number of studies: 72)

RQ4—Main Findings: It turns out that the most frequently used keywords are design and construction. Which indicates a prominence of those life cycle stages for research efforts. This observation supports the results of RQ2 in chapter 4.1.3. Moreover, the abbreviation TIM is not part of the resulting list of frequently used keywords. The analysis also suggests that the development of processes as well as systems for data collection and information exchange is a predominant topic in the research domain of tunnel information modelling.

4.1.6 Validity evaluation

Descriptive validity

According to Wolhin et al. [WRH+12] descriptive validity is the extent to which observations are described objectively and accurately. The main threats are subjective classifications of papers which may lead to inaccurate or biased results. The threat of subjective classification is an immanent part of all studies conducted by a single author. Our mitigation strategy is to involve domain experts in the activities *Screening of Papers*

and *Classification using Abstracts*. Hence, we received a test-set of studies from domain experts which must be covered by the result set. Therefore, we encourage further research activities in order to evaluate the findings and results of this study.
Additionally, Wohlin et al. [WRH+12] identify the risk of *publication bias* which occurs when certain study types are not submitted by their authors or systematically rejected by reviewers or editors. The risk is, that for example opinion studies are often rejected or unpublished and therefore the mapping study concludes that opinion papers are less frequent in the research domain. Our measure to minimize the risk of publication bias is to use different research publication databases with various scopes, e.g. Engineering and Information Systems.

Internal validity

According to Wohlin et al. [WRH+12] this aspect of validity is of interest when causal relationships are examined. Therefore, the objective is that the research methods used within the mapping study are causing the results of the survey. Also, Wohlin et al. [WRH+12] identified the selection of publications and the instrumentation caused by design of artifacts as influencing factors.

- *Publication selection:*
  - *Keywords:* we applied an iterative process to define groups of keywords and combinations in close cooperation with domain experts. The objective of the keyword groups is to identify relevant studies and describe publications of interest in the domain of tunnel information models. The resulting definition is shown in Chapter 3.2.2, Figure 3.3. We used those keyword sets combined with boolean operators to search abstracts and titles in several publication databases.
  - *Time frame:* we restricted the time period from 2002 to 2019, since our pilot searches found the first publication in regards to information model application in 2002 [EH02]. We excluded 2020 as it is the year of conducting the study and there may still be publications until the end of the year which would not be part of the mapping study and therefore be a threat for reproducibility.
  - *Literature repositories:* before conducting the mapping study we analyzed the review methodologies used by published studies in the domain of BIM, see Chapter 2.2. This set of papers consists of the reviews [VSS14], [BLLD16] and [WPL19] and has been delivered by domain experts. Hence, we identified that there are no well-established and regularly used literature repositories. Based on domain experts recommendations and results from pilot searches we took three different repositories into account: Scopus, Springer Link and Wiley.
  - *Publication language:* only studies with an abstract written in English or German were considered, even though the publication databases provided

search results containing the keywords in abstracts or titles written in another language or time frame.

– *Manual filtering:* Duplicates, theses, books, papers without available abstracts as well as publications without relation to BIM or TIM were removed from the result set.

- *Instrumentation:* publication databases may delay previously published studies (*timeliness*) or may deliver incomplete venues (*completeness*). Therefore, we developed a thorough mitigation strategy. We use (i) multiple keyword groups and combinations to identify relevant studies, (ii) a specific time frame from 2002 to 2019, and (iii) three different publication databases of different research domains.

Moreover, another important threat to validity is researcher bias which may occur during the activities *Screening of Papers* and *Classification using Abstracts* as shown in Chapter 3.2 Fig. 3.2. As elaborated before, the screening of papers and other activities were conducted by the author of the thesis which is the main threat to validity. Additionally, another major issue is that important and prominent papers from the domain of Tunnel Information Modelling (TIM) may not have been identified by the aforementioned activity. Our mitigation strategy to overcome this threat was the involvement of domain experts especially to define keywords which identify relevant studies. The domain experts and the co-advisor provided a test-set of relevant papers which may be covered by the result set of the screening of papers activity. After we identified relevant papers by screening papers we used this test-set to evaluate if our result set contained all relevant studies and therefore was successful. Also, researcher bias is a major threat during the classification using abstracts activity. In contrast to other mapping studies, e.g. Petersen et al. [PVK15] or Wolny et al. [WMC+20], this study is conducted by a single author. There is no second author or team to support the mapping process including the classification of papers and as a consequence this threat is immanent to theses with a single author. Our mitigation strategy for the researcher bias during the classification activity was to discuss inconclusive classifications with domain experts.

Generalizability

Wohlin et al. [WRH+12] describe generalizability or external validity with the question, if the findings of the study can be generalized outside the scope of the study. The findings of this mapping study is not generalizable outside of the scope of the study. This is due to the very specific research domain, the start and end of the examination period as well as the thorough definition of relevant literature by using keyword groups. Therefore, our mapping study focuses on research in the domain of tunnel information models and may not be generalized to any closely related scientific field.

Interpretive validity

According to Petersen et al. [PVK15] interpretive validity means, that conclusions and reasons drawn from the data are reasonable. A major threat in interpreting the data is researcher bias and may be encountered during the *Mapping of Papers* activity. Our mitigation strategy for this threat is the review of the thesis through the advisor, co-advisor and reviewer.

Repeatability

We documented the research process from the definition of the research questions to the mapping of papers thoroughly in order to ensure that our applied processes and research results are reproducible. We apply well-established guidelines to conduct the mapping study, elaborate on decisions and actions during the mapping process and describe mitigation strategies in order to reduce threats to validity.

4.2 Design Science Artifact

In this section we analyze and measure the performance of the data science models and evaluate the classifications based on the results of the mapping study in order to answer RQ5. According to Chapter 2.3.2, Figure 2.2, this section represents the *Evaluation* sub-process of the CRISP-DM data analysis process model. Hence, the evaluation of the data mining artifact results is based on the *Data Preparation*, the *Data Science Models* and the proposed *Test Design*.

4.2.1 RQ5: Evaluation of Study Classification

Table 4.3 shows the different classification models built for the binary classification task to classify English studies which are relevant in the domain of Tunnel Information Modelling. In order to measure the predicted results of each classifier we use the propose metrics described in Chapter 3.3.5. In detail, the prediction performance of the classifiers is measured using the macro-average of precision (π), recall (ρ) and F_1 during a 10-fold cross-validation. Additionally to the macro-average values the standard deviation (σ) is provided. The Support Vector Classifier (SVC) reaches the highest value of precision with a total average of 0.82 and a standard deviation of 0.11. Resulting in a higher spread of the macro-averaged precision value when compared to the results of other classifiers. In regards to the highest macro-averaged recall value the LR peaks to 0.81 with a standard deviation of 0.1. In terms of the macro-averaged F_1^M value the SVC reaches the high point of 0.78 and a standard deviation of 0.09. The low standard deviation in the F_1^M value of the Gradient Boosting Tree classifier is the result of accurate recall and precision values. In regards to the MCC value the classifier predictions are positive which indicates a positive agreement between prediction and the classification of the mapping study. Moreover, the averaged MCC value shows a minor increased standard deviation for

Table 4.3: Averaged 10-fold cross-validation results by classifier

Classifier	Averaged measures			
	π^M	ρ^M	F_1^M	MCC
Logistic Regression	0.76, $\sigma = 0.06$	0.81, $\sigma = 0.1$	0.76, $\sigma = 0.08$	0.57, $\sigma = 0.15$
Support Vector Classifier	0.82, $\sigma = 0.11$	0.79, $\sigma = 0.09$	0.78, $\sigma = 0.09$	0.60, $\sigma = 0.17$
Gradient Boosting Tree	0.68, $\sigma = 0.08$	0.75, $\sigma = 0.1$	0.67, $\sigma = 0.01$	0.42, $\sigma = 0.17$

Table 4.4: Results of the pairwise McNemar's χ^2 evaluation of classifiers

$H0$	χ^2
$LR = SVC$	0
$LR = GBTree$	0.5
$SVC = GBTree$	0.52

all three classifiers, indicating a larger spread. Overall the classifiers reach a positive classification performance according to the macro-averaged metrics as well as the MCC.

The measurements indicate that the classifiers in average do perform equally with minor deviations. In order to find a better or worse classifier we conduct a McNamara's test with the null hypothesis, that the results of the two classifiers are equal. The results presented in Table 4.4 show, that the null hypothesis is not rejected as the differences are not statistically significant in regards to a probability of 5%. Both, the classifier measurements as well as the statistical null hypothesis test, show no significant difference. Therefore, all three classifiers qualify for the binary classification to identify relevant studies in the TIM domain.

Due to the low number of six German written studies it is not effective to train and test a German-specific model in a satisfying manner.

The upcoming evaluation provides detailed information for the selection of search terms in order to classify studies in regards to excavation types and tunnel life cycle phases. Therefore, we start by using the keywords identified by domain experts to identify studies in the field of continuous and conventional tunnelling. In order to identify relevant studies in the TIM domain, we combine these search terms with keywords received by the analysis of TIM-related studies, i.e. *BIM*, *Information Model*, see Chapter 3.2.2. Table 4.5 shows a binary classification using search term combinations, the confusion matrix and the resulting quality measurements. The confusion matrix and the quality metrics π, ρ, F_1 and MCC are defined in Chapter 3.3.5. The measurements of the keywords used to classify studies in regards to conventional tunnelling indicate a performance comparable to random selection. Moreover, the combination $TBM+BIM$ shows a low to medium classification performance. Other keywords used to identify studies with relations to mechanized tunnelling do not reach a comparable level. Especially the amount of successfully identified studies in relation to the number of relevant documents, the recall

Table 4.5: Evaluation of search terms to classify excavation types

Keyword combination	TP	FP	TN	FN	π	ρ	F_1	MCC
$TBM + BIM$	6	2	355	18	0.75	0.25	0.375	0.41
$mechanized + BIM$	4	1	356	20	0.8	0.16	0.27	0.34
$continu + IM$	1	0	357	23	1.0	0.04	0.08	0.19
$NATM + BIM$	0	2	355	24	0.0	0.0	0.0	-0.01
$NATM + IM$	0	1	356	24	0.0	0.0	0.0	-0.01
$austrian + BIM$	0	4	353	24	0.0	0.0	0.0	-0.02
$austrian + IM$	1	2	355	23	0.33	0.04	0.07	0.09
$conventional + IM$	0	4	353	24	0.0	0.0	0.0	-0.02
$conventional + BIM$	0	5	352	24	0.0	0.0	0.0	-0.02

Table 4.6: Evaluation of terms to classify project life cycle phases

Keyword combination	TP	FP	TN	FN	π	ρ	F_1	MCC
plan + BIM	4	27	330	20	0.12	0.16	0.14	0.08
design + IM	11	33	324	13	0.25	0.45	0.32	0.27
construct + IM	13	41	316	11	0.24	0.54	0.33	0.29
construct + BIM	11	55	302	13	0.16	0.45	0.24	0.19
mainten + BIM	0	8	349	24	0.0	0.0	0.0	-0.03
operat + IM	2	10	347	22	0.16	0.08	0.11	0.07

value, is very low in this category. Overall the keyword-based identification of studies, relating to a specific excavation type, shows low performance measures. Therefore, we conclude that the multi-class classification using the derived keywords does not qualify for model development.

We continue the evaluation of identifying keywords for the classification of studies in regards to the project life cycle phases of a tunnelling project. Therefore, our evaluation of descriptive search terms resulted in the keywords (i) *plan & design*, (ii) *construct* and (iii) *mainten & operat*. For the planning phase we compare the number of identified studies of the search terms *design* and *plan* in the abstracts of the DM artifact to the labeled data of the SMS. The classification results of the planning phase show a low precision, indicating an increased amount of false-positives. The MCC score of 0.27 shows a low positive classification result of the DM artifact. Moreover, the construction focused search terms show a similar pattern with minor increased MCC values and a low precision score. In detail the MCC value of 0.3 indicates a minor positive classification. The negative MCC value as well as the low value of the other metrics of the *maintenance*

phase show that the classification using the selected keywords is not effective. Therefore, our conclusion regarding the project phase classification using the proposed keywords shows low performance measures. Whereas studies which are relevant to the maintenance and operation phase are classified as good as random sampling.

RQ5—Main Findings: Several well established classification models show good performance by identifying relevant studies in the Tunnel Information Modelling domain. Moreover, the metrics and hypothesis tests of the models show an equal and, therefore, no statistical significant difference in classification performance. In regards to the classification of papers written in German it is not effective to train and test a model due to the low amount of available German studies in the field of TIM. The multi-class classifications of studies in classes relating to tunnel life cycle phases and tunnel excavation methods show a low classification performance. Especially the classification results of the classes *conventional tunnelling* and the *tunnel operation phase* do not differ from random selection. A low study identification quality is directly connected to the performance of the study ranking and ultimately results in classification models with equally low performance metrics.

CHAPTER 5

Discussion

In this study, we first conducted a mapping study in the research domain TIM and then designed a data science artifact to identify and classify relevant studies. Accordingly, we wanted to achieve two objectives: (i) provide an overview of research activity in the domain of applied information models in subsurface engineering and (ii) automatize the process of literature identification and classification to provide state-of-the-art research works.

By realizing the first objective, we identified a constant increasing interest in the research area since 2011 only interrupted in 2015. We also recognized, that the most prominent conference, according to the number of published studies, is focusing on automation and robotics while the most prominent journals are thematically positioned in the domain of construction and engineering. Additionally, we observe that governmental digitization strategies and, therefore, research funding is a primary driver to increase the number of published, TIM-related studies. A prominent example for this observation is the number of studies published by authors with affiliations to Germany, China and Austria. It can be summarized, that the design and planning phase as well as the continuous excavation are the most prominent classes in regards to life cycle phases and excavation methods. Moreover, the relative high amount of studies not specifying an excavation method can be explained through the higher abstraction level of BIM strategies and principles. Regarding the research type facets, we see a trend towards validation research, solution and experience papers focusing on the planning phase of a tunnel facility. The low amount of studies in the field of applying BIM during operation of a tunnel can be explained by the fact, that most of the tunnel facilities which apply BIM principles are currently designed or in construction. Finally, the mapping study identified, that the dominant topics in the research area are related to processes and systems for data collection and information exchange.

On accomplishing the second objective, we applied the well established CRISP-DM data analysis process to develop a data science software artifact. The results show, that common

text classification methods yield positive performance measures by identifying relevant studies in the domain of Tunnel Information Modelling. In detail, Logistic Regression, Support Vector Classifier and the boosting tree model have an equal classification performance. The classification of studies based on excavation type and tunnel life cycle phase shows low performance metrics, due to the inability to classify relevant studies. A major reason for this is the keyword-based identification approach which is not able to extract the fine granular context, e.g. life cycle phase and excavation method, of a study. Based on the results of the data science artifact and main findings of the mapping study we identified the following research directions for future work.

Research direction 1: Tunnel Operation

The mapping study results show, that the operation phase of a tunnel facility is underrepresented in terms of published studies. Therefore, a future research direction is as-build Tunnel Information Modellings in order to create digital representations based on current physical conditions of the tunnel facility. These representations are fundamental for operation and maintenance and enable a historical documentation of the tunnels condition.

Research direction 2: Continuous Tunnelling

The amount of studies in the field of conventional tunnelling is significantly low compared to the continuous excavation method. We identify the digitization of the conventional tunnelling as the primary challenge for future research work. A major field of future research is the analysis and prediction of the rock face and therefore the use of geologic information systems in the design and construction phase.

Research direction 3: Feature construction

The chosen vector-based document representation of our study eliminates context during the data cleaning to reduce the complexity of the corpus and extract the essence of each abstract. Therefore, it is well known that document representation is a vital issue in the area of text classification. A different approach for document representation is to use a word embedding in combination with a vectorized representation. Examples to learn an embedding from a corpus are word2vec and extensions such as GloVe or fastText. This representation may be used to train a Recurrent Neural Network for text classification.

Tunnel Information Modelling Papers identified in the Mapping Study

[S1] A. Alsahly, V.E. Gall, A. Marwan, J. Ninić, G. Meschke, A. Vonthron, and M. König. From building information modeling to real-time simulation in mechanized tunneling: An integrated approach applied to the wehrhahn-line düsseldorf. In *ITA-AITES World Tunnel Congress 2016, WTC 2016*, volume 4, pages 3178–3186. Society for Mining, Metallurgy and Exploration, 2016.

[S2] N. Baum, S. Boxheimer, D. Krause, F. Renz, B. Hoffmann, J. Wächter, and T. Klingeberg. Drill & blast and special foundation in gothenburg. *Bautechnik*, 96(7):549–557, 2019.

[S3] R. Bernard, J. Pacovský, and I. Zemánek. Geo - monitoring performed during the construction of the valik highway tunnels. *Tunnelling and Underground Space Technology*, 21(3-4):226–227, 2006.

[S4] A. Borrmann, T.H. Kolbe, A. Donaubauer, H. Steuer, J.R. Jubierre, and M. Flurl. Multi-scale geometric-semantic modeling of shield tunnels for gis and bim applications. *Computer-Aided Civil and Infrastructure Engineering*, 30(4):263–281, 2015.

[S5] H. G. Bui, A. Alsahly, J. Ninic, and G. Meschke. Bim-based model generation and high performance simulation of soil-structure interaction in mechanized tunnelling. *Civil-Comp Proceedings*, 111, 2017.

[S6] J. Cesnik, M. Zibert, M. Lah, and M. Skalja. Required model content and information workflows enabling proficient bim usage. In *IOP Conference Series: Materials Science and Engineering*, volume 603. Institute of Physics Publishing, 2019.

[S7] Y. J. Cheng, W. G. Qiu, and D. Y. Duan. Automatic creation of as-is building information model from single-track railway tunnel point clouds. *Automation in Construction*, 106, 2019.

[S8] D. Cho, N.-S. Cho, H. Cho, and K.-I. Kang. Parametric modelling based approach for efficient quantity takeoff of natm-tunnels. In *2012 Proceedings of the 29th International Symposium of Automation and Robotics in Construction, ISARC 2012*, 2012.

[S9] J. Daller, M. Žibert, C. Exinger, and M. Lah. Implementation of bim in the tunnel design – engineering consultant's aspect. *Geomechanik und Tunnelbau*, 9(6):674–683, 2016.

[S10] Christoph Deporta, Joachim Wondre, and Alexander Zöhrer. Modern risk management applied to a large infrastructure site. *ce/papers*, 2(2-3):951–956, 2018.

[S11] B. Du, Y. Du, F. Xu, and P. He. Conception and exploration of using data as a service in tunnel construction with the natm. *Engineering*, 4(1):123–130, 2018.

[S12] J. Du, R. He, and V. Sugumaran. Clustering and ontology-based information integration framework for surface subsidence risk mitigation in underground tunnels. *Cluster Computing*, 19(4):2001–2014, 2016.

[S13] A. S. Elkadi and M. Huisman. 3d-gsis geotechnical modelling of tunnel intersection in soft ground: The second heinenoord tunnel, netherlands. *Tunnelling and Underground Space Technology*, 17(4):363–369, 2002.

[S14] C. Exinger, G. Mulitzer, R. Felsner, J. Lemmerer, R. Matt, and E. Griesser. Bim pilot project granitztal tunnel chain – development of data structures for tunnel structure and track superstructure. *Geomechanik und Tunnelbau*, 11(4):348–356, 2018.

[S15] J. Fernandes, M.L. Tender, and J.P. Couto. Using bim for risk management on a construction site. In *Occupational Safety and Hygiene V - Proceedings of the International Symposium on Occupational Safety and Hygiene, SHO 2017*, pages 269–272. CRC Press/Balkema, 2017.

[S16] Matthias Flora, Georg Fröch, and Werner Gächter. Optimierung des baumanagements im untertagebau mittels digitaler infrastruktur-informationsmodelle. *Bautechnik*, 2020.

[S17] G. Fröch, M. Flora, W. Gächter, F. Harpf, and A. Tautschnig. Application possibilities of a digital ground model in tunnel construction. *Bautechnik*, 96(12):885–895, 2019.

[S18] Stephan Frodl and Peter-Michael Mayer. Bim modelling in mined tunnelling – discussions and recommendation for the structural geometry to be considered in 3d models. *Geomechanics and Tunnelling*, 11(4):357–365, 2018.

[S19] M. Ghaznavi and S. Abourizk. An integrated multi-dimensional information model framework for tunneling projects using ifc data model. In *Proceedings, Annual*

Conference - Canadian Society for Civil Engineering, volume 1, pages 908–919. Canadian Society for Civil Engineering, 2013.

[S20] Gerald Goger and Tobias Bisenberger. Tunnelling 4.0 – construction-related future trends. *Geomechanics and Tunnelling*, 11(6):710–721, 2018.

[S21] C. Gruber, T. Weiner, and R. Zuchtriegel. Bim for tunnelling for a company – approaches and strategies. *Geomechanik und Tunnelbau*, 11(4):366–373, 2018.

[S22] R. Gueulet and L. Milesy. A 4d visualization tool for tbm worksites using cap: Integration of 3d models and real-time modeling thanks to database connections. In *ISARC 2018 - 35th International Symposium on Automation and Robotics in Construction and International AEC/FM Hackathon: The Future of Building Things*. International Association for Automation and Robotics in Construction I.A.A.R.C, 2018.

[S23] T. Hartmann, R. Amor, and E. W. East. Information model purposes in building and facility design. *Journal of Computing in Civil Engineering*, 31(6), 2017.

[S24] F. Hegemann, K. Lehner, and M. König. Ifc-based product modeling for tunnel boring machines. In *eWork and eBusiness in Architecture, Engineering and Construction - Proceedings of the European Conference on Product and Process Modelling 2012, ECPPM 2012*, pages 289–296. CRC Press, 2012.

[S25] R. Heikkilä, A. Kaaranka, and T. Makkonen. Information modelling based tunnel design and construction process. In *31st International Symposium on Automation and Robotics in Construction and Mining, ISARC 2014 - Proceedings*, pages 672–675. University of Technology Sydney, 2014.

[S26] M. Hu and Z. Huang. Ontology-driven tunnel construction information retrieval and extraction. In *26th Chinese Control and Decision Conference, CCDC 2014*, pages 4741–4746. IEEE Computer Society, 2014.

[S27] M. Hu and Y. Liu. E-maintenance platform design for public infrastructure based on ifc and semantic web technologies. *Advances in Intelligent Systems and Computing*, 842:517–526, 2019.

[S28] M. Hu, Y. Liu, V. Sugumaran, B. Liu, and J. Du. Automated structural defects diagnosis in underground transportation tunnels using semantic technologies. *Automation in Construction*, 107, 2019.

[S29] Min Hu and Yunru Liu. E-maintenance platform design for public infrastructure maintenance based on ifc ontology and semantic web services. *Concurrency and Computation: Practice and Experience*, 32(6):e5204, 2020.

[S30] M. Žibert, S. Keniston, and J. Karlovšek. Tunnel information modelling in most recent form: Applying bim technologies and procedures in tunnelling environment.

In *ITA-AITES World Tunnel Congress 2016, WTC 2016*, volume 1, pages 1–10. Society for Mining, Metallurgy and Exploration, 2016.

[S31] M. Žibert, M. Lah, and S. Saje. Challenges and opportunities in implementing bim methodology in tunnelling. *Geomechanik und Tunnelbau*, 11(4):335–339, 2018.

[S32] M. König, T. Rahm, F. Nagel, and L. Speier. Building information modeling in tunneling – digital design and construction of tunneling projects. *Bautechnik*, 94(4):227–231, 2017.

[S33] C. Koch, A. Vonthron, and M. König. A tunnel information modelling framework to support management, simulations and visualisations in mechanised tunnelling projects. *Automation in Construction*, 83:78–90, 2017.

[S34] B. Kohlböck, E. Griesser, S. Hillisch, H. Birgmann, and A. Fasching. The bim pilot project köstendorf – salzburg. *Geomechanik und Tunnelbau*, 11(4):325–334, 2018.

[S35] H. Kontrus and M. Mett. High-speed 3d tunnel inspection. In *Proceedings - Rapid Excavation and Tunneling Conference*, volume 2019-June, pages 809–816. Society for Mining, Metallurgy and Exploration, 2019.

[S36] P. C. Lee, Y. Wang, T. P. Lo, and D. Long. An integrated system framework of building information modelling and geographical information system for utility tunnel maintenance management. *Tunnelling and Underground Space Technology*, 79:263–273, 2018.

[S37] S.-H. Lee and B.-G. Kim. Ifc extension for road structures and digital modeling. In *Procedia Engineering*, volume 14, pages 1037–1042, 2011.

[S38] Sang-Ho Lee, Sang I. Park, and Junwon Park. Development of an ifc-based data schema for the design information representation of the natm tunnel. *KSCE Journal of Civil Engineering*, 20(6):2112–2123, 2016.

[S39] R. Lensing. Enrichment of bim with construction process data in mechanized tunnel construction. *gis.Science - Die Zeitschrift fur Geoinformatik*, 3:110–117, 2018.

[S40] X. Li and H. Zhu. Development of a web-based information system for shield tunnel construction projects. *Tunnelling and Underground Space Technology*, 37:146–156, 2013.

[S41] Y. Loo, M. Sykes, C. Sturzaker, J. Osborne, and C. Cook. Early-stage bim for cern's future circular collider studies. *Structural Engineer*, 93(7):12–18, 2015.

[S42] S.-R. Lu, I.-C. Wu, and B.-C.B. Hsiung. Applying building information modelling in environmental impact assessment for urban deep excavation projects. In *2012 Proceedings of the 29th International Symposium of Automation and Robotics in Construction, ISARC 2012*, 2012.

[S43] S. Mahdi, F.-Z. Houmymid, and E. Chiriotti. Use of numerical modelling and gis to analyse and share the risks related to urban tunnelling: Greater paris - red line - south section. In *ITA-AITES World Tunnel Congress 2016, WTC 2016*, volume 3, pages 2187–2197. Society for Mining, Metallurgy and Exploration, 2016.

[S44] U. Maidl and J. Stascheit. Real time process controlling for epb shields / echtzeit-prozesscontrolling bei erddruckschilden. *Geomechanik und Tunnelbau*, 7(1):64–71, 2014.

[S45] S. Mao, J.-L. Lebrun, O. Doukari, R. Aguejdad, and Y. Yuan. 3d bim multi-scale modeling for a tunnel construction project [modélisation 3d bim multi-échelle d'un projet btp tunnel]. In *CEUR Workshop Proceedings*, volume 1535, pages 135–149. CEUR-WS, 2015.

[S46] Peter-Michael Mayer, Stephan Frodl, and Felix Hegemann. Bim as a process in tunnelling / bim als prozess im tunnelbau. *Geomechanics and Tunnelling*, 9(6):684–695, 2016.

[S47] G. Morin, S.L. Deaton, R. Chandler, and S. Miles. Silvertown tunnel, london, england - a case study applying bim principles to the geotechnical process. In *Geotechnical Special Publication*, number GSP 277, pages 587–595. American Society of Civil Engineers (ASCE), 2017.

[S48] Z. Ning, L. Galisson, and P. Smith. Case study: Geotechnical instrumentation and monitoring of alaskan way viaduct replacement project. In *Geotechnical Special Publication*, volume 2019-March, pages 277–286. American Society of Civil Engineers (ASCE), 2019.

[S49] J. Ninić, H. G. Bui, C. Koch, and G. Meschke. Computationally efficient simulation in urban mechanized tunneling based on multilevel bim models. *Journal of Computing in Civil Engineering*, 33(3), 2019.

[S50] J. Ninić, C. Koch, and J. Stascheit. An integrated platform for design and numerical analysis of shield tunnelling processes on different levels of detail. *Advances in Engineering Software*, 112:165–179, 2017.

[S51] J. Ninic, H.G. Bui, and G. Meschke. Parametric design and isogeometric analysis of tunnel linings within the building information modelling framework. In *CEUR Workshop Proceedings*, volume 2394. CEUR-WS, 2019.

[S52] J. Ninic, C. Koch, and W. Tizani. Parametric information modelling of mechanised tunnelling projects for multi-level decision support. In *Digital Proceedings of the 24th EG-ICE International Workshop on Intelligent Computing in Engineering 2017*, pages 228–238. European Group for Intelligent Computing in Engineering (EG-ICE), 2017.

[S53] A. Osello, N. Rapetti, and F. Semeraro. Bim methodology approach to infrastructure design: Case study of paniga tunnel. In *IOP Conference Series: Materials Science and Engineering*, volume 245. Institute of Physics Publishing, 2017.

[S54] M. Pleshko, I. Voinov, and A. Revyakin. Assessment of the impact of railway tunnel lining defects with a long working lifespan on its carrying capacity. In *MATEC Web of Conferences*, volume 106. EDP Sciences, 2017.

[S55] S. Providakis, C. D. F. Rogers, and D. N. Chapman. Predictions of settlement risk induced by tunnelling using bim and 3d visualization tools. *Tunnelling and Underground Space Technology*, 92, 2019.

[S56] S. Schindler, F. Hegemann, A. Alsahly, T. Barciaga, M. Galli, K. Lehner, and C. Koch. An interaction platform for mechanized tunnelling. application on the wehrhahn-line in düsseldorf (germany) / eine interaktionsplattform für maschinelle tunnelvortriebe. anwendung am beispiel der wehrhahn-linie in düsseldorf: Application on the wehrhahn-line in düsseldorf (germany) / anwendung am beispiel der wehrhahn-linie in düsseldorf. *Geomechanik und Tunnelbau*, 7(1):72–86, 2014.

[S57] Jürgen Schwarz. Risikomanagement im tunnelbau - steuerung von kosten und terminen mit einem baubetrieblichen referenzmodell über den gesamten lebenszyklus. *ce/papers*, 3(2):27–32, 2019.

[S58] J. Sherry. Digital engineering enables multinational input on bergen's light rail extension, norway. *Proceedings of the Institution of Civil Engineers: Civil Engineering*, 171(5):49–56, 2018.

[S59] J. Stascheit, J. Ninić, G. Meschke, F. Hegemann, and U. Maidl. Building information modelling in mechanised shield tunnelling – a practitioner's outlook to the near future. *Geomechanik und Tunnelbau*, 11(1):34–49, 2018.

[S60] M. Stelzer, N. Radoncic, P. L. Iserte Llacer, A. Tatar, and M. Holmberg. Bim processes and workflows using the example of the subway extension in stockholm. *Geomechanik und Tunnelbau*, 11(4):340–347, 2018.

[S61] V. Vierhub-Lorenz, K. Predehl, S. Wolf, C.S. Werner, F. Kühnemann, and A. Reiterer. A multispectral tunnel inspection system for simultaneous moisture and shape detection. In *Proceedings of SPIE - The International Society for Optical Engineering*, volume 11157. SPIE, 2019.

[S62] S. Vilgertshofer, J. Amann, B. Willenborg, A. Borrmann, and T.H. Kolbe. Linking bim and gis models in infrastructure by example of ifc and citygml. In *Congress on Computing in Civil Engineering, Proceedings*, volume 0, pages 133–140. American Society of Civil Engineers (ASCE), 2017.

[S63] N. Vossebeld and T. Hartmann. Supporting tunnel safety assessment with an information model. In *Computing in Civil and Building Engineering - Proceedings of*

the 2014 International Conference on Computing in Civil and Building Engineering, pages 57–64. American Society of Civil Engineers (ASCE), 2014.

[S64] F. P. Weichenberger and G. Pischinger. Geological documentation – conditions, status quo and future development. *Geomechanik und Tunnelbau*, 10(5):567–573, 2017.

[S65] X. Wu, M. Lu, S. Mao, and X. Shen. Real-time as-built tunnel product modeling and visualization by tracking tunnel boring machines. In *ISARC 2013 - 30th International Symposium on Automation and Robotics in Construction and Mining, Held in Conjunction with the 23rd World Mining Congress*, pages 857–865. Canadian Institute of Mining, Metallurgy and Petroleum, 2013.

[S66] N. Yabuki, T. Aruga, and H. Furuya. Development and application of a product model for shield tunnels. In *ISARC 2013 - 30th International Symposium on Automation and Robotics in Construction and Mining, Held in Conjunction with the 23rd World Mining Congress*, pages 435–447. Canadian Institute of Mining, Metallurgy and Petroleum, 2013.

[S67] C.-I. Yen, J.-H. Chen, and P.-F. Huang. The study of bim-based mrt structural inspection system. In *Proceedings of the 28th International Symposium on Automation and Robotics in Construction, ISARC 2011*, pages 130–135, 2011.

[S68] C.W. Yu and J.C. Chern. Expert system for d&b tunnel construction. In *"Proceedings of the 33rd ITA-AITES World Tunnel Congress - Underground Space - The 4th Dimension of Metropolises"*, volume 1, pages 799–803, 2007.

[S69] L. Zhang, X. Wu, L. Ding, M. J. Skibniewski, and Y. Lu. Bim-based risk identification system in tunnel construction. *Journal of Civil Engineering and Management*, 22(4):529–539, 2016.

[S70] Chenyang Zhao, Arash Alimardani Lavasan, and Tom Schanz. Application of submodeling technique in numerical modeling of mechanized tunnel excavation. *International Journal of Civil Engineering*, 17(1):75–89, 2019.

[S71] Y. Zhou, Y. Wang, L. Ding, and P. E. D. Love. Utilizing ifc for shield segment assembly in underground tunneling. *Automation in Construction*, 93:178–191, 2018.

[S72] G. Zwittnig. From the lavanttal into the jauntal – project implementation and first experience of design and build 4.0: Vom lavanttal ins jauntal – projektumsetzung und erste erfahrungen zum planen und bauen 4.0. *Geomechanik und Tunnelbau*, 10(6):722–729, 2017.

List of Figures

List of Tables

List of Algorithms

Acronyms

AECOO Architecture, Engineering, Construction, Owner and Operator. 1, 13, 65

API Application Programming Interface. 14, 25–28, 30, 33, 34, 36, 40–42, 46, 48, 89

BIM Building Information Modelling. xi, xiii, 1, 3, 5–7, 16, 20, 21, 32, 38, 45, 49, 50, 54–56, 68–71, 73, 74, 79

BM25 Best Match 25. 57

CRISP-DM Cross Industry Standard Process for Data Mining. xi, xiii, 9, 10, 13, 14, 23, 26, 30, 32, 33, 40, 42, 45, 50, 52, 75, 79

CSV Comma-separated values. 35–37, 40, 89

DM Data Mining. xi, xiii, 3, 23, 54–56, 58, 77

DOI Digital Object Identifier. 14, 25, 26, 29, 33–38, 40–42, 47, 51–53, 62, 93

EBSE Evidence Based Software Engineering. 5

EG-ICE International Workshop on Intelligent Computing in Engineering. 67

GIS Geographical Information System. 7

HTTP Hypertext Transfer Protocol. 35, 41, 46

IAARC International Associtation for Automation and Robotics in Construction. 66

IFC Industry Foundation Classes. 32, 68

IR Information Retrieval. 57

IS Information Systems. xiii, 1, 2, 5, 13–15, 24, 57, 70

ISARC International Symposium on Automation and Robotics in Construction. 66

ITA International tunnelling and Underground Space Association. 67

JSON JavaScript Object Notation. 36, 42

KDD Knowledge Discovery in Database. 8–10, 89

LR Logistic Regression. 57, 60, 75, 80

MCC Matthews Correlation Coefficient. 59, 60

N/A Not Available. 33, 34, 38–40, 42, 44–46, 48, 50–53, 89

NATM New Austrian tunnelling Method. 22, 65

RQ Research Questions. 15, 16, 24, 62, 75

SE Software Engineering. 5

SLR Systematic Literature Review. 5, 6

SMS Systematic Mapping Study. xi, xiii, 1–3, 15, 24, 35, 54–56, 77

SRU Search/Retrieval via URL. 46

SVC Support Vector Classifier. 75, 80

SVM Support Vector Machine. 57, 60

TBM Tunnel Boring Machine. 20, 24, 31, 37, 45, 49, 50, 69

TIM Tunnel Information Modelling. xi–xiii, 1, 2, 5, 13, 15, 16, 21, 24, 30, 31, 37, 44, 48–50, 52, 54, 56–58, 61–63, 65–68, 70–72, 74–76, 78–80

TU Technische Universität. 26

URL Uniform Resource Locator. 62

VPN Virtual Private Network. 26

WTC World Tunnel Congress. 66

XML Extensible Markup Language. 36, 46, 47

References

[AG21] Springer Nature Switzerland AG. SpringerNature meta api. http://api.springernature.com/meta/v2/json, 2021. Accessed: 2021-01-20.

[AMB14] Dietmar Adam, Roman Markiewicz, and Markus Brunner. Block-in-Matrix Structure and Creeping Slope: Tunneling in Hard Soil and/or Weak Rock. *Geotechnical and Geological Engineering*, 32(6):1467–1476, December 2014.

[Bil12] Bilbao metro line two gets support. *Tunnels and Tunnelling International*, (SEP):31–34, 2012.

[BKD+15] A. Borrmann, T. H. Kolbe, A. Donaubauer, H. Steuer, J. R. Jubierre, and M. Flurl. Multi-Scale Geometric-Semantic Modeling of Shield Tunnels for GIS and BIM Applications. *Computer-Aided Civil and Infrastructure Engineering*, 30(4):263–281, 2015.

[Bla02] Blast design using measurement while drilling parameters. *Coal International Mining and Quarry World*, 250(2):82–85, 2002.

[BLLD16] Alex Bradley, Haijiang Li, Robert Lark, and Simon Dunn. BIM for infrastructure: An overall review and constructor perspective. *Automation in Construction*, 71:139–152, November 2016.

[Bre15] Florent Brenguier. Noise-Based Seismic Imaging and Monitoring of Volcanoes. In Michael Beer, Ioannis A. Kougioumtzoglou, Edoardo Patelli, and Siu-Kui Au, editors, *Encyclopedia of Earthquake Engineering*, pages 1561–1566. Springer, Berlin, Heidelberg, 2015.

[Bro20] Dorian Brown. Rank-BM25: A Collection of BM25 Algorithms in Python. `https://doi.org/10.5281/zenodo.4520057`, 2020.

[Bun15] Bundesministerium für Verkehr und digitale Infrastruktur. Stufenplan Digitales Planen und Bauen. `https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/stufenplan-digitales-bauen.pdf?__blob=publicationFile`, 2015. Accessed: 2021-01-20.

[Cab16] Cabinet Office and Infrastructure and Projects Authority. Government Construction Strategy: 2016-2020. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/510354/Government_Construction_Strategy_2016-20.pdf, 2016. Accessed: 2021-01-20.

[CCK+00] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. R. Shearer, and R. Wirth. Crisp-dm 1.0: Step-by-step data mining guide. https://www.the-modeling-agency.com/crisp-dm.pdf, 2000. Accessed: 2021-01-20.

[CCK14] Daegu Cho, Hunhee Cho, and Daewon Kim. Automatic data processing system for integrated cost and schedule control of excavation works in natm tunnels. *Journal of Civil Engineering and Management*, 20:132–141, 03 2014.

[CG16] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[CT94] William Cavnar and John Trenkle. N-Gram-Based Text Categorization, May 1994.

[Dan20] Michal Danilák. Port of nakatani shuyoś language-detection library, version from 03-03-2014 to python. https://github.com/Mimino666/langdetect, 2020. Accessed: 2020-10-22.

[DFSS14] Jürgen Demharter, Sebastian Fuchs, Sven-Eric Schapke, and Raimar J. Scherer. Multimodell und Multimodellcontainer. In Raimar J. Scherer and Sven-Eric Schapke, editors, *Informationssysteme im Bauwesen 1: Modelle, Methoden und Prozesse*, VDI-Buch, pages 39–63. Springer, Berlin, Heidelberg, 2014.

[DiEL16] Josef Daller, Marko Žibert, Christoph Exinger, and Martin Lah. Implementation of bim in the tunnel design – engineering consultant's aspect. *Geomechanics and Tunnelling*, 9(6):674–683, 2016.

[DMRS14] K.D. Dhote, K.P.S. Murthy, K.M. Rajan, and M.M. Sucheendran. Quantification of projection angle in fragment generator warhead. In *28th International Symposium on Ballistics*, volume 10, pages 177–183, 2014.

[EH02] A.S. Elkadi and M. Huisman. 3D-GSIS geotechnical modelling of tunnel intersection in soft ground: The Second Heinenoord Tunnel, Netherlands. *Tunnelling and Underground Space Technology*, 17(4):363–369, 2002.

[Els20a] Elsevier B.V. Automation in Construction. https://www.journals.elsevier.com/automation-in-construction, 2020. Accessed: 2021-01-20.

[Els20b] Elsevier B.V. Elsevier Developer Portal. https://dev.elsevier.com/api_docs.html, 2020. Accessed: 2020-10-15.

[Els20c] Elsevier B.V. Elsevier Developer Portal - Scopus Search Views. https://dev.elsevier.com/sc_search_views.html, 2020. Accessed: 2020-10-15.

[Els20d] Elsevier B.V. Scopus - Welcome to Scopus. https://www.scopus.com/, 2020. Accessed: 2020-10-22.

[FFG+19] Georg Fröch, Matthias Flora, Werner Gächter, Florian Harpf, and Arnold Tautschnig. Anwendungsmöglichkeiten eines digitalen Baugrundmodells im Tunnelbau. *Bautechnik*, 96(12):885–895, 2019.

[FPS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37–37, March 1996.

[GB18] Gerald Goger and Tobias Bisenberger. Tunnelling 4.0 – Construction-related future trends. *Geomechanics and Tunnelling*, 11(6):710–721, 2018.

[gir05] Ausführungsvorbereitung. In Gerhard Girmscheid, editor, *Angebots- und Ausführungsmanagement — Leitfaden für Bauunternehmen: Erfolgsorientierte Unternehmensführung vom Angebot bis zur Ausführung*, pages 89–206. Springer, Berlin, Heidelberg, 2005.

[Gir10a] Gerhard Girmscheid. Bauhof- und Bauinventarmanagement. In Gerhard Girmscheid, editor, *Strategisches Bauunternehmensmanagement: Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft*, VDI-Buch, pages 849–939. Springer, Berlin, Heidelberg, 2010.

[Gir10b] Gerhard Girmscheid. Industrielle Bauprozesse. In Gerhard Girmscheid, editor, *Strategisches Bauunternehmensmanagement: Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft*, VDI-Buch, pages 525–551. Springer, Berlin, Heidelberg, 2010.

[Gir10c] Gerhard Girmscheid. Kooperations- und Outsourcingstrategien. In Gerhard Girmscheid, editor, *Strategisches Bauunternehmensmanagement: Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft*, VDI-Buch, pages 287–362. Springer, Berlin, Heidelberg, 2010.

[Gir10d] Gerhard Girmscheid. Organisation von Bauunternehmen. In Gerhard Girmscheid, editor, *Strategisches Bauunternehmensmanagement: Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft*, VDI-Buch, pages 363–424. Springer, Berlin, Heidelberg, 2010.

[Gir10e] Gerhard Girmscheid. Risikomanagement in Bauprojekten und Bauunternehmen. In Gerhard Girmscheid, editor, *Strategisches Bauunternehmensmanagement: Prozessorientiertes integriertes Management für Unternehmen in der Bauwirtschaft*, VDI-Buch, pages 697–806. Springer, Berlin, Heidelberg, 2010.

[GPUA14] J.J. García Adeva, J.M. Pikatza Atxa, M. Ubeda Carrillo, and E. Ansuategi Zengotitabengoa. Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, 41(4, Part 1):1498–1508, 2014.

[GUP14] G. K. GUPTA. *INTRODUCTION TO DATA MINING WITH CASE STUDIES*. PHI Learning Pvt. Ltd., third edition, June 2014.

[GWZ18] Christian Gruber, Thorsten Weiner, and Ralf Zuchtriegel. BIM for tunnelling for a company – Approaches and strategies. *Geomechanics and Tunnelling*, 11(4):366–373, 2018.

[Hev07] Alan Hevner. A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, 19, January 2007.

[HMVLB20] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

[HRM+04] Alan Hevner, Alan R, Salvatore March, Salvatore T, Park, Jinsoo Park, Ram, and Sudha. Design Science in Information Systems Research. *Management Information Systems Quarterly*, 28:75, March 2004.

[Int15] International DOI Foundation. Digital Object Identifier System Resources. `https://www.doi.org/resources.html`, May 2015. Accessed: 2020-10-22.

[ISO16] ISO Central Secretary. Codes for the representation of names of languages. Standard ISO/IEC TR 29110-1:2016, International Organization for Standardization, Geneva, CH, 2016.

[Joh16] John Eynon. *Construction Manager's BIM Handbook*. Wiley Blackwell, 2016.

[Joh21a] John Wiley & Sons, Inc. Search Retrieve via Url. `https://onlinelibrary.wiley.com/action/sru`, 2021. Accessed: 2021-01-31.

[Joh21b] John Wiley & Sons, Inc. Wiley Online Library. https://onlinelibrary.wiley.com/, 2021. Accessed: 2021-01-31.

[KC07] Barbara Kitchenham and Stuart Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical report, Keele University, Durham, UK, January 2007.

[Kit04] Barbara Kitchenham. Procedures for Performing Systematic Reviews. https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf, July 2004. Accessed: 2021-01-22.

[KPBB+09] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, 51(1):7–15, January 2009.

[KRNS17] Markus König, Tobias Rahm, Felix Nagel, and Ludger Speier. BIM-Anwendungen im Tunnelbau. *Bautechnik*, 94(4):227–231, 2017.

[Lin19] Jimmy Lin. The Simplest Thing That Can Possibly Work: Pseudo-Relevance Feedback Using Text Classification. *arXiv:1904.08861 [cs]*, April 2019.

[Lov68] Julie Beth Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11(nos. 1 and 2):22–31, 1968.

[Mat75] B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, October 1975.

[MP18] Marcin Michał Mirończuk and Jarosław Protasiewicz. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106:36–54, September 2018.

[MSK95] Lawrence Mann, Anuj Saxena, and Gerald M. Knapp. Statistical-based or condition-based preventive maintenance? *Journal of Quality in Maintenance Engineering*, 1(1):46–59, January 1995.

[Nak10] Shuyo Nakatani. Language Detection Library for Java. http://code.google.com/p/language-detection/, 2010. Accessed: 2020-10-22.

[PBC+19] Marc Peaucelle, Cédric Bacour, Philippe Ciais, Nicolas Vuichard, Sylvain Kuppel, Josep Peñuelas, Luca Belelli Marchesini, Peter D. Blanken, Nina Buchmann, Jiquan Chen, Nicolas Delpierre, Ankur R. Desai, Eric Dufrene, Damiano Gianelle, Cristina Gimeno-Colera, Carsten Gruening, Carole Helfter, Lukas Hörtnagl, Andreas Ibrom, Richard Joffre, Tomomichi Kato, Thomas E. Kolb, Beverly Law, Anders Lindroth, Ivan Mammarella, Lutz

Merbold, Stefano Minerbi, Leonardo Montagnani, Ladislav Šigut, Mark Sutton, Andrej Varlagin, Timo Vesala, Georg Wohlfahrt, Sebastian Wolf, Dan Yakir, and Nicolas Viovy. Covariations between plant functional traits emerge from constraining parameterization of a terrestrial biosphere model. *Global Ecology and Biogeography*, 28(9):1351–1365, 2019.

[PBR+16] Tony Printemps, Nicolas Bernier, Eric Robin, Zineb Saghi, and Lionel Hervé. 3D Elemental and interdependent reconstructions based on a novel compressed sensing algorithm in electron tomography. In *European Microscopy Congress 2016: Proceedings*, pages 67–68. American Cancer Society, 2016.

[PFMM08] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, pages 68–77, Italy, June 2008. BCS Learning & Development Ltd.

[PRTV12] Ken Peffers, Marcus Rothenberger, Tuure Tuunanen, and Reza Vaezi. Design Science Research Evaluation. In Ken Peffers, Marcus Rothenberger, and Bill Kuechler, editors, *Design Science Research in Information Systems. Advances in Theory and Practice*, Lecture Notes in Computer Science, pages 398–410, Berlin, Heidelberg, 2012. Springer.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[PVK15] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18, August 2015.

[QYHW16] Siyu Qian, Ping Yu, David M. Hailey, and Ning Wang. Factors influencing nursing time spent on administration of medication in an Australian residential aged care home. *Journal of Nursing Management*, 24(3):427–434, 2016.

[Raf18] K. Rafie. Interpretation of EPB TBM graphical data. In *North American Tunneling Conference, NAT 2018*, volume 1, pages 111–120, 2018.

[Rau17] Andreas Rauber. VU Business Intelligence Data Mining Supervised ML Classification, February 2017.

[Rau19] Andreas Rauber. Research Methods in Data Analysis: Process Models and Reproducibility, November 2019.

[RF21] Kenneth Reitz and Python Software Foundation. Requests: a simple, yet elegant HTTP library. https://doi.org/10.5281/zenodo.1212303, 2021. Accessed: 2021-01-22.

[RK19] Michael E. Rose and John R. Kitchin. Pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX*, 10:100263, July 2019.

[RMj+21] Jeff Reback, Wes McKinney, jbrockmendel, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, gfyoung, Simon Hawkins, Sinhrks, Matthew Roeschke, and et al. pandas-dev/pandas: Pandas 1.2.2. https://doi.org/10.5281/zenodo.4524629, Feb 2021.

[SDvB+99] M. P. M. Steultjens, J. Dekker, M. E. van Baar, R. a. B. Oostendorp, and J. W. J. Bijlsma. Internal consistency and validity of an observational method for assessing disability in mobility in patients with osteoarthritis. *Arthritis Care & Research*, 12(1):19–25, 1999.

[SKA+17] D.K. Sheet, O. Kaiwartya, A.H. Abdullah, Y. Cao, A.N. Hassan, and S. Kumar. Location information verification using transferable belief model for geographic routing in vehicular ad hoc networks. *IET Intelligent Transport Systems*, 11(2):53–60, 2017.

[Sme90] P. Smets. The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, May 1990.

[SNM+18] Janosch Stascheit, Jelena Ninić, Günther Meschke, Felix Hegemann, and Ulrich Maidl. Building Information Modelling in mechanised shield tunnelling – A practitioner's outlook to the near future. *Geomechanics and Tunnelling*, 11(1):34–49, 2018.

[Spr21a] Springer Nature Switzerland AG. Advanced Search. https://link.springer.com/advanced-search, 2021. Accessed: 2021-01-31.

[Spr21b] Springer Nature Switzerland AG. CSV Export. https://link.springer.com/search/csv, 2021. Accessed: 2021-01-31.

[Spr21c] Springer Nature Switzerland AG. Example API Responses. https://dev.springernature.com/example-metadata-response, 2021. Accessed: 2021-01-31.

[Spr21d] Springer Nature Switzerland AG. SpringerLink. https://link.springer.com/, 2021. Accessed: 2021-01-31.

[Spr21e] Springer Nature Switzerland AG. SpringerNature API Portal. https://dev.springernature.com/, 2021. Accessed: 2021-01-31.

[SSJ+17] Elizabeth Schenk, Ruth Schleyer, Cami R. Jones, Sarah Fincham, Kenn B. Daratha, and Karen A. Monsen. Time motion analysis of nursing work in ICU, telemetry and medical-surgical units. *Journal of Nursing Management*, 25(8):640–646, 2017.

[TA17] M. Kimura T. Adachi, K. Tateyama, editor. *Modern Tunneling Science and Technology : Volume 1*. 2017.

[TK19] Thomas Tschickardt and Daniel Krause. BIM im Verkehrswegebau am Beispielprojekt „Verfügbarkeitsmodell A 10/A 24“. *Bautechnik*, 96(3):259–268, 2019.

[TPB14] Andrew Trotman, Antti Puurula, and Blake Burgess. Improvements to BM25 and Language Models Examined. In *Proceedings of the 2014 Australasian Document Computing Symposium on - ADCS '14*, pages 58–65, Melbourne, VIC, Australia, 2014. ACM Press.

[UK 11] UK Cabinet Office. Government Construction Strategy. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/61152/Government-Construction-Strategy_0.pdf`, May 2011. Accessed: 2020-09-13.

[UK 12] UK Government. Building information modelling. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/34710/12-1327-building-information-modelling.pdf`, 2012. Accessed: 2021-02-09.

[VB17] Simon Vilgertshofer and André Borrmann. Using graph rewriting methods for the semi-automatic generation of parametric infrastructure models. *Advanced Engineering Informatics*, 33:502–515, August 2017.

[VCP+13] P. Vos, P. De Cock, K. Petry, W. Van Den Noortgate, and B. Maes. Investigating the relationship between observed mood and emotions in people with severe and profound intellectual disabilities. *Journal of Intellectual Disability Research*, 57(5):440–451, 2013.

[VSS14] Rebekka Volk, Julian Stengel, and Frank Schultmann. Building Information Modeling (BIM) for existing buildings — Literature review and future needs. *Automation in Construction*, 38:109–127, March 2014.

[WJMB11] D.J. Wald, K.S. Jaiswal, K.D. Marano, and D. Bausch. Earthquake impact scale. *Natural Hazards Review*, 12(3):125–139, 2011.

[WMC+20] Sabine Wolny, Alexandra Mazak, Christine Carpella, Verena Geist, and Manuel Wimmer. Thirteen years of SysML: A systematic mapping study. *Software and Systems Modeling*, 19(1):111–169, January 2020.

[WMMR06] Roel Wieringa, Neil Maiden, Nancy Mead, and Colette Rolland. Requirements engineering paper classification and evaluation criteria: A proposal and a discussion. *Requirements Engineering*, 11(1):102–107, March 2006.

[WPL19] Hao Wang, Yisha Pan, and Xiaochun Luo. Integration of BIM and GIS in sustainable built environment: A review and bibliometric analysis. *Automation in Construction*, 103:41–52, July 2019.

[WRH+12] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in Software Engineering*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[ZDKB13] Konrad Zilch, Claus Jürgen Diederichs, Rolf Katzenbach, and Klaus J. Beckmann. *Allgemeine Grundlagen*. Springer, Berlin, Heidelberg, 2013.