Conference Paper, Published Version

Huang, Jiuru

Aufbau und Erweiterung eines Wissensbasis-gestützten Datenmodells mit Semantic Web Technologie und Verarbeitung natürlicher Sprache im Kontext vom Multiprojektmanagement im Verkehrswasserbau

Aufbau und Erweiterung eines Wissensbasis-gestützten Datenmodells mit Semantic Web Technologie und Verarbeitung natürlicher Sprache im Kontext vom Multiprojektmanagement im Verkehrswasserbau

Jiuru Huang[1]

[1]Bundesanstalt für Wasserbau

Kußmaulstraße 17, 76187 Karlsruhe, Germany

E-Mails: jiuru.huang@baw.de

Abstract: Im Bereich des Verkehrswasserbaus entsteht aktuell das Informationssystem „Multiprojektmanagement der Wasserstraßen- und Schifffahrtsverwaltung des Bundes (WSV)“, welches einen aktuellen und umfassenden Überblick aller projektrelevanten Informationen von Baumaßnahmen bieten soll. Dabei bezieht das Informationssystem die Daten aus den anderen etablierten Informationssystemen. Diese Daten werden jedoch in heterogener Art und Qualität gespeichert und für unterschiedliche Nutzungsanforderungen verwertet. Der Überführung relevanter Informationen in maschinenlesbare und -verarbeitbare Daten kommt künftig aber mehr Bedeutung zu. In einem Forschungs- und Entwicklungsprojekt der Bundesanstalt für Wasserbau wird der Ansatz der Semantic Web Technologie (SWT) zum Aufbau eines Wissensbasis-gestützten Datenmodells zur Integration der heterogenen Daten untersucht. Im Zuge dessen wurde ein Ansatz zur Verarbeitung natürlicher Sprache (engl. Natural Language Processing, kurz NLP) verwendet, um neue Entitäten für die Wissensbasis aus den Daten der natürlichen Sprache zu gewinnen. Verschiedene Techniken der NLP, wie z. B. Erkennung bekannter Entitäten (engl. Named Entity Recognition, kurz NER), wurden auf die zuvor erstellte Wissensbasis angewendet, um z. B. in einem vortrainierten Modell neues Vokabular zur Verarbeitung hinzuzufügen. Der Beitrag zeigt exemplarische Ergebnisse dieses Ansatzes und schließt mit einer fachlichen Evaluation.

Keywords: Wissensbasis, SWT, NLP, NER, Verkehrswasserbau

1 Problemstellung und Motivation

Die IT-Anwendungslandschaft in der Wasserstraßen- und Schifffahrtsverwaltung des Bundes (WSV) ist aufgrund ihrer langen Entwicklungshistorie technologisch heterogen. Es entsteht aktuell das Informationssystem “Multiprojektmanagement der WSV”, welches einen aktuellen und umfassenden Überblick aller projektrelevanten Informationen von Baumaßnahmen bieten soll. Dabei bezieht das Informationssystem die Daten manuell aus den anderen etablierten Informationssystemen. Diese Daten werden jedoch in heterogener Art und Qualität gespeichert und für unterschiedliche Nutzungsanforderungen verwertet. Der Überführung relevanter Informationen in maschinenlesbare und -verarbeitbare Daten kommt künftig aber mehr Bedeutung zu. Ein Informationsmodell, das eine effizient vernetzte IT-Landschaft zur optimalen Bedienung der Informationsanforderungen der Nutzer bündelt und integriert, existiert nicht. In einem Forschungs- und Entwicklungsprojekt der Bundesanstalt für Wasserbau (BAW) wird der Ansatz der Semantic Web Technologie (SWT) zum Aufbau eines Wissensbasis-gestützten Datenmodells zur Integration der heterogenen Daten untersucht. In Huang [1] wurde dieser Ansatz hierzu bereits erläutert. In diesem Beitrag werden die Fortführung sowie der aktuelle Stand dieser Entwicklung vorgestellt.

In der vorherigen Untersuchung in Huang [1] wurde festgestellt, dass die SWT für die Integration der Daten in der Qualität eines Datenbanken-ähnlichen Systems einerseits geeignet ist und andererseits liegen größtenteils relevante Informationen in Daten der natürlichen Sprache, wie z. B. Dokumente, Entwürfe, vor. Basierend auf dieser Feststellung wird ein kombinierter Ansatz – SWT mit Unterstützung von Methoden des maschinellen Lernens (ML) – angestrebt. Mit der Entwicklung des ML kommen Sprachmodelle zur Verarbeitung natürlicher Sprache (engl. Natural Language Processing, kurz NLP) zum Einsatz, welche anhand großer Datensätze trainiert wurden und an den Domänenbereich benutzerdefiniert angepasst werden können [2, 3]. Verschiedene Techniken der NLP sollen auf die zuvor erstellte Wissensbasis angewendet werden, um z. B. in einem vortrainierten Modell neues Vokabular zur Verarbeitung hinzuzufügen und neue Entitäten für eine Wissensbasis aus den Daten zu gewinnen. Kapitel 2 führt die Modellierungsmethode der SWT ein und erläutert den aktuellen Stand des Wissensbasis-gestützten Datenmodells. Im Kapitel 3 wird der kombinierte Ansatz mit NLP zur Erweiterung der Wissensbasis prinzipiell demonstriert. Der Beitrag schließt mit einer fachlichen Evaluation und Disskusion.

2 Datenmodell als Wissensbasis

2.1 Modellierung der Wissensbasis

In Huang [1] wurde die Modellierungsmethode zur Überführung relevanter Informationen in Daten mit SWT bereits beschrieben, welche zum besseren Verständnis im Folgenden zusammengefasst wird:

Es wird ein Domänenbereich, in dem die Wissensbasis aufgebaut werden soll, festgelegt. Es werden Abfragen, die das Datenmodell bedienen sollen, als Anwendungsszenarien definiert. Auf diesen Abfragen aufbauend wird ein Assoziationsnetz als Entwurf für die Wissensbasis konzipiert, welches die im Datenmodell befindlichen Klassen bildet. Die Bestandteile der Wissensbasis werden in formaler Sprache mittels der Methode SWT definiert. Diese Definitionen werden durch eine umfangreiche Recherche der Verwaltungsvorschriften und der Regelwerke widerspruchfrei mit Quellenangaben belegt. Das entstandene Datenmodell wird fortlaufend konsistent und kohärent gehalten. Anschließend werden Quellspeicherorte der Wissensbasis und deren Datenstrukturen in den Datenbeständen der verteilten IT-Systeme analysiert, um die Durchsuchungsmöglichkeit der vorhandenen IT-Systeme über die Wissensbasis zu skizzieren. Final soll eine Wissensbasis entstehen, in der alle integrierten Daten mit der Sprache SPARQL abgefragt werden können. Die technologischen Grundlagen sind in Hitzler et al. [4] beschrieben.

2.2 Stand der Wissensbasis

Es wurde eine Wissensbasis für das Informationssystem Multiprojektmanagement der WSV konzipiert. Darin wurden sukzessiv Entitäten nach festgelegten Anwendungsszenarien definiert. Aktuell befinden sich in der Wissensbasis 397 Klassen, welche über 17 Properties miteinander in Beziehungen gesetzt wurden. Davon sind 132 Klassen von Bauwerken und Bauteilen, 99 Klassen von Maßnahmen, 47 Klassen von Wasserstraßen. Organisationen, abstrakte Personen, Bauwerksinspektionsnoten, sowie beschreibende Klassen wie Funktion und Zweck der Bauwerke und der Wasserstraße sind enthalten. Abbildung 1 zeigt den in WebVOWL [5] visualisierten Stand der Wissensbasis, deren Individuen nach und nach in das Datenmodell integriert werden.

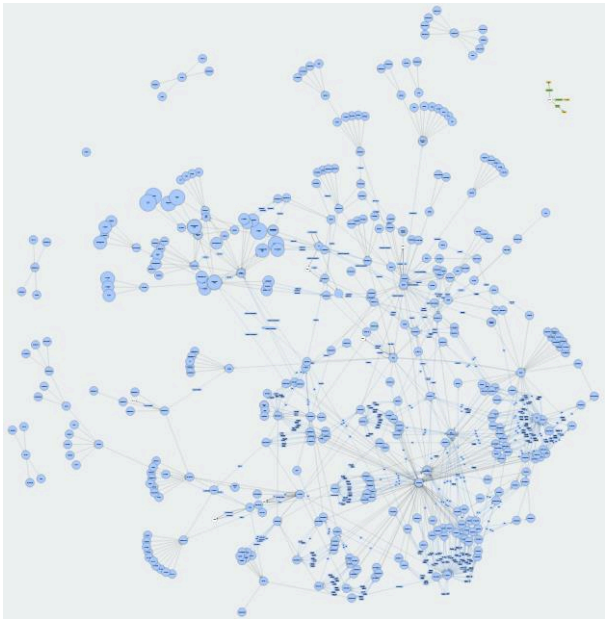

Abbildung 1: Visualisierung der Wissensbasis in WebVOWL [5]

Diese Wissensbasis wird laufend durch das Programm “Debugger” von Protégé auf Konsistenz und Kohärenz geprüft. Die Datenintegration betrifft sowohl die Klassen und Properties als auch die Individuen und deren Datenwerte. Eine Datenbestandsanalyse dient dazu, das Vorhandensein und die Qualität der Daten, die als Klassen, Properties, Individuen vorfinden, zu dokumentieren. Die in Quelldatenbanken befindlichen Daten werden als Klasse, Property oder als Individuum in der Wissensbasis definiert. Bei der Integration müssen die Daten in der Wissensbasis und in den Quellsystemen eindeutig annotiert sein, sodass diese über Application Programming Interface (API) ausgetauscht werden können. Oftmals stellen Quellsysteme kein API zur Verfügung und diese Daten sind nicht identifiziert, d. h. uneindeutig oder widersprüchlich. Um das Problem zu lösen, muss eine Festlegung der Identifikation der zu integrierenden Daten getroffen werden, um die Daten mit einer eindeutigen Kennung zu annotieren. Z. B. wird für jeden Schaden die Schadensnummer, adiert mit Objektidentnummern zwecks eindeutiger Identifikation vergeben – eine Art der Datenveredelung. Der obengenannte Prozess wird derzeit noch manuell bearbeitet. Ein Konzept zur automatisierten Datenkonvertierung wird aufgestellt. Die Datenumwandelung soll so programmiert werden, dass die Daten unabhängig von den Quellsystemen und aus den Quellsystemen direkt automatisch konvertiert und aktualisiert werden, ohne die Quelldaten zu verändern.

Zwei ausgewählte Beispiele demonstrieren das Prinzip einer Wissensbasis. Abbildung 2 stellt die Klasse Projektklasse “Fischpass Wehr Synergie” und ihre Klassenbeziehungen und Abbildung 3 stellt das Individum Wasserstraße “Rhein, Wehrarm Burkheim” dar. Beide sind in WebProtégé [6] visualisiert.

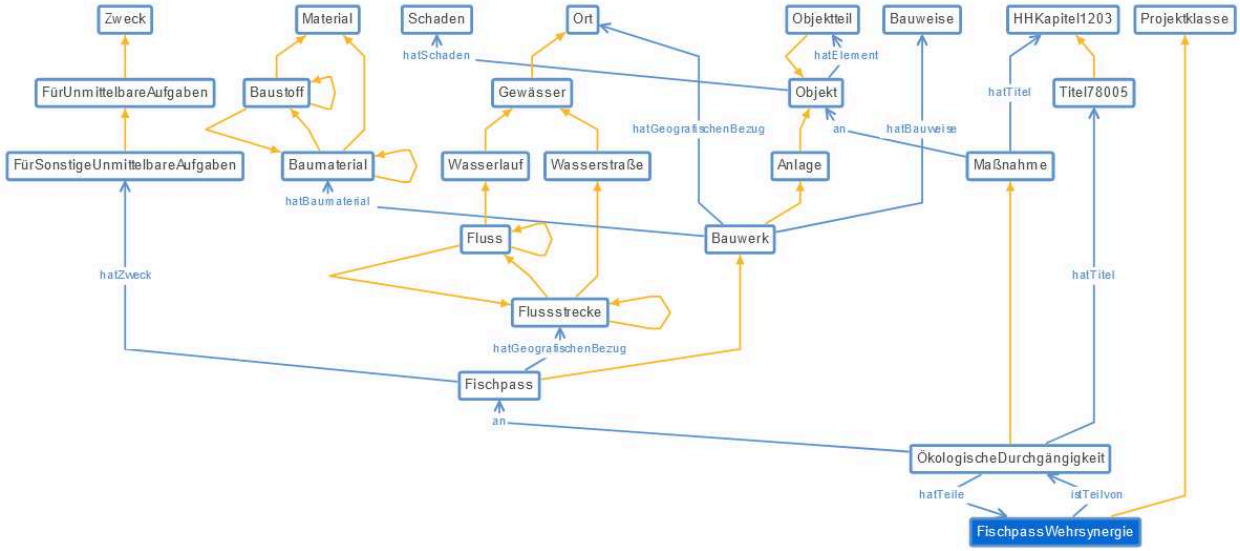

Abbildung 2: Klasse Projektklasse “Fischpass Wehr Synergie” und ihre Klassenbeziehungen dargestelt in WebProtégé [6]

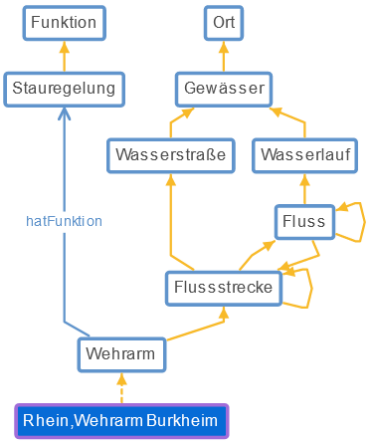

Abbildung 3: Individuum Wasserstraße “Rhein, Wehrarm Burkheim” und seine Klasse sowie Klassenbeziehungen, dargestellt in WebProtégé [6]

2.3 Anwendung der Wissensbasis

Die Wissensbasis kann recherchiert, durchsucht und kollaborativ via Web Protégé [6] editiert werden. Mit den Abfragesprachen DL Query und SPARQL Query können alle in der Wissensbasis befindenlichen Informationen, wie Individuen bestimmter Klasse oder abgeleitete Klassen, abgefragt werden.

Zwar erfüllt der Ontology-Editor Protégé die Grundvoraussetzung eines Datenbankensystems, um relevante Informationen in Daten zu speichern, zu ändern und abzufragen, ist solch eine Nutzung jedoch abhängig vom Programm Protégé und damit auch nur eingeschränkt bei der Fortschreibung und Pflege der darin befindlichen Daten. Eine unabhängige Programmierung bis hin zum Lizenzmodell mit Zugriffssystem ist daher notwendig, um eine breite Anwendung zu gewährleisten. Eine Webanwendung für die Nutzung der Wissensbasis u. a. in Datenformaten RDF/OWL (auch triplestore genannt) wird derzeit umgesetzt.

3 Wissensbasis zur Verarbeitung natürlicher Sprache

3.1 Einführung

Die manuelle Erstellung einer Wissensbasis ist aufwendig. Mit der Entwicklung im Bereich des maschinellen Lernens (ML) sind neue Technologien entstanden – u. a. NLP, die für die Daten der natürlichen Sprachen speziell entwickelt ist [2, 3]. Es existieren bereits vortrainierte Sprachmodelle, z. B. spaCy [7], die einige Entitäten in solch beliebigen Daten mit hoher Genauigkeit erkennen können [8]. Daher wird diese Technologie im Hinblick auf die Frage untersucht, in wieweit diese zur Unterstützung der Erstellung einer Wissensbasis angewendet werden können. Ziel ist es, die

Erkennung relevanter Entitäten und Relationen aus Daten für die Wissensbasis zu automatisieren. Zunächst wird ein vortrainiertes Sprachmodell gewählt, dann die erste manuell erstellte Wissensbasis für das Hinzufügen von Vokablen verwendet. Die Trainingsdaten werden anhand der Wissensbasis generiert und anschließend zum Erlernen des neuen Modells in den maschinellen Lernprozess eingegeben. Das erlernte Modell, welches spezielle Vokabeln der Wissensbasis erkennt, kann anschließend auf weitere Daten zur Erkennung von Entitäten und Relationen eingesetzt werden, welche durch Domänenexperten geprüft der erweiterten Wissensbasis hinzugefügt werden.

3.2 Anwendungsbeispiele

Nachfolgend werden die untersuchten Techniken der NLP anhand der Algorithmen 1-3 erläutert. Zunächst wurde das Python-Paket der Organisation spaCy importiert, dann das auf den deutschen Nachrichten trainierte mittelgroße Sprachmodell [9] geladen. Ein paar eigene Muster und ein Wörterbuch wurden zu dem Modell hinzugefügt, um die Named Entity Recognition (NER) auf die ausgewählten Datensätzen mit angepassten benutzerdefinierten Attributen anzuwenden.

Algorithmus 1: Hinzufügen von Mustern und Wörtern

```
1  import spacy
2  nlp = spacy.load("de_core_news_md")
3  ruler = nlp.add_pipe("entity_ruler")
4  patterns = [{"label": "ORG", "pattern": "Generaldirektion
   Wasserstraßen und Schifffahrt"},{"label": "ORG", "pattern":
   "Wasser- und Schifffahrtsdirektion Mitte"}]
5  ruler.add_patterns(patterns)
6  DICTIONARY = {"GDWS":"Genernaldirektion Wasserstraßen und
   Schifffahrt","WSD Mitte":"Wasser- und Schifffahrtsdirektion
   Mitte"}
```

Algorithmus 2: Erkennen von bekannten Entitäten mit Named Entity Recognition (NER)

```
1  text = ("current_dataset")
2  doc = nlp(text)
3  for entity in doc.ents:
   print(entity.text,entity.label_,spacy.explain(entity.label_))
```

Algorithm 3: Inspizieren von Nominalphrasen und Verben mit Part-of-speech Tagging (POS)

```
1   print("Noun phrases:", [chunk.text for chunk in
    doc.noun_chunks])
    print("Verbs:", [token.lemma_ for token in doc if token.pos_ ==
    "VERB"])
```

Die Ausgabe der Algorithmen 1-3 auf den ausgewählten Datensätzen ist in Abbildung 4 zu sehen. Abbildung 5 stellt die von spaCy visualisierten NER dar. Die Dependenz, visualisiert in Abbildung 6, kann für das Finden von Subjekt-Objekt-Beziehungen (S-P-O) verwendet werden. Die gefundenen S-P-O können der Wissensbasis hinzugefügt werden und als Trainingssatz zum maschinellen Lernprozess dienen. Dabei werden die Nominalphrasen mit den dazugehörigen Verben markiert. Im Beispielsatz lauten die S-P-O "WSA Verden"-"aufstellen"-"Entwurf-AU" und "WSD Mitte"-"genehmigen"-"Entwurf-AU". Alle richtig erkannten Entitäten und S-P-O sind potenziell die neuen Kandidaten der Klasse und Properties in der Wissensbasis.

```
Entwurf-AU Nr. I 3312.000.2013.05.A.23 MISC Miscellaneous entities, e.g.
events, nationalities, products or works of art
Mittelwesennrehren Petershagen, Schlüsselburg und Lang-wedel LOC Non-GPE
locations, mountain ranges, bodies of water
WSA Verden ORG Companies, agencies, institutions, etc.
WSD Mitte ORG Companies, agencies, institutions, etc.
Noun phrases: ['Der Entwurf-AU', '3312.000.2013.05.A.23', 'den Ersatz',
'der Laufeinrichtungen', 'den Mittelwesennrehren Petershagen', 'Schlüsse
lburg', 'Lang-wedel', 'WSA Verden', '15. April', 'der WSD', 'Mitte', '2
5. Juni']
Verbs: ['aufstellen', 'genehmigen']
```

Abbildung 4: Ausgaben der ausgeführten Algorithmen 1-3 auf Beispiel-Datensätzen

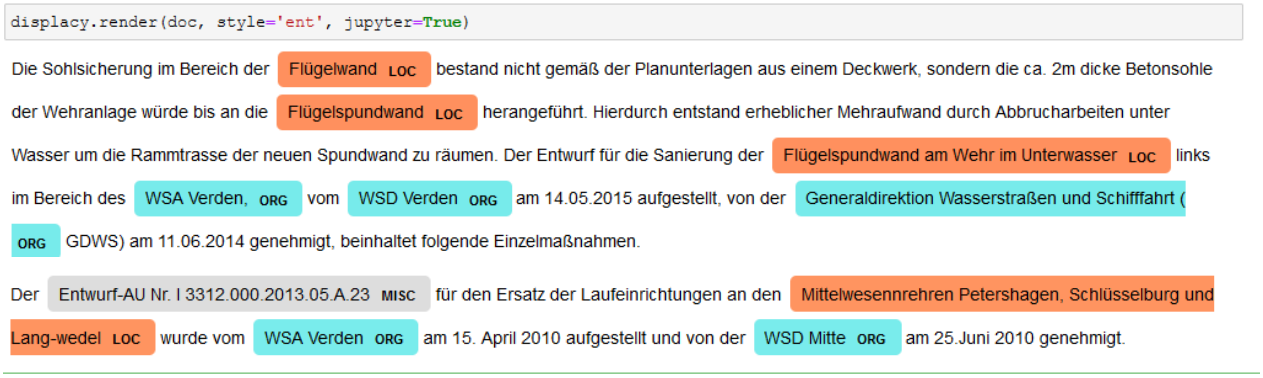

Abbildung 5: Visualisierung der erkannten Entitäten

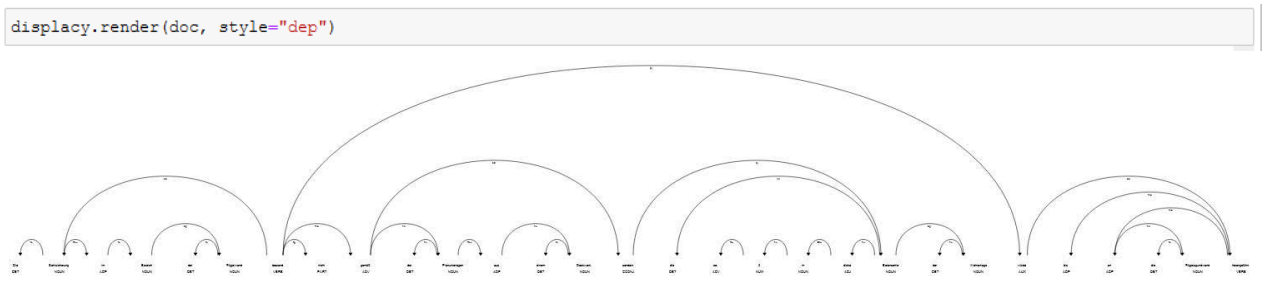

Abbildung 6: Visualisierung der Dependenz der Tokens

4 Fachliche Evaluation, Diskussion und Fazit

Dieser Beitrag zeigt den kombinierten Ansatz mit SWT und NLP. Die SWT ist nützlich für die Erstellung einer Wissensbasis für den Domänenbereich mit Datenbanken-ähnlichen Daten. Diese Wissensbasis enthält relevante Informationen, die auf eine machinenlesbare Art und Weise spezifiziert wurden. Die Herausforderung besteht darin, die Daten in den Quellsystemen eindeutig zu identifizieren, sodass diese mit der Wissensbasis verbunden und von der Wissensbasis abgefragt werden können. Dadurch wird eine Durchsuchung der IT-Landschaft möglich. Bis jetzt wurde die NER der NLP mit angepassten Attributen, d. h. ohne das Modell zu trainieren, angewendet. Erkannt wurden auf den Testdatensätzen insgesamt Organisationen (ORG), Orte (LOC) und sonsitge Begriffe (MISC). Die im Kapitel 2 gezeigten Klassen (z. B. Wasserstraße, Bauwerk, Maßnahme und abstrakte Person) werden als neue Muster oder Wörter für das Modell hinzugefügt. Die hinzugefügten benutzerdefinierten Muster und Wörter wurden richtig erkannt. Fehl- und Falscherkennungen sind vorhanden. Um eine Domänen-spezifische Genauigkeit der NER festzulegen, bedarf es mehr Testdaten. Diese Untersuchung soll durchgeführt werden, um im nächsten Schritt das angepasste Sprachmodell auf größeren Datensätzen anzuwenden. Die Generierung der neuen Wissensbasis und des Trainingssatzes wurden vorgestellt. Eine Automatisierung soll diese Prozesse beschleunigen.

Literatur

[1] J. Huang, „Datenabfrage und -integration im Kontext vom Multiprojektmanagement im Verkehrswasserbau mit Semantic Web Technologie“ in *32. Forum Bauinformatik 2021*, S. 157–165. [Online] Verfügbar unter: https://hdl.handle.net/20.500.11970/108246.

[2] T. Hoppe, *Semantische Suche*. Wiesbaden: Springer Fachmedien Wiesbaden, 2020.

[3] W. Ertel, *Grundkurs Künstliche Intelligenz*. Wiesbaden: Springer Fachmedien, 2021.

[4] P. Hitzler, M. Krötzsch, S. Rudolph und Y. Sure, *Semantic Web: Grundlagen*. Springer-Verlag Berlin Heidelberg, 2008.

[5] *WebVOWL: Web-based Visualization of Ontologies.* [Online] Verfügbar unter: http://vowl.visualdataweb.org/webvowl.html. Zugriff am: 6. Juli 2022.

[6] *The WebProtege site.* [Online] Verfügbar unter: https://webprotege.stanford.edu.

[7] *Industrial-Strength Natural Language Processing in Python.* [Online] Verfügbar unter: https://spacy.io/. Zugriff am: 6. Juli 2022.

[8] *Facts & Figures: The hard numbers for spaCy and how it compares to other tools.* [Online] Verfügbar unter: https://spacy.io/usage/facts-figures. Zugriff am: 6. Juli 2022.

[9] *Available trained pipelines for German: de_core_news_md.* [Online] Verfügbar unter: https://spacy.io/models/de#de_core_news_md. Zugriff am: 6. Juli 2022.

Lehrstuhl für Computergestützte Modellierung und Simulation | Prof. Dr.-Ing. André Borrmann
Lehrstuhl für Architekturinformatik | Prof. Dr.-Ing. Frank Petzold

Proceedings of

33. Forum*Bauinformatik*

07. – 09. September 2022

Herausgeber:
Martin Slepicka, Lothar Kolbeck,
Sebastian Esser, Kasimir Forth,
Florian Noichl, Jonas Schlenger

Technische Universität München

Herausgeber
Sebastian Esser, Kasimir Forth
Florian Noichl, Jonas Schlenger
Martin Slepicka, Lothar Kolbeck

Layout & Satz
Martin Slepicka
Lothar Kolbeck

Rein digitale Veröffentlichung über das mediaTUM-Portal der Technischen Universität München

DOI: 10.14459/2022md1686600
URL: https://mediatum.ub.tum.de/1686600

Logo: Forum Bauinformatik | www.forum-bauinformatik.de
Umschlagfoto: Astrid Eckert | Technische Universität München | Leonhard Obermeyer Centrum