CARL VON OSSIETZKY universität OLDENBURG

Fakultät II - Informatik, Wirtschafts- und Rechtswissenschaften

Oldenburger Schriften zur Wirtschaftsinformatik
Hrsg.: Prof. Dr.-Ing. Jorge Marx Gómez

Felix Kruse

End-to-End-Datenintegration von Realwelt-Entitäten

Konzeption eines Record Linkage-Prozesses, Feldexperimente, unternehmensrelevante Implikationen

Abteilung Wirtschaftsinformatik
Carl von Ossietzky Universität Oldenburg

Carl von Ossietzky Universität Oldenburg

End-to-End-Datenintegration von Realwelt-Entitäten

Konzeption eines Record Linkage-Prozesses, Feldexperimente,
unternehmensrelevante Implikationen

Von der
Fakultät für Informatik, Wirtschafts- und Rechtswissenschaften der
Carl von Ossietzky Universität Oldenburg zur Erlangung des
Grades und Titels eines

Doktors der Ingenieurwissenschaften **(Dr.-Ing.)**

angenommene Dissertation
von

Felix Kruse

geboren am
18.07.1991 in Damme

Gutachter: **Prof. Dr.-Ing. habil. Jorge Marx Gómez**
Weiterer Gutachter: **Prof. Dr. Peter Loos**

Tag der Disputation: 13. Oktober 2022 in Oldenburg

Oldenburger Schriften zur Wirtschaftsinformatik

Band 32

Felix Kruse

End-to-End-Datenintegration von Realwelt-Entitäten

Konzeption eines Record Linkage-Prozesses, Feldexperimente, unternehmensrelevante Implikationen

Shaker Verlag
Düren 2022

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Zugl.: Oldenburg, Univ., Diss., 2022

ISBN 978-3-8440-8836-6
ISSN 1863-8627

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren
Telefon: 02421 / 99 0 11 - 0 • Telefax: 02421 / 99 0 11 - 9
Internet: www.shaker.de • E-Mail: info@shaker.de

Danksagung

Zu aller erst bedanke ich mich bei meinem Doktorvater Prof. Dr.-Ing. habil. Jorge Marx Gómez. Jorges großartige Unterstützung in den vergangenen Jahren ist die elementare Grundlage gewesen, damit ich diese Dissertation erfolgreich meistern konnte. Neben den konstruktiven Fachgesprächen zu meiner Dissertation, haben auch die Diskussionen über Werder Bremen und etwaige Tischtennis-Matches dazu beigetragen, dass die Zeit am Lehrstuhl unfassbar viel Spaß gemacht hat.

Ich bedanke mich auch bei meinem Zweitgutachter Prof. Dr. Peter Loos, der mir mit seinem konstruktiven Feedback ebenfalls wichtige Anregungen und Hinweise geliefert hat, um meine Promotion erfolgreich zu beenden.

Ein besonderer Dank geht an alle Kolleginnen und Kollegen der Abteilung VLBA der Universität Oldenburg und an meine ehemaligen Arbeitskollegen aus der OLB, die mich in den letzten Jahren mit konstruktivem Feedback und vielen tollen Erlebnissen während der Promotion begleitet haben. Besonders hervorzuheben ist hier unsere Doktoranden-Selbsthilfegruppe, die aus Marius, Christian, Hauke, Christoph und René bestand.

Weiterhin bedanke ich mich bei den vielzähligen Praxispartnern, die mich während meiner Forschung mit ihrer Expertise unterstützt haben. Ganz besonders bedanke ich mich bei Sven Wittfoth und Jonas Frischkorn, die mich mit konstruktiven Diskussionen und wertvollem Feedback unterstützt haben.

René, JP und Jan-Hendrik, bei euch möchte ich mich nochmal ausdrücklich bedanken. Unsere gemeinsame Zeit in einem Büro an der Uni kann nicht mit Worten beschrieben werden. Ihr habt mir in den letzten Jahren unfassbar geholfen und die Zeit mit euch werde ich niemals vergessen.

Ein riesen großer Dank geht an Philipp, Raphael, Martina, Lars, AK und Christoph. Ihr seid die besten Freunde, die man sich als Unterstützung während einer so langen Zeit voller Höhen und Tiefen wünschen kann. Eure Unterstützung ist entscheidend dafür gewesen, dass ich die Dissertation erfolgreich beendet habe.

Ganz zum Schluss sind natürlich meine Freundin Kathi, meine Eltern Renate und Hubertus sowie meine Schwester Johanna der entscheidende Rückhalt in den letzten Jahren für mich gewesen, um die Dissertation erfolgreich zu meistern. Ohne eure Unterstützung wäre ich niemals soweit gekommen.

Zusammenfassung

Unternehmen wird zunehmend bewusst, dass eine hohe Stammdatenqualität die entscheidende Grundlage für die erfolgreiche Durchführung von KI- und Digitalisierungsprojekten ist. Die oftmals schlechte Stammdatenqualität liegt darin begründet, dass einzelne Fachabteilungen Stammdaten oftmals dezentral in mehreren Datenbanken anlegen, verändern und löschen. Dadurch entstehen und existieren in den jeweiligen Datenbanken Inkonsistenzen und Stammdatenduplikate, wie doppelte Kund:innen oder Lieferant:innen. Diese über mehrere Datenbanken verteilten Stammdaten mit ihren Inkonsistenzen und Duplikaten führen zu zahlreichen Prozessfehlern. Weiterhin existiert kein Gesamtüberblick über die Stammdaten, sodass Mitarbeiter:innen nur mit sehr hohem manuellem Aufwand alle vorhandenen Informationen aus den verteilten Datenbanken in ihre Geschäftsaktivitäten einfließen lassen können.

Um die notwendige Stammdatenqualität herzustellen, führen Unternehmen Datenintegrationsprojekte durch. Bei der Datenintegration werden Stammdaten aus mehreren Quellen zusammengeführt und einheitlich dargestellt. Zur Durchführung dieser Projekte werden IT-Fachexpert:innen benötigt, die einen enormen manuellen Aufwand für die Identifzierung und Implementierung geeigneter Algorithmen betreiben. Zudem verstärkt das exponentielle Wachstum der Datenquellen den Bedarf der Datenintegration von Unternehmen.

Der bisherige Stand der Forschung der Datenintegration wird kaum in Unternehmen eingesetzt, da der manuelle Aufwand zur Bewertung der Ergebnisqualität nach wie vor zu hoch ist. Bestehende Datenintegrationssysteme wie Magellan oder BigGorilla stellen eine Vielzahl an Algorithmen und Verfahren für die Datenintegration zur Verfügung, schaffen es allerdings nicht die Automatisierung weiter voranzutreiben. Laut Gartner existiert durch den Einsatz von KI in der Datenintegration ein Einsparpotenzial von 45% des manuellen Aufwands.

In dieser Arbeit wird ein Record Linkage-System erforscht und entwickelt welches die Ziele hat, (1) Datenquellen-unabhängig zu sein und (2) den manuellen Prüfaufwand zu reduzieren, um die Automatisierung der Datenintegration weiter voranzutreiben. Die Kernidee des Konzepts zur Entwicklung des Record Linkage-Systems basiert auf der Einschränkung der zu integrierenden Datenquellen über eine Realwelt-Entität wie bspw. Unternehmen. Für diese Realwelt-Entität können dann die häufig vorkommenden Informationen abgeleitet werden und von diesen können die häufig vorkommenden Datenintegrationsprobleme abgeleitet werden, für die geeignete Algorithmen und Verfahren entwickelt werden können. Innerhalb von neun Feldexperimenten wurde das Konzept in ein prototypisches Record Linkage-System überführt. Das prototypische Record Linkage-System, der Unternehmen-Matcher, wurde in vier Fallstudien mit den Industriepartnern Volkswagen AG, EWE TEL GmbH, CEWE Stiftung & Co. KGaA und Oldenburgisch Ostfriesischer Wasserverband evaluiert. Die Evaluation des Unternehmen-Matcher hat gezeigt, dass dieser die Automatisierung der Datenintegration ermöglicht und diese Arbeit damit einen Beitrag zur Forschung der Datenintegration leistet.

Abstract

Companies are increasingly aware that high master data quality is crucial for successfully implementing AI and digitization projects. The often poor master data quality is caused by the fact that individual departments often create, change and delete master data decentrally in several databases. As a result, inconsistencies and master data duplicates, such as duplicate customers or suppliers, arise and exist in the respective databases. These master data are distributed over several databases, with their inconsistencies and duplicates leading to numerous process errors. Furthermore, there is no overall view of the master data, so employees can only incorporate all existing information from the distributed databases into their business activities with much manual effort.

Companies carry out data integration projects to create the necessary master data quality. Data integration involves merging master data from multiple sources and presenting it uniformly. IT experts are needed to carry out these projects, who spend an enormous amount of manual effort to identify and implement suitable algorithms. In addition, the exponential growth of data sources reinforces the need for enterprise data integration.

The existing state of the art of data integration research is hardly used in enterprises because the manual effort required to evaluate the quality of results is still too high. Existing data integration systems such as Magellan or BigGorilla provide a variety of algorithms and procedures for data integration but fail to drive automation further. According to Gartner, a potential savings of 45% of manual effort exists through AI in data integration.

In this PhD thesis, a record linkage system is explored and developed, which has the goals of (1) be data source-independent and (2) reduce manual checking effort to drive data integration automation further. The core idea for developing the record linkage system is to restrict the data sources to be integrated via a real-world entity such as a company. For this real-world entity, the commonly occurring information can then be derived, and from these, the commonly occurring data integration problems can be derived, for which suitable algorithms and procedures can be developed. The concept was transformed into a prototype record linkage system within nine field experiments. The prototypical record linkage system, the company-matcher, was tested in four case studies with the industry partners Volkswagen AG, EWE TEL GmbH, CEWE Stiftung & Co. KGaA and Oldenburgisch Ostfriesischer Wasserverband. The evaluation of the company-matcher has shown that it enables the automation of data integration and thus this work contributes to the research of data integration.

Inhaltsverzeichnis

Abkürzungen

CRF Conditional Random Fields

DI Datenintegration

EI Experteninterview

FG Fokusgruppe

FN False Negative

FP False Positive

GTIN Global Trade Item Number

HF Hauptforschungsfrage

IKT Informations- und Kommunikationssysteme

ILS Illustratives Szenario

IQ Information Quality

IS Information Systems

KI Künstliche Intelligenz

MAMBA Multiple Algorithm Matching for Better Analysis

ML Machine Learning

PC Pair Completeness

PM Potenzieller-Match

PQ Pair Quality

PT Personentage

RL Record Linkage

RR Reduction Ratio

SB Standard Blocking

SM Sicherer-Match

SN Sorted Neighborhood

SVC Support Vector Classifier

TF Teilforschungsfrage

TN True Negative

TP True Positive

UM Unternehmen-Matcher

UPC Universal Product Code

WI Wirtschaftsinformatik

Abbildungen

Tabellen

Listings

1 Einleitung

1.1 Motivation

Big Data ist ein Phänomen, das die Vielzahl der verfügbaren Datenquellen beschreibt. Jedoch wird Big Data nicht nur durch die Datenmenge (Volume) definiert. Zur vollständigen Definition von Big Data dient das Modell der 5Vs. Dies umfasst die Dimensionen Volume, Velocity, Variety, Value und Veracity (vgl. Blazquez & Domenech, 2018b, S. 1; Fasel & Meier, 2016, S. 4-6; Dong & Srivastava, 2013, S. 1245). Durch die Digitalisierung aller Lebensbereiche entsteht dieses Phänomen und lässt Daten zu einer entscheidenden Ressource für Unternehmen werden, wenn diese die Informationen in unternehmerischen Entscheidungen berücksichtigen (vgl. Gluchowski & Chamoni, 2016, S. 55-56; Blazquez & Domenech, 2018b, S. 1-2; Fasel & Meier, 2016, S. 4-5).

Um die Ressource Daten zu nutzen, werden Informationssysteme benötigt, die die Datenmengen speichern, verarbeiten und analysieren (vgl. Deloitte, 2018; Mühlroth & Grottke, 2018, S. 2; Dong & Srivastava, 2013, S. 1245). In der Entstehung von Big Data bestand die Herausforderung zunächst darin, wie relevante Daten identifiziert und gespeichert werden können. Eine fortwährende Herausforderung besteht darin, wie aus diesen Datenmengen relevante und nützliche Erkenntnisse extrahiert werden können (vgl. Blanco, Enriquez, Dominguez-Mayo, Escalona & Tuya, 2018, S. 1). Um diese Erkenntnisse in Form von Data Products in unternehmerische Entscheidungsprozesse einfließen zu lassen, wird die Disziplin Data Science genutzt (vgl. Fasel & Meier, 2016, S. 65-66; Kruse, Dmitriyev & Marx Gómez, 2018, S. 1-3; Kessler & Marx Gómez, 2020). Zur Entwicklung der Data Products existieren etablierte Vorgehensmodelle wie der CRISP-DM (vgl. Wirth & Hipp, 2000). Damit die unternehmerischen Entscheidungen auf verlässlichen und wertschöpfenden Data Products basieren, sind die zugrunde liegenden Datenquellen und deren Qualität die entscheidende Basis. Diese Basis wird im CRISP-DM Prozessschritt *Data Preparation* aufgebaut. Die *Data Preparation* umfasst u.a. das Integrieren, Bereinigen und Aufbereiten der Daten (vgl. Saluja, 2018; Christophides, Efthymiou, Palpanas, Papadakis & Stefanidis, 2021, S. 1). In Data Science Projekten werden oftmals Informationen aus verschiedenen Datenquellen benötigt, um das Informationspotenzial ausschöpfen zu können (vgl. Christen & Winkler, 2016).

Durch Big Data stehen den Unternehmen viele interne und externe Datenquellen zur Verfügung (vgl. Christophides et al., 2021, S. 1; Blazquez & Domenech, 2018a, S. 108). Die hohe Anzahl der Datenquellen liegt darin begründet, dass diese oftmals für eine bestimmte Aufgabe erstellt worden sind (vgl. Dong & Srivastava, 2013, S. 1245; Blanco et al., 2018, S. 1; Rahm, 2016). Daher können sowohl die internen als auch externen Datenquellen komplementäre, sich ergänzende oder unterschiedliche Informationen beinhalten, die gemeinsam einen Informati-

onsmehrwert liefern können (vgl. Rahm et al., 2019, S. 12; Pershina, 2016, S. 3; Lin, Wang, Li & Gao, 2016, S. 1). Dieses Potenzial der Datenquellen gilt es für Unternehmen zu nutzen, um aus den Informationen nützliche Erkenntnisse zu extrahieren. Dabei steht die Erschließung und das Nutzbarmachen von neuen externen Datenquellen in Unternehmen noch am Anfang, obwohl die Unternehmen das Potenzial dieser als sehr hoch ansehen (vgl. Meyn, Sock, Adan & Stüben, 2019; Iffert & Derwisch, 2018; Fasel & Meier, 2016, S. 51-52; Witte, Gerberding, Melching & Marx Gómez, 2021).

Bevor die internen und externen Datenquellen in Data Science Projekten gemeinsam genutzt werden können, müssen diese zunächst integriert werden (vgl. Christen & Winkler, 2016, S. 1; Rahm, 2016, S. 1). Das Ziel der Datenintegration ist es, einen einheitlichen Zugriff auf die verschiedenen Datenquellen zu ermöglichen. Durch Big Data steigen die Herausforderungen der Datenintegration (vgl. Dong & Srivastava, 2013, S. 1245; Rahm, 2016; Golshan, Halevy, Mihaila & Tan, 2017, S. 101-102). Die Herausforderungen der Datenintegration durch Big Data bestehen aus (1) der rasant wachsenden Anzahl an Datenquellen, (2) der hohen Dynamik in den Datenquellen, bedingt durch neue oder sich ändernde Daten, (3) der strukturellen und semantischen Heterogenität der Datenquellen und (4) der unterschiedlichen Datenqualität hinsichtlich Vollständigkeit, Genauigkeit und Aktualität (vgl. Dong & Srivastava, 2013, S. 1245).

Die Datenintegration umfasst die Prozessschritte (1) Schema Matching, (2) Record Linkage und (3) Data Fusion, um diese Herausforderungen zu bewältigen (vgl. Dong & Srivastava, 2015, S. 9-10). Die Prozessschritte der Datenintegration sind komplex und in vielen Fällen noch nicht automatisiert, da die Herausforderungen mit der Anzahl der zu integrierenden Datenquellen steigen. Daher basieren die meisten Datenintegrationsverfahren auf wenigen Datenquellen (vgl. Dong & Srivastava, 2013; Rahm, 2016, S. 1; Blazquez & Domenech, 2018b; González Enríquez, 2017, S. 1-2). Um das angesprochene Potenzial der integrierten Datenquellen nutzen zu können, ist das Bewältigen dieser Herausforderungen erforderlich (vgl. Dong & Srivastava, 2015, S. 1). Daher investieren Industrie und Wissenschaft viel Aufwand in die Forschung des Datenintegrationsprozesses (vgl. Rahm, 2016, S. 1; El-Ghafar, Gheith, El-Bastawissy & Nasr, 2017, S. 225).

Der wichtigste Prozessschritt der Datenintegration ist das Record Linkage (RL) (vgl. Dong & Rekatsinas, 2018; Blazquez & Domenech, 2018b; Christen & Winkler, 2016). Dong und Rekatsinas (2018) bezeichnen RL als „unavoidable and arguably the most important problem in integrating data from different sources“(Dong & Rekatsinas, 2018, S. 1646). Die Aufgabe von RL besteht in der Identifizierung von Datensätzen, die zur gleichen Realwelt-Entität gehören. Realwelt-Entitäten sind bspw. Unternehmen, Personen oder Produkte. Dabei existiert das RL-Problem seit Beginn der relationalen Datenbanken und ist durch die Herausforderungen und das Potenzial von Big Data wieder in den Fokus gerückt (vgl. Dong & Rekatsinas,

2018; Ebraheem, Thirumuruganathan, Joty, Ouzzani & Tang, 2018, S. 1454). Oftmals müssen mehrere tausend Entitäten aus verschiedenen Datenquellen verknüpft werden, weshalb die manuelle Durchführung von RL aufgrund des hohen Aufwandes selten zielführend ist. Für die Automatisierung des Datenintegrationsprozesses, insbesondere RL, wird deshalb versucht vermehrt Machine Learning (ML) einzusetzen (vgl. Dong & Rekatsinas, 2018, S. 1645-1646; Barlaug & Atle Gulla, 2020; Behnen, Kruse & Marx Gómez, 2021). Der Trend der Automatisierung des Datenintegrationsprozesses durch ML oder Künstliche Intelligenz (KI) ist neben der Wissenschaft auch bei den Unternehmen angekommen. Gartner bezeichnen den Einsatz von ML und KI in der Datenintegration als „augmented data integration“ und prognostizieren, dass dadurch bis Ende 2022 bis zu 45% des manuellen Aufwands reduziert werden kann (vgl. Gartner, 2021).

1.2 Problemstellung

RL wurde 1959 von Newcombe, Kennedy, Axford und James definiert und zehn Jahre später von Fellegi und Sunter formalisiert, seit dieser Zeit wird RL erforscht (vgl. Ebraheem et al., 2018, S. 1). Dennoch gibt es durch die Herausforderungen von Big Data weiteren Forschungsbedarf, da die Unternehmen vermehrt Interesse daran haben, die neuen, potenziell nützlichen internen und externen Datenquellen für Analysezwecke zu integrieren (vgl. Mudgal et al., 2018, S. 30; Christen, 2012a, S. 3)

RL wird benötigt, wenn Datenquellen integriert werden sollen, die keine gemeinsame Identifikationsnummer besitzen. Häufig ist dies im Kontext von Big Data bei internen und externen Datenquellen der Fall. In der Industrie gibt es einige Bestrebungen global standardisierte Identifikationsnummern zu schaffen, wie beispielsweise für Produkte mit dem Universal Product Code (UPC) oder der Global Trade Item Number (GTIN) (vgl. Schmidt, 2010, S. 49-50). Dennoch sind diese Identifikationsnummern oftmals nicht in allen Datenquellen vorhanden oder vollständig, sodass diese nicht als verknüpfendes Attribut ausreichen (vgl. Köpcke, Thor, Thomas & Rahm, 2012, S. 546; Jupin & Shi, 2014, S. 1; Schild & Schultz, 2017, S. 1-2). Um die Datenquellen dennoch integrieren zu können, wird der RL-Prozess angewandt. Der RL-Prozess beschreibt die Integration von zwei strukturierten Datenquellen und besteht aus den Prozessschritten (1) Data Preparation, (2) Blocking, (3) Record Pair Comparison, (4) Classification und (5) Evaluation (vgl. Christen, 2012a, S. 24). Im RL-Prozess werden die verfügbaren Attribute der Entitäten herangezogen, um die Ähnlichkeit zwischen den Datensätzen zu bestimmen. Bei der Entität Produkt könnten dies beispielsweise die Modellnummer, die Größe oder der Hersteller sein (vgl. Talburt, 2011, S. 1).

RL ist eine herausfordernde Aufgabe, wenn Entitäten über ihre beschreibenden Attribute verglichen werden müssen, da in den einzelnen Datenquellen beispielsweise duplizierte, un-

vollständige oder fehlerhafte Datensätze vorliegen können (vgl. Köpcke, 2014 S. 5; Kooli, Allesiardo & Pigneul, 2018 S. 3; Enríquez et al., 2015 S. 3; Kong, Gao, Xu, Qian & Zhou, 2016 S. 2). Probleme stellen auch die strukturelle und semantische Heterogenität zwischen den Datenquellen dar. Dabei kann die Heterogenität auf Schema- und auf Datenebene auftreten. Beispielsweise können die Entitäten durch eine unterschiedliche Anzahl von Attributen repräsentiert werden oder es können auf Schema- oder Datenebene Probleme wie strukturelle Konflikte, Synonyme oder Homonyme existieren (vgl. Bleiholder & Schmid, 2018, S. 121-122; Rahm & Hai Do, 2000, S. 4-5). Diese Probleme steigen mit der Anzahl und der vielfältigen Struktur der zu integrierenden Datenquellen und müssen in dem jeweiligen RL-Prozess und den einzelnen RL-Prozessschritten bewältigt werden (vgl. Rahm, 2016, S. 1; Kruse, 2019).

Mit dem zunehmenden Einsatz von ML und der Entwicklung von RL als Teilaufgabe der Data Science wurden viele RL-Forschungsergebnisse publiziert. Dabei haben die meisten Arbeiten Algorithmen für den RL-Prozess entwickelt, was die Vielzahl der vorhandenen Lösungsmöglichkeiten begründet. Ein weiterer Teil der Arbeiten hat sich mit der Entwicklung von RL-Systemen befasst (vgl. Govind et al., 2019, S. 3). Für den RL-Prozessschritt *Classification* wurden zunächst regelbasierte Verfahren entwickelt. Darauf folgten unsupervised ML-Verfahren und supervised ML-Verfahren, wie Entscheidungsbäume, die logistische Regression oder die Support Vector Machine. Seit 2018 werden auch vermehrt Deep Learning Verfahren für RL in Publikationen betrachtet (vgl. Dong & Rekatsinas, 2018, S. 1646; Köpcke, 2014, S. 97-105; Köpcke & Rahm, 2010; Mudgal et al., 2018, S. 19-20; Ebraheem et al., 2018, S. 1454). Für den Prozessschritt *Record Pair Comparison* existieren für jeden Datentyp, wie Zeichenketten, Datumsangaben oder numerische Werte, verschiedene Ähnlichkeitsmaße. Zum Vergleich von Zeichenketten existieren beispielsweise im RL-System Magellan[1] 23 verschiedene Ähnlichkeitsmaße, wie Levenshtein, Jaro-Winkler, Jaccard oder Tf-idf (vgl. Konda et al., 2016b; Kooli et al., 2018, S. 4; Peng, Li & Kennedy, 2012, S. 1). Auch die stetige Entwicklung interoperabler RL-Tools, bedingt durch die Entwicklung von monolithischen RL-Systemen hinzu einem RL-Ecosystem, fördert die steigende Anzahl an Algorithmen für die RL-Prozessschritte (vgl. Govind et al., 2019, S. 8).

Aus dieser Vielzahl an Algorithmen in den jeweiligen Prozessschritten muss ein Data Scientist wählen. Dabei ist die Auswahl aktuell immer für zwei zu integrierende Datenquellen zu treffen (vgl. Ebraheem et al., 2018 S. 1454 - 1455; Theodoros I. Rekatsinas, Xin Dong, Lise Getoor & Divesh Srivastava, 2015 S. 1; Köpcke, 2014 S. 14; Peng et al., 2012 S. 1). Da oftmals mehr als zwei Datenquellen in einem Data Science Projekt integriert werden und eine Vielzahl von Data Science Projekten in Unternehmen existieren, ist die Auswahl der Algorithmen eine wiederkehrende Aufgabe. Auch für einen Data Scientist ist diese wiederkehrende Auswahl von Verfahren und Algorithmen innerhalb des RL-Prozesses nicht trivial. RL-Systeme wie

[1] `https://sites.google.com/site/anhaidgroup/projects/magellan`

Magellan beinhalten einen Leitfaden, um die Experten im RL-Prozess zu unterstützen (vgl. Konda et al., 2016b). Doch in der Anwendung des Systems mit dem Leitfaden zur Lösung von realweltlichen Problemen zeigte sich der Leitfaden als nicht zielführend, da die RL-Probleme vor allem im realweltlichen Kontext sehr verschieden waren und die Nutzer viel experimentierten, um zu einer Lösung zu kommen (vgl. Govind et al., 2019, S. 7-12). Im RL-Prozess existieren viele wiederkehrende manuelle Aufgaben, um einen RL-Prozess für die Integration von zwei Datenquellen zu entwickeln (vgl. Ebraheem et al., 2018, S. 1454).

Neben dem manuellen Aufwand zur Entwicklung eines RL-Prozesses existiert zudem manueller Aufwand bei der Bewertung der Ergebnisse eines RL-Prozesses. Für eine vollständige Evaluation und Bewertung der Ergebnisqualität werden Ground Truth Daten benötigt. Ground Truth Daten für einen RL-Prozess lassen sich nur über den manuellen Vergleich des Kreuzproduktes der beiden zu integrierenden Datenquellen erstellen. Bei zwei Datenquellen, die jeweils 100 Datensätze enthalten, bedeutet dies 10.000 Vergleiche. Wenn für die Prüfung eines Datensatzpaares im Durchschnitt 30 Sekunden benötigt werden, beträgt der Aufwand insgesamt 10,4 Personentage (PT). Gleichzeitig bedeutet das Erstellen einer Ground Truth Datenmenge das manuelle Durchführen des RL-Prozesses und steht damit der gewünschten Automatisierung des RL-Prozesses gegenüber. Daher werden die klassifizierten Ergebnisse des RL-Prozesses häufig stichprobenartig geprüft, was manuellen Aufwand bedeutet (vgl. Christen, 2012a; Köpcke, 2014). Laut Barlaug und Atle Gulla (2020) ist der manuelle Aufwand zur Bewertung der Ergebnisqualität ein Grund dafür, warum der State-of-the-Art der RL-Forschung bisher kaum in Unternehmen eingesetzt wird.

Daher existiert Forschungsbedarf wie der RL-Prozess weiter automatisiert und der Data Scientist unterstützt werden kann, um externe unternehmensrelevante Datenquellen zu integrieren (vgl. Konda, Subramanian Seshadri, Segarra, Hueth & Doan, 2019, S. 499; Kruse, Hassan, Awick & Marx Gómez, 2020).

1.3 Ziel und Forschungsdesign

Das Ziel der Arbeit leitet sich aus der in Abschnitt 1.2 beschriebenen Problemstellung ab. Diese hat aufgezeigt, dass RL für die Datenintegration in Data Science Projekten ein entscheidender und wiederkehrender Prozessschritt ist. Probleme bereiten die verschiedenartigen zu integrierenden Datenquellen und die hohe Anzahl existierender Algorithmen für die RL-Prozessschritte. Um diese Probleme zu lösen, soll ein RL-System konzeptioniert und prototypisch implementiert werden. Zum Einen soll dieses dabei helfen, die Datenintegrationsprobleme zu klassifizieren. Zum Anderen soll dieses geeignete RL-Verfahren und Algorithmen bereitstellen, um die identifizierten Datenintegrationsprobleme zu lösen. Damit dieses Ziel erreicht wird, ist das folgende Forschungsdesign entwickelt worden:

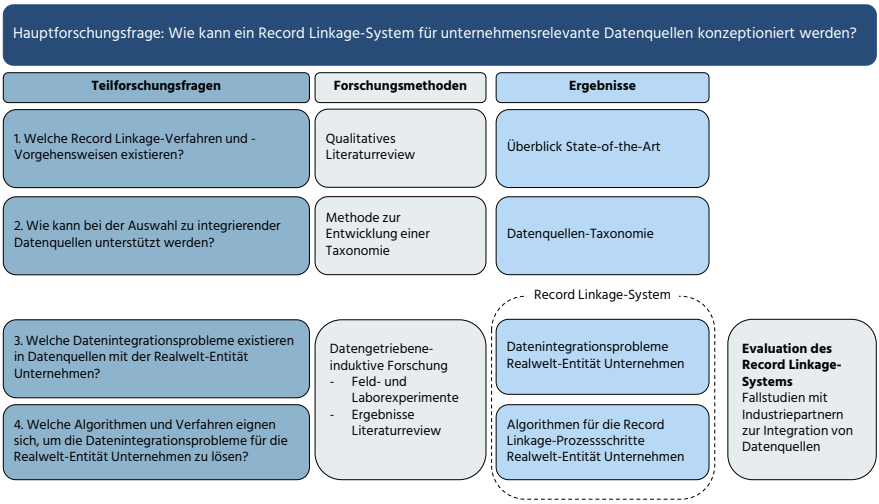


Abbildung 1.1: Forschungsdesign der Dissertation - eigene Darstellung

Hauptforschungsfrage (HF): Wie kann ein Record Linkage-System für unternehmensrelevante Datenquellen konzeptioniert werden?

Um die Hauptforschungsfrage beantworten zu können, sind die drei folgenden Teilforschungsfragen abgeleitet worden:

Teilforschungsfrage 1 (TF1): Welche Record Linkage-Verfahren und -Vorgehensweisen existieren?

Für die Beantwortung der TF1 soll eine qualitative Literaturanalyse durchgeführt werden. Mit dieser soll ein Überblick über den aktuellen Stand der Forschung im Bereich RL erstellt werden. Zudem soll ein Überblick über die eingesetzten Verfahren und Algorithmen, mit denen Datenquellen integriert werden können, erstellt werden.

Teilforschungsfrage 2 (TF2): Wie kann bei der Auswahl zu integrierender Datenquellen unterstützt werden?

Für die Beantwortung der TF2 soll eine Datenquellen-Taxonomie entwickelt werden. Dazu wird die Methode zur Entwicklung einer Taxonomie von Nickerson, Varshney und Muntermann (2013) verwendet.

Teilforschungsfrage 3 (TF3): Welche Datenintegrationsprobleme existieren in Datenquellen mit der Realwelt-Entität Unternehmen?

Für die Beantwortung von TF3 sollen die häufig vorkommenden Datenintegrationsprobleme für die Realwelt-Entität Unternehmen identifiziert werden. Durch die mit TF2 entwickelte Datenquellen-Taxonomie können die relevanten Datenquellen, die die Realwelt-Entität Unternehmen enthalten, identifiziert werden. Mit den identifizierten Datenquellen werden mit der datengetriebenen-induktiven Forschung die Datenintegrationsprobleme identifiziert.

Teilforschungsfrage 4 (TF4): Welche Algorithmen und Verfahren eignen sich, um die Datenintegrationsprobleme für die Realwelt-Entität Unternehmen zu lösen?

Die Ergebnisse aus TF1 liefern den aktuellen Stand der Forschung von Record Linkage-Verfahren und -Algorithmen. Die Ergebnisse aus TF3 liefern die Datenintegrationsprobleme für die Realwelt-Entität Unternehmen. Auf Basis dieser Ergebnisse sollen zunächst zur Beantwortung der TF4 Feldexperimente durchgeführt werden, um Algorithmen und Verfahren auszuwählen und zu implementieren. Mit den implementierten Algorithmen und Verfahren wird dann ein RL-Prozess entwickelt, der die Datenintegrationsprobleme für die Realwelt-Entität Unternehmen löst.

Die Evaluation des entwickelten RL-Systems, das einen RL-Prozess für die Realwelt-Entität Unternehmen enthält, wird einer Fallstudie mit Industriepartnern evaluiert. Innerhalb der Fallstudie wird das RL-System auf Datenquellen der Industriepartner angewendet und von den Experten der Industriepartner qualitativ bewertet.

1.4 Einordung in das Forschungsgebiet der Wirtschaftsinformatik

„Gegenstand der Wirtschaftsinformatik (WI) sind Informationssysteme (IS), oft auch als Informations- und Kommunikationssysteme (IKS) bezeichnet, in Wirtschaft, öffentlicher Verwaltung und im Privathaushalt. IS umfassen menschliche und maschinelle Komponenten (Teilsysteme)“(vgl. Mertens, 2019). Die Konzeption, Entwicklung, Einführung, Nutzung und Wartung von Anwendungssystemen sowie generell das Management des Produktionsfaktors Information stehen im Mittelpunkt der WI. Dabei versteht sich die WI als interdisziplinäres Fach zwischen der Betriebswirtschaftslehre und der Informatik (vgl. Mertens, 2019).

Die Wirtschaftsinformatik ist eine Wissenschaft, „die sich mit der Beschreibung, Erklärung, Prognose und Gestaltung rechnergestützter Informationssysteme und deren Einsatz in Wirtschaft, Verwaltung und zunehmend dem unmittelbaren privaten Lebensumfeld befasst. Sie versteht sich als eigenständiges interdisziplinäres Fach im Wesentlichen zwischen Betriebswirtschaftslehre und Informatik“(vgl. Laudon, Laudon & Schoder, 2016, S. 57). Dabei fokussiert

die Wirtschaftsinformatik traditionell die praktische Relevanz von Erkenntnissen. Daher werden vornehmlich Lösungen für Unternehmen entwickelt, die Informationstechnik sowie Hard- und Software für diverse betriebliche Aufgaben bestmöglich einsetzen (vgl. Robra-Bissantz & Strahringer, 2020, S. 163).

In der Wirtschaftsinformatik können zwei Forschungsrichtungen unterschieden werden (vgl. Wilde & Hess, 2006, S. 1-3; Laudon et al., 2016, S. 60-62; Robra-Bissantz & Strahringer, 2020, S. 164-165):

Behavioral Science: Die Behavioral Science verfolgt traditionell nach der englischsprachigen Forschung des Information Systems (IS) Research die empirische, verhaltensorientierte Forschung. Hauptziel ist es, das Verhalten und die Auswirkungen von Informationssystemen auf Organisationen oder Menschen eher sozialwissenschaftlich zu analysieren. Dabei fokussiert sie nicht ausschließlich das Lösen von Problemen aus der Praxis.

Design Science: Die Design Science Research wird im deutschsprachigen Raum als gestaltungsorientierte Wirtschaftsinformatik bezeichnet. Hauptziel ist es mit wissenschaftlicher Güte Modelle, Methoden oder Systeme zu entwickeln, die die unternehmerischen Aufgaben und die beteiligten Menschen berücksichtigen. Daher fokussiert die Design Science die Forschung für die Praxis und ermöglicht dabei die Verknüpfung von praktischer Nutzbarkeit und wissenschaftlichem Erkenntnisgewinn.

Dabei erfolgt der Erkenntnisgewinn in der Wirtschaftsinformatik oftmals nach dem Muster, dass (Referenz-) Theorien aus Referenzdisziplinen auf einen bestimmten Informationssystem-Kontext angewendet werden. Dadurch entstehen Erkenntnisse zwischen der hohen Abstraktionsebene, die (Referenz-) Theorien enthält, und der niedrigen Abstraktionsebene, die die konkreten Daten des betrachteten Kontextes abbildet. Grover und Lyytinen bezeichnen die Ergebnisse dieser Forschung als Midrange-Theorien (vgl. Grover & Lyytinen, 2015, S. 271-274; Robra-Bissantz & Strahringer, 2020, S. 184). Zur Weiterentwicklung der Wirtschaftsinformatik-Forschung sollte laut Grover und Lyytinen mehr Forschung auf der hohen Abstraktionsebene, die mutige Theorieentwicklung (Blue Ocean), und auf der niedrigen Abstraktionsebene, die datengetriebene Forschung (Data Driven), stattfinden (vgl. Grover & Lyytinen, 2015, S. 285; Robra-Bissantz & Strahringer, 2020, S. 184). Die mutige Theorieentwicklung soll Nutzen für die Forschung bringen, indem originelle und disziplinspezifische Theorien entwickelt werden. Die datengetriebene Forschung soll die Nutzung und den Nutzen in der Praxis stärken, da durch datengetriebene induktive Forschung, interessante, neue und praxisrelevante Phänomene entdeckt werden können (vgl. Robra-Bissantz & Strahringer, 2020, S. 184-185). Grover und Lyytinen definieren die datengetriebene Forschung wie folgt: „Data-driven, inductive research emphasizes the interpretation of patterns inherent in (new) empirical data. It aims to discern regularized empirical patterns of importance, and thereby

to influence the intellectual framing of what we need to know "(vgl. Grover & Lyytinen, 2015, S. 285). Der Einsatz von IT ist ein dynamisches Feld, in dem der Forscher, unabhängig von bereits existierenden Theorien, ein aufmerksamer Beobachter der Praxis sein sollte. Die Beobachtung, Beschreibung und Identifizierung von Mustern sind gültige Forschungsmethoden und führen zu einer robusten Theoriebildung (vgl. Grover & Lyytinen, 2015, S. 287). Die Autoren Grover und Lyytinen und Robra-Bissantz und Strahringer zeigen in ihren Beiträgen, dass die Wirtschaftsinformatik Forschung und Praxis in praxisrelevanter Forschung vereinen kann (vgl. Robra-Bissantz & Strahringer, 2020, S. 185-186).

Diese Dissertation folgt dem Design Science Ansatz. Gemäß den Prinzipien der Design Science, liefert diese Dissertation Erkenntnisgewinn für Experten aus der Wissenschaft und der Praxis, die sich mit der Datenintegration und im speziellen mit RL beschäftigen. Da in dieser Dissertation die datengetriebene induktive Forschung eingesetzt werden soll, die laut Robra-Bissantz und Strahringer (2020) verstärkt eingesetzt werden sollte, wird großer Erkenntnisgewinn für die Praxis gewonnen.

Für die Praxis ergibt sich der Nutzen aus dem in dieser Arbeit entwickelten RL-System für die Datenquellen-unabhängige und nahezu vollständig automatisierte Integration von Datenquellen mit der Realwelt-Entität Unternehmen. Bisher hat keine dem Autor bekannte Arbeit den State-of-the-Art aus der RL-Forschung auf die realen RL-Probleme der Unternehmen, wie sie Barlaug und Atle Gulla (2020) aufführt, übertragen und versucht zu lösen. In dieser Dissertation wird dies mit dem Konzept und der prototypischen Implementierung des RL-Systems erstmalig erforscht. Weiterhin wurde der Datenintegrationsprozess um einen für Unternehmen wichtigen Prozessschritt, die Datenquellenauswahl, erweitert und eine Datenquellen-Taxonomie entwickelt, die Unternehmen und Wissenschaft in diesem Prozessschritt unterstützen kann.

Für die Wissenschaft leistet diese Dissertation Erkenntnisgewinn im Bereich der Datenintegration und im speziellen des RL. Bisher wurde durch keine dem Autor bekannte Arbeit der Begriff RL-System definiert. Somit wird der Begriff RL-System in dieser Dissertation erstmalig definiert. Weiterhin liefert die Dissertation einen Überblick über den State-of-the-Art der RL-Forschung. Durch die Übertragung der wissenschaftlichen Konzepte und Algorithmen auf die praktischen RL-Probleme zeigt diese Arbeit einen neuen Ansatz für die RL-Forschung, wie Datenquellen-unabhängige RL-Prozesse entwickelt und der Automatisierungsgrad durch RL-Systeme erhöht werden kann.

2 Grundlagen der Datenintegration

2.1 Datenintegration

Den Unternehmen stehen große Datenmengen mit unternehmensrelevantem Bezug zur Verfügung, die stetig anwachsen. Dabei handelt es sich um strukturierte und unstrukturierte Datenquellen, die intern als auch extern vorliegen können. Das Nutzen dieser Datenmengen, um Informationsmehrwerte zu generieren und diese in Entscheidungsprozessen zu berücksichtigen, wird immer entscheidender für den wirtschaftlichen Erfolg von Unternehmen (vgl. Heinrich & Stühler, 2018, S. 77-78; Kölbl, Mühlroth, Wiser, Grottke & Durst, 2019, S. 1-2; Dong & Srivastava, 2015, S. 1; Stonebraker & Ilyas, 2018, S. 3-4).

Die Information Value Chain beschreibt die Abfolge dieser zyklischen Aktivitäten: (1) Daten zu Informationen, (2) Informationen zu Wissen, (3) das Wissen wird genutzt um Entscheidungen zu treffen, die wiederum (4) zu einer Aktion in der realen Welt führen, um einen Mehrwert zu erzielen (siehe Abb. 2.1).

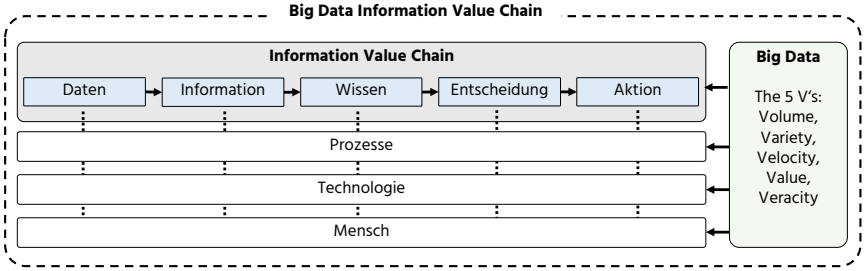


Abbildung 2.1: Big Data Information Value Chain - eigene Darstellung in Anlehnung an (vgl. Abbasi et al., 2016)

Entlang der gesamten Aktivitäten sind Menschen, Prozesse und Technologien beteiligt (vgl. Abbasi et al., 2016, S. 3). Durch den Einfluss von Big Data wird die Information Value Chain als Big Data Value Chain definiert (vgl. Abbasi et al., 2016). Abbasi et al. definieren Big Data mit den vier V's: Volume, Variety, Velocity und Veracity (siehe Abb. 2.1) (vgl. Abbasi et al., 2016, S. 5-7). Big Data führt zur steigenden Komplexität des Verwaltens, Speicherns und Integrierens von Daten (vgl. Abbasi et al., 2016, S. 15). Wesentliche Kernfragen in den Schritten Daten und Information der Big Data Value Chain beziehen sich darauf, welche internen oder externen Datenquellen integriert und analysiert werden können und welche Unternehmensinfrastruktur dazu benötigt wird (vgl. Abbasi et al., 2016, S. 6).

Die Datenintegration (DI) ist eine entscheidende Aufgabe in der Big Data Value Chain, um

relevante Informationen in unternehmerischen Entscheidungen nutzbar machen zu können. Die benötigten Informationen befinden sich oftmals in verschiedenen Datenquellen, sodass diese nicht in einer Analyse gemeinsam verwendet werden können. Um die Potenziale der Daten zu nutzen, sollten die Daten integriert werden (vgl. Dong & Srivastava, 2015, S. 1). Die Data Integration (DI) befasst sich mit der Integration von Daten, die aus verschiedenen Datenquellen stammen. In Tabelle 2.1 sind Definitionen für den Begriff DI aufgeführt und werden im Folgenden diskutiert.

Tabelle 2.1: Definitionen Data Integration

Autor	**Definition**
Doan, Halevy & Ives, 2012, S. 6	„[...] the goal of a data integration system is to offer uniform access to a set of autonomous and heterogeneous data sources.“
Christen, 2012a, S. 6-7	„[...] data integration is the overall process of integrating heterogeneous databases, datawarehouses or data repositories to provide a unified view of the available data.“
Dong & Srivastava, 2015, S. 2	„Data integration has the goal of providing unified access to data residing in multiple, autonomous data sources. While this goal is easy to state, achieving this goal has proven notoriously hard, even for a small number of sources that provide structured data—the scenario of traditional data integration.“
Rahm, 2016, S. 11	„Data integration aims at providing uniform access to data from multiple sources. It has become a pervasive task for data analysis in business and scientific applications. The most popular data integration approaches such as data warehouses or big data platforms utilize a physical data integration where the source data is combined within a new dataset or database tailored for analysis tasks.“
Golshan et al. (2017)	„At the outset, the goal of data integration was to build systems that provide a uniform query (and for the ambitious, update) interface to multiple databases within an enterprise. As the Web came into prominence, data integration addressed challenges such as incorporating data from external business partners, data exchange between multiple sources, peer-to-peer data sharing architectures, and querying multiple deep-web data sources.“

Doan et al. (2012) definieren DI als ein System mit dem Ziel, einen einheitlichen Zugriff auf autonome und heterogene Datenquellen zu ermöglichen.

Christen (2012a) definiert DI als einen ganzheitlichen Prozess, um eine einheitliche Sicht auf die verfügbaren Datenquellen wie heterogene Datenbanken, Data Warehouses oder Daten-Repositories zu erstellen. Die beiden Definitionen unterscheiden sich darin, dass Doan et al. DI als ein System und Christen DI als einen Prozess versteht.

Dong und Srivastava (2015) definieren DI ebenfalls mit dem Ziel, einen einheitlichen Zugriff auf Daten zu ermöglichen, die sich in vielen autonomen Datenquellen befinden. Zusätzlich unterscheidet sich diese Definition von den anderen, da sie anführen, dass das DI Ziel einfach zu formulieren ist, aber in der Realität notorische Probleme bei der Umsetzung existieren. Zudem führen sie an, dass auch die Integration von wenigen und strukturierten Datenquellen, der traditionellen DI, eine schwierige Aufgabe ist.

Rahm (2016) definiert das Ziel der DI ebenfalls als das Bereitstellen eines einheitlichen Zugriffs auf Daten aus verschiedenen Datenquellen. Weiterhin definiert Rahm die DI als eine weitverbreitete Aufgabe für Datenanalysen in unternehmerischen als auch in wissenschaftlichen Anwendungsfällen. Als Zielsysteme der DI nennt Rahm Ansätze wie bspw. das Data Warehouse oder Big Data Plattformen, in denen die integrierten Datenquellen physisch und für Analysezwecke optimiert gespeichert werden.

Golshan et al. (2017) definieren DI als Systeme, die eine einheitliche Abfrage auf mehrere Datenbanken innerhalb eines Unternehmens bieten. Sie erweitern in ihrer Definition die DI Herausforderung, da nicht nur interne Datenquellen sondern auch externe Datenquellen wie bspw. von Geschäftspartnern oder von Datenaustauschplattformen vermehrt integriert werden sollen.

Für diese Arbeit soll die folgende Definition für DI gelten, die sich an alle vier Definitionen anlehnt:

Defintion 2.1

Data Integration (DI) beschreibt einen Prozess, um verschiedene, heterogene interne und externe Datenquellen miteinander zu verbinden. Das Ziel ist es, einen einheitlichen Zugriff auf die Datenquellen für weitere Analysen bereitzustellen.

2.1.1 Herausforderungen der Datenintegration

Damit die entstehenden Mehrwerte der integrierten Datenquellen genutzt werden können, sollten die Herausforderungen der Datenintegration gelöst werden (vgl. Dong & Srivastava, 2015, S. 1). Die Herausforderungen bei der Datenintegration entstehen dadurch, dass die

Datenquellen (1) verteilt, (2) autonom, (3) heterogen und oftmals in einer (4) Vielzahl vorhanden sind. Das Problem der physischen Verteilung der Datenquellen auf mehreren Servern kann durch die zunehmend bessere Vernetzung vernachlässigt werden. Die größeren Herausforderungen stellen die Autonomie, Heterogenität und Vielzahl der Datenquellen dar. Die Autonomie der Datenquellen beschreibt, dass diese oftmals von verschiedenen Personen und Organisation innerhalb und außerhalb von Unternehmen erstellt und gepflegt werden. Die Autonomie der Datenquellen führt zur Heterogenität der Datenquellen und steigert diese. Die Heterogenität der Datenquellen kann in die technische, strukturelle und semantische Heterogenität unterteilt werden (vgl. Schmidt, 2010, S. 16; vgl. Doan et al., 2012, S. 6-8; Hildebrand, Gebauer, Hinrichs & Mielke, 2018, S. 121-122; Schildgen & Deßloch, 2016, S. 7-8; Dong & Srivastava, 2015, S. 8-9):

Technische Heterogenität: Die technische Heterogenität beschreibt Hard- und Softwareunterschiede der Systeme. Beispielsweise können Datenquellen in relationalen Datenbanken, in No-SQL Datenbanken oder in Dateien auf einem Server vorliegen. Die Systeme können unterschiedliche Schnittstellen wie REST, JDBC oder Webservices besitzen. Die Abfragesprache innerhalb der Schnittstelle kann verschieden sein wie bspw. SQL, SPARQL oder XQuery. Die Datenaustauschformate können CSV, JSON, XML, Text oder binär sein.

Strukturelle Heterogenität: Ein Anwendungskonzept wie Personen oder Unternehmen können durch viele verschiedene Möglichkeiten in einem Datenmodell umgesetzt werden. Wird dasselbe Anwendungskonzept in zwei Systemen unterschiedlich modelliert, spricht man von struktureller Heterogenität. Dies kann bedeuten, dass eine unterschiedliche Anzahl von Entitäten modelliert ist, sodass ein unterschiedlicher Normalisierungsgrad vorliegt. Auch die Attribute der Entitäten können in einer unterschiedlichen Anzahl modelliert worden sein. Die strukturelle Heterogenität wird in der Literatur auch als schematische Heterogenität bezeichnet.

Semantische Heterogenität: Die semantische Heterogenität kann auf Schema- und Datenebene auftreten. Auf Schemaebene können für dasselbe Konzept bspw. Synonyme, fremdsprachige Übersetzungen oder Homonyme für semantische Heterogenität sorgen. Auf Datenebene kann diese bspw. durch Abkürzungen, unterschiedliche Datumsformate und Währungsangaben entstehen. Beispielsweise können Unternehmensnamen in der einen Datenquellen abgekürzt sein, bspw. VW, und in der anderen Datenquelle ist der Unternehmensname ausgeschrieben, bspw. Volkswagen.

Neben der Heterogentität ist die mögliche Inkonsistenz der Daten in den Datenquellen eine weitere Herausforderung bei der Integration. Beispielsweise ist die Anzahl der Mitarbeiter für

die Entität Volkswagen in Datenquelle A mit 450.000 angegeben und in Datenquelle B ist die Mitarbeiteranzahl für die gleiche Entität mit 652.300 angegeben. Es gilt zu lösen, welcher Wert der korrekte ist (vgl. Dong & Srivastava, 2015, S. 9).

Die Vielzahl der existierenden und neu entstehenden Datenquellen führt zu mehr technischer, struktureller und semantischer Heterogenität, da die Datenquellen oftmals autonom entstehen und betrieben werden. Auf der einen Seite bietet die Vielzahl der existierenden und entstehenden Datenquellen potenziell wertvolle Informationen für unternehmerische Entscheidungen. Auf der anderen Seite steigen die Herausforderungen der Datenintegration, die bewältigt werden müssen, um die Informationsmehrwerte in Entscheidungsprozessen nutzen zu können. (vgl. Doan et al., 2012, S. 6-8; Hildebrand et al., 2018, S. 121-122).

2.1.2 Datenintegrationsprozess

Die Herausforderungen der Datenintegration sollen mit Hilfe des Datenintegrationsprozesses bewältigt werden. Der Datenintegrationsprozess besteht aus den drei Prozessschritten (1) Schema Matching, (2) Record Linkage und (3) Data Fusion (s. Abb. 2.2) (vgl. Dong & Srivastava, 2015, S. 9).

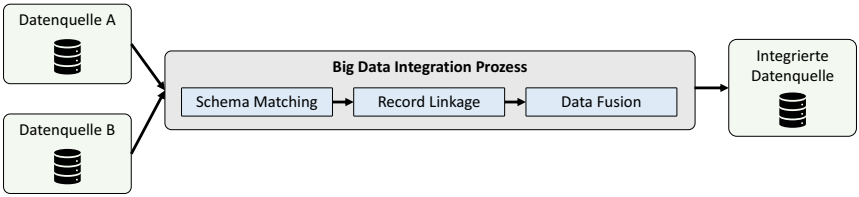


Abbildung 2.2: Big Data Integration Prozess - eigene Darstellung in Anlehnung an (Dong & Srivastava, 2015)

Im Folgenden wird der Datenintegrationsprozess vorgestellt. Da der Fokus dieser Arbeit auf dem Prozessschritt Record Linkage liegt, wird dieser in Abschnitt 2.2 ausführlich behandelt.

Zunächst werden im Data Integration Prozess zwei Datenquellen ausgewählt, die miteinander integriert werden sollen. Im Prozessschritt Schema Matching werden die Tabellen und Attribute aus den beiden Datenquellen identifiziert, die die gleichen Informationen beinhalten. Ein Beispiel für das Schema Matching ist in Abbildung 2.3 dargestellt.

Abbildung 2.3 zeigt zwei Datensätze, die jeweils die Volkswagen AG repräsentieren. Die Volkswagen AG wird mit unterschiedlich normalisierten und benannten Attributen repräsentiert, sodass die zu vergleichenden Attribute, wie bspw. Firma und Name, identifiziert werden müssen.

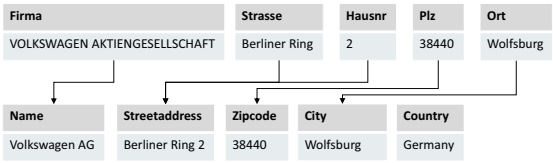


Abbildung 2.3: Record Linkage-Prozessschritt Schema Matching

Im Prozessschritt Record Linkage, welcher in dieser Arbeit fokussiert wird, werden die Datensätze aus den beiden Datenquellen identifiziert, die die gleiche Realwelt-Entität repräsentieren. Wenn die Datensätze dieselbe Realwelt-Entität repräsentieren, wird dies als Match bezeichnet. Repräsentieren die Datensätze nicht dieselbe Realwelt-Entität, wird dies als Kein-Match bezeichnet.

Der Prozessschritt Data Fusion hat das Ziel die als Match klassifizierten Datensätze in einen aufbereiteten und konsistenten Datensatz zu überführen, der die entsprechende Realwelt-Entität repräsentiert (vgl. Christen, 2012a, S. 3-4). Die größte Herausforderung der Data Fusion ist das Auflösen von Konflikten zwischen den Datensätzen, wenn bspw. die gleichen Attribute der Datensätze unterschiedliche Werte beinhalten (vgl. Christen, 2012a, S. 8).

2.2 Record Linkage

Record Linkage (RL) stellt den zweiten Schritt des Datenintegrationsprozesses nach Dong und Srivastava (2015) dar. In der folgenden Tabelle 2.2 werden verschiedene Definitionen zu RL aufgeführt und anschließend diskutiert.

Tabelle 2.2: Definitionen Record Linkage

Autor	**Definition**
Christen (2012b), S. 1537	„Record linkage is the process of matching records from several databases that refer to the same entities. When applied on a single database, this process is known as deduplication.“
Köpcke (2014), S. 21	„[Record Linkage] is the process of determining whether two objects are referring to the same entity or two different entities. Objects to be resolved may reside in distributed, typically heterogeneous data sources or may be stored in a single data source, e.g., in a database or search engine store.“

Doan et al. (2020), S. 83	„Entity matching (EM) finds data instances that refer to the same real-world entity, such as tuples (David Smith, UW-Madison) and (D. Smith, UWM). This problem, also known as entity resolution, record linkage, deduplication, data matching, et cetera, has been a long-standing challenge in the database, AI, KDD, and Web communities.“

Die Definition von Christen (2012b) definiert RL als einen Prozess, der Datensätze aus verschiedenen Datenquellen zur selben Entität integriert. Zudem wird mit Deduplication ein RL Synonym aufgeführt. Der Begriff Deduplication wird verwendet, wenn der RL-Prozess mit einer Datenquelle durchgeführt wird.

Köpcke (2014) definieren RL ebenfalls als Prozess. Ziel des Prozesses ist es, zu bestimmen, ob zwei Objekte zur selben oder zu unterschiedlichen Entitäten gehören. Köpcke (2014) legt den Fokus darauf, dass der RL-Prozess auf einer oder mehreren verteilten und heterogenen Datenquellen ausgeführt wird.

Doan et al. (2020) definieren RL als Vorgehen, das Datensätze identifiziert, die zur selben Realwelt-Entität gehören. Weiterhin führen sie Synonym verwendete Begriffe wie Entity Matching, Entity Resolution, Deduplication und Data Matching auf. In ihrer Definition betonen sie, dass RL eine langjährige Herausforderung in den Domänen der Künstlichen Intelligenz oder der Webentwicklung darstellt.

Für diese Arbeit soll die folgende Definition für RL gelten, die sich an den drei zuvor dargestellten Definitionen anlehnt:

Defintion 2.2

Record Linkage (RL) bezeichnet den Prozess, Datensätze in heterogenen und verteilten Datenquellen zu identifizieren, die zur selben Realwelt-Entität gehören.

Reocrd Linkage hat eine weit zurückreichende Historie, da bereits vor der Innovation der modernen Computer Statistiker und Forscher interessiert daran waren, Datensätze aus verschiedenen und verteilten Datenquellen, die zur selben Realwelt-Entität gehören, zusammenzuführen (vgl. Christen, 2012a, S. 9). Das erste Mal wurde der Begriff RL in der Publikation von Dunn (1946) im Jahr 1946 genutzt. Die Publikation von Dunn (1946) behandelt die Idee eines „book of life“ für jedes Individuum der Welt, in dem Informationen über den Geburts- und Todestag sowie Interaktionen mit dem Gesundheits- und Sozialsystem dokumentiert

werden. Bereits Dunn (1946) realisierte, dass die Datenqualität für ein derartiges „book of life“ eine große Herausforderung darstellt. In der Publikation von Newcombe et al. (1959) wurde im Jahr 1959 erstmals der Einsatz von Computern zur Automatisierung des RL vorgeschlagen. Die Grundidee des probabilistischen RL-Ansatzes geht ebenfalls auf Newcombe et al. (1959) zurück. In seinem probabilistischen Ansatz nutzt er das Soundex Ähnlichkeitsmaß, um Namensvariationen zu bewältigen. Über die Verteilung der Attributwerte werden dann Gewichte für die einzelnen Attribute bestimmt, um letztendlich zu entscheiden, ob ein Datensatzpaar zur selben Realwelt-Entität gehört oder nicht (vgl. Newcombe et al., 1959). Auf Basis der Publikation von Newcombe et al. (1959) haben Fellegi und Sunter (1969) in ihrer Publikation dargelegt, dass der optimale probabilistische RL-Prozess nur unter der Annahme gefunden werden kann, dass die Attribute, die für den Vergleich der Datensätze genutzt werden, unabhängig voneinander sind. Der Ansatz von Fellegi und Sunter (1969) ist bis heute noch die Basis für viele RL-Systeme und Softwareprodukte, wie bspw. das System SPLINK[2] (vgl. Christen, 2012a, S. 10).

RL wurde über die letzten Jahre in verschiedenen Forschungsbereichen unabhängig voneinander erforscht, was zu einer Vielzahl von Synonymen für den Begriff RL geführt hat. Für RL werden u.a. die folgenden Synonyme verwendet: Deduplication, Duplicate Detection, Entity Identification, Entity Matching, Entity Resolution, Object Matching oder Reference Matching (vgl. Christen, 2012a, S. 11; Köpcke, 2014, S. 21-22).

Durch die digitale Transformation und die daraus resultierend stark steigenden Datenmengen wird bis heute viel RL-Forschung betrieben. In den letzten Jahren wurden neue Verfahren wie Machine Learning (ML), Natural Language Processing (NLP) und neuronale Netze für den Einsatz im RL erforscht, um die Qualität des RL-Prozess zu optimieren (Christen, 2012b; Doan et al., 2020; Barlaug & Atle Gulla, 2020; Barlaug, 2020; Papadakis, Ioannou, Thanos & Palpanas, 2021). Obwohl viel im Bereich RL geforscht worden ist und geforscht wird, werden die erforschten Ansätze kaum in der Praxis eingesetzt. Stattdessen basiert die Datenintegration in Unternehmen auf ad-hoc Projekten, die von Fall zu Fall durchgeführt werden (vgl. Barlaug & Atle Gulla, 2020).

2.2.1 Record Linkage-Prozess

Der RL-Prozess besteht aus den fünf Prozessschritten (1) Data Preparation, (2) Blocking, (3) Comparison, (4) Classification und (5) Evaluation (siehe Abb. 2.4) (vgl. Christen, 2012a).

Im Prozessschritt Data Preparation wird dafür gesorgt, dass die Daten aus beiden Datenquellen in einem einheitlichen Format vorliegen. Der Prozessschritt Blocking hat das Ziel die quadratische Komplexität des RL-Prozesses zu reduzieren. Dies wird versucht, indem die

[2] `https://github.com/moj-analytical-services/splink`

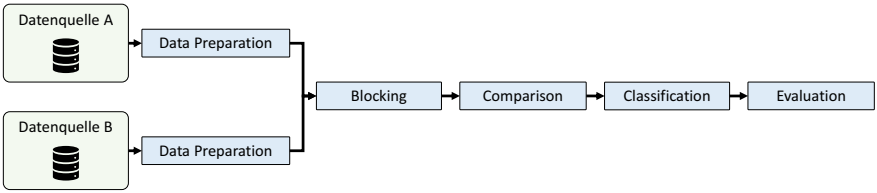


Abbildung 2.4: Record Linkage-Prozess - eigene Darstellung in Anlehnung an Christen (2012a)

zur Verfügung stehenden Attribute der Datenquellen genutzt werden, um Kandidatenpaare zu generieren, die wahrscheinlich zur selben Realwelt-Entität gehören. Anschließend werden die auf der Basis des Blocking generierten Datensatzpaare im Prozessschritt Comparison mit Ähnlichkeitsmaßen verglichen. Im Prozessschritt Classification erfolgt dann auf Basis der ermittelten Ähnlichkeiten die Klassifikation der Datensatzpaare in die Kategorien Match, Kein-Match und potenzieller-Match. Im letzten Prozessschritt der Evaluation wird die Qualität und Vollständigkeit der Ergebnisse des RL-Prozesses bewertet (vgl. Christen, 2012a, S. 23).

Die einzelnen RL-Prozessschritte werden in den folgenden Abschnitten näher beschrieben. Damit die Herausforderungen und Aufgaben der einzelnen Prozessschritte anschaulich beschrieben werden können, soll das in Abbildung 2.5 dargestellte Beispiel verwendet werden. Abbildung 2.5 zeigt ein CRM-System und eine Vertragsdatenbank. Beide Datenquellen enthalten Unternehmensdatensätze, deren Unternehmensname und Adressdaten unterschiedlich strukturiert und repräsentiert sind. Die Datenquellen besitzen keine gemeinsame ID, sodass der RL-Prozess auf die vorhandenen Attribute zum Abgleich zurückgreifen muss.

2.2.1.1 Data Preparation

Im Prozessschritt Data Preparation werden zunächst die technische, semantische und syntaktische Heterogenität (siehe Abschnitt 2.1.1), die zwischen den zu integrierenden Datenquellen bestehen, gelöst. Das Ziel der Data Preparation ist es, die beiden zu integrierenden Datenquellen zu bereinigen und zu standardisieren. Dies gilt vor allem für die Attribute, die für den Vergleich der Datensatzpaare herangezogen werden. Daher sehen Christen (2012a) und I. Koumarelas, Jiang und Naumann (2020) den Prozessschritt Data Preparation als einen der wichtigsten im gesamten RL-Prozess. Da die zu integrierenden Datenquellen und deren Datenqualität sehr verschieden sein können, muss die Data Preparation oftmals individuell für die jeweiligen Datenquellen entwickelt werden. Christen (2012a) nennt drei Data Preparation Verfahren, die häufig verwendet werden:

Datenquelle A – CRM System

ID-A	Name	Adresse	Land	Mitarbeiter
1	Volksbank Oldenburg EG	Lange Straße 8/9 – Staulinie 3 26122 Oldenburg	DE	160
2	Volkswagen A.G.	Berliner Ring 2, 38440 Wolfsburg	DEU	662.600
3	OOWV	Georgstraße 4 26919 Brake	DE	Null
4	CEWE Color	Meerweg 30-32, 26133 Oldenburg	DE	4.199
5	EWE Tel GmbH	Cloppenburger Str. 310 26133 Oldenburg	Null	1.230

Datenquelle B - Vertragsdatenbank

ID-B	Name	Straße	Hausnummer	Stadt	Land	Anz. Veträge
A	Volksbank e.G.	Lange Str.	8	Oldenburg	Deutschland	5
B	Volkswagen Aktiengesellschaft	Berliner Ring	2	38440	Germany	10
C	Oldenburgisch-Ostfriesischer Wasserverband	Georgstrasse		Brake	Deutschland	3
D	Cewe Stiftung & Co. KGaA	Meerweg	30-32	Oldenburg	Deutschland	12

Abbildung 2.5: Beispiel Datensätze für den Record Linkage-Prozess

Entfernen von unerwünschten Zeichen und Worten: In Datenquellen existieren oftmals unerwünschte Zeichen, wie Kommas, Doppelpunkte, Semikolons, Punkte oder Anführungszeichen, die entfernt werden sollten. In den Datenquellen können auch ganze Wörter entfernt werden, wenn diese keine für den RL-Prozess relevanten Informationen beinhalten (vgl. Christen, 2012a). In den Beispieldatensätzen in Abbildung 2.5 sollten bspw. die Kommata im Attribut Adresse in Datenquelle A oder die Punkte im Attribut Name in Datenquelle A entfernt werden.

Abkürzungen auflösen und Rechtschreibfehler korrigieren: Die Daten der zu integrierenden Datenquellen sind häufig durch Abkürzungen, Namensvarianten, Spitzname oder häufige Schreibfehler unterschiedlich repräsentiert. Diese unterschiedlichen Repräsentationen sollen in diesem Schritt standardisiert werden, um die Datenqualität für die Integration zu erhöhen (vgl. Christen, 2012a). In den Beispieldatensätzen in Abbildung 2.5 sollte bspw. die Abkürzung OOWV (ID-A = 3) des Unternehmensnamens Oldenburgisch-Ostfriesischer Wasserverband (ID-B = C) aufgelöst werden.

Normalisieren von Attributen in atomare Attribute: Dieser Schritt befasst sich mit dem Problem, dass Attribute oftmals nicht normalisiert sind und mehrere Informationen enthalten. Ein Beispiel ist, wenn die Adresse in einem Attribut dargestellt wird, wie in Abbildung 2.5 in Datenquelle A dargestellt. Das Attribut Adresse enthält Informationen zur Straße, Hausnummer, Postleitzahl und Stadt. Für den RL-Prozess sollten die Attribute normalisiert werden, sodass die Wertebereiche atomar sind. Die Adresse der Datenquelle A

sollte also in die Attribute Straße, Hausnummer, Postleitzahl und Stadt unterteilt werden (vgl. Christen, 2012a).

Die durch die Data Preparation Verfahren aufbereiteten Daten sollten in den Datenquellen nicht überschrieben werden, damit die Ursprungsdaten erhalten bleiben. Die Ursprungsdaten werden benötigt, wenn bswp. Fehler in den Data Preparation Verfahren gemacht worden sind. Daher sollten die aufbereiteten Daten in einer weiteren Datenbanktabelle oder Datei persistiert werden, sodass die aufbereiten Daten für den nächsten RL-Prozessschritt Blocking leicht zugänglich sind (vgl. Christen, 2012a).

2.2.1.2 Blocking

Nachdem die Daten im RL-Prozessschritt Data Preparation aufbereitet worden sind, folgt das Blocking. Auch das Blocking ist ein wichtiger Prozessschritt, der für die Zeiteffizienz und Skalierbarkeit sorgt, da RL ohne Blocking das Problem der quadratischen Zeitkomplexität, $O(n^2)$, besitzt, da alle Datensätze miteinander verglichen werden müssen. Für das Beispiel in Abbildung 2.5 bedeutet dies 20 Vergleiche[3]. Bei derart kleinen Datenmengen stellt die quadratische Komplexität keine Probleme dar. Bei Datenquellen die 100.000 und 50.000 Datensätze enthalten, sind 5.000.000.000 Vergleiche notwendig. Wenn 100.000 Datensätze pro Sekunde verglichen werden können, würde der Vergleich insgesamt ca. 14 Stunden dauern (vgl. Papadakis, Skoutas, Thanos & Palpanas, 2020).

Um die quadratische Zeitkomplexität zu lösen, werden im Prozessschritt Blocking ähnliche Datensätze aus den zu integrierenden Datenquellen in einen Block gruppiert. Die Ähnlichkeitsberechnungen werden ausschließlich zwischen den Datensätzen, die sich in einem gemeinsamen Block befinden, durchgeführt, sodass die Effizienz erhöht wird (vgl. Christen, 2012a).

Für das Blocking wurden einige Verfahren und Algorithmen entwickelt. Eine umfassende Übersicht ist in der Publikation von Papadakis, Skoutas et al. (2020) erstellt worden. Beispielsweise gruppieren Papadakis, Skoutas et al. (2020) die Blocking Verfahren in die Kategorie „learning-based“ und „nonlearning“. Die „learning-based“ Verfahren berechnen mit Hilfe von ML den besten Blocking Key und benötigten daher Trainingsdaten. Die „nonlearning“ Verfahren basieren auf einem manuell selektierten Blocking Key der durch Expertenwissen und Datenstatistiken erstellt worden ist (vgl. Papadakis, Skoutas et al., 2020). Die Auswahl der Attribute für den Blocking Key sollte sorgfältig getroffen werden. Bei der Auswahl sollte die Qualität, die Vollständigkeit sowie die Verteilung der Werte in den Attributen berücksichtigt werden (vgl. Christen, 2012a). Die für diese Arbeit relevanten Blocking Verfahren Standard Blocking und Sorted Neighborhood werden im Folgenden näher beschrieben:

[3] Datenquelle A mit 5 Datensätzen x Datenquelle B mit 4 Datensätzen

Standard Blocking (SB): Das Standard Blocking stellt das einfachste Verfahren dar, da ein Experte die am besten geeigneten Attribute auswählt, die für die Bildung des Blocking Key herangezogen werden. Es können auch Bestandteile, wie die ersten drei Buchstaben eines Attributs, genutzt werden und mit weiteren Attributen oder Attributsbestandteilen zu einem Blocking Key kombiniert werden. Jeder eindeutige Blocking Key bildet einen Block, in dem die Datensatzpaare miteinander verglichen werden. In Abbildung 2.6 ist dargestellt, welche Datensatzpaare beim Standard Blocking über das Attribut Stadt für die Datenquellen aus dem Beispiel in Abbildung 2.5 generiert werden würden. Es existieren die vier eindeutige Blocking Keys „Oldenburg“, „Wolfsburg“, „Brake“ und „38440“. Das Standard Blocking ist gegenüber Datenqualitätsproblemen sehr empfindlich. Das Beispiel der Datensätze mit dem Blocking Key „Wolfsburg“ und „38440“ zeigt dieses Problem. Da ein Datensatz die Stadt Angabe textuell und der andere Datensatz die Stadt Angabe als Postleitzahl enthält, werden die Datensätze fälschlicherweise nicht zu einem Datensatzpaar gruppiert.

Blocking Key „Oldenburg“

ID-A	Name-A	ID-B	Name-B
1	Volksbank Oldenburg EG	A	Volksbank e.G.
1	Volksbank Oldenburg EG	D	Cewe Stiftung & Co. KGaA
4	CEWE Color	A	Volksbank e.G.
4	CEWE Color	D	Cewe Stiftung & Co. KGaA
5	EWE Tel GmbH	A	Volksbank e.G.
5	EWE Tel GmbH	D	Cewe Stiftung & Co. KGaA

Blocking Key „Wolfsburg“

ID-A	Name-A	ID-B	Name-B
2	Volkswagen A.G.	NULL	NULL

Blocking Key „Brake“

ID-A	Name-A	ID-B	Name-B
3	OOWV	C	Oldenburgisch-Ostfriesischer Wasserverband

Blocking Key „38440“

ID-A	Name-A	ID-B	Name-B
NULL	NULL	B	Volkswagen Aktiengesellschaft

Abbildung 2.6: Beispiel Standard Blocking

Sorted Neighborhood (SN): Für das Sorted Neighborhood Verfahren muss ebenfalls ein Blocking Key ausgewählt und erstellt werden. Anschließend werden die Datensätze über den Blocking Key alphabetisch sortiert. Daraufhin wird ein Fenster der Größe w über die sortierte Liste gelegt und alle Datensatzpaare innerhalb des Fensters werden für den Vergleich selektiert. Die Grundannahme des Verfahren ist, dass die am wahrscheinlichsten übereinstimmenden Datensätze nah beieinander liegen. Das Ergebnis des Sorted Neighborhood Verfahrens über den Blocking Key Name mit einer Fenster Größe von w = 3 für das Beispiel aus Abbildung 2.5 ist in Abbildung 2.7 dargestellt. Insgesamt wurden sieben Datensatzpaare generiert, in denen alle korrekten Matches enthalten sind. Ein Nachteil des

Sorted Neighborhood Verfahren ist die zu konfigurierende Fenstergröße w. Diese ist manuell zu konfigurieren und die Auswahl der optimalen Fenstergröße ist ebenfalls schwierig zu bestimmen (vgl. Papadakis, Skoutas et al., 2020).

Blocking Key
Name

ID-A	Name-A	ID-B	Name-B
4	CEWE Color	D	Cewe Stiftung & Co. KGaA
5	EWE Tel GmbH	D	Cewe Stiftung & Co. KGaA
5	EWE Tel GmbH	C	Oldenburgisch-Ostfriesischer Wasserverband
3	OOWV	C	Oldenburgisch-Ostfriesischer Wasserverband
3	OOWV	B	Volkswagen Aktiengesellschaft
1	Volksbank Oldenburg EG	A	Volksbank e.G.
2	Volkswagen A.G.	B	Volkswagen Aktiengesellschaft

Abbildung 2.7: Beispiel Sorted Neighborhood - w = 3

Für die Evaluation der Blocking Verfahren wird zum Vergleich das Kreuzprodukt der beiden zu integrierenden Datenquellen gebildet ($|A|x|B|$). Weiterhin ist die Anzahl der gesamten Matches zwischen den Datenquellen $|M|$, sowie die Anzahl der korrekten Matches in den selektierten Datensatzpaaren durch das Blocking $|N_m|$ relevant. Mit diesen Angaben können die klassischen Metriken zur Evaluation von Blocking Verfahren berechnet werden, die im Folgenden beschrieben werden (vgl. Papadakis, Skoutas et al., 2020):

Pair Completeness (PC): Die Pair Completeness zeigt, wie viele der korrekten Matches durch das Blocking Verfahren selektiert worden sind, siehe Formel 2.1.

$$PC = \frac{|N_m|}{|M|} \tag{2.1}$$

Pair Quality (PQ): Die Pair Quality gibt an, wie hoch der Anteil der Datensatzpaare, die ein korrektes Match sind, gegenüber der insgesamt generierten Datensatzpaare $|N|$ ist, siehe Formel 2.2.

$$PQ = \frac{|N_m|}{|N|} \tag{2.2}$$

Reduction Ratio (RR): Die Reduction Ratio gibt an, um wie viel Prozent die Menge der zu vergleichenden Datensatzpaare durch das Blocking Verfahren gegenüber dem Kreuzprodukt reduziert worden ist, siehe Formel 2.3.

$$RR = 1 - \frac{N}{|A|x|B|} \tag{2.3}$$

Nachdem ein Blocking Verfahren ausgewählt und angewendet wurde, werden die generierten Datensatzpaare an den darauf folgenden RL-Prozessschritt Comparsion übergeben.

2.2.1.3 Comparison

Die im RL-Prozessschritt Blocking generierten Datensatzpaare werden im Prozessschritt Comparison genutzt, um die Ähnlichkeit zwischen den Datensätzen zu berechnen. Um die Ähnlichkeit zwischen den Datensätzen zu bestimmen, werden die Attribute der beiden Datensätze herangezogen. Im Beispiel aus Abbildung 2.5 kann der Name sowie die Adressdaten für den Vergleich der Datensätze genutzt werden. Generell gilt, je mehr gleiche Attributwerte existieren, desto ähnlicher sind sich die Datensätze (vgl. Christen, 2012a, S. 29-30).

Im Beispiel in Abbildung 2.5 bilden Datensatz mit der ID-A '1' und Datensatz mit der ID-B 'A' ein Match. Auch wenn beide Namen Attribute durch die Data Preparation klein geschrieben und die Sonderzeichen entfernt werden, unterscheiden sich die Werte der Attribute, obwohl die Datensätze zusammengehören. Daher können die Attribute nicht ausschließlich auf Gleichheit, 100% Ähnlichkeit, und keine Gleichheit, 0% Ähnlichkeit, überprüft werden. Um dieses Problem zu lösen existieren einige Ähnlichkeitsmaße, die eine Ähnlichkeitswahrscheinlichkeit zwischen 0 und 100% zwischen den Attributwerten berechnen (vgl. Christen, 2012a, S. 101).

Es existieren Ähnlichkeitsverfahren für Attribute die Zeichenketten, numerische Werte, Zeitangaben oder geografische Koordinaten enthalten (vgl. Christen, 2012a). Da für diese Arbeit die Ähnlichkeitsverfahren für Zeichenketten und geografische Koordinaten relevant sind, werden diese näher beschrieben. Köpcke (2014) unterteilt die Ähnlichkeitsverfahren für Zeichenketten in die Kategorien (1) zeichenbasierte-Verfahren, (2) tokenbasierte-Verfahren und (3) hybride-Verfahren. Die zeichenbasierten-Verfahren behandeln die zu vergleichenden Werte der Attribute als einen gesamten String. Tokenbasierte-Verfahren sehen die Werte der zu vergleichenden Attribute nicht als gesamten String, sondern als eine Liste einzelner Token, die bspw. über das Leerzeichen getrennt werden. Die hybriden-Verfahren kombinieren die zeichenbasierten- und tokenbasierten-Verfahren (vgl. Köpcke, 2014). In dieser Arbeit werden die zeichenbasierten-Verfahren Levenshtein und Jaro-Winkler, das tokenbasierte-Verfahren Jaccard, das hybride-Verfahren Monge-Elkan sowie die Haversine-Distanz für geografische Koordinaten genutzt und im Folgenden näher beschrieben.

Levenshtein: Die Levenshtein-Distanz berechnet zwischen den zu vergleichenden Strings σ_1 und σ_2 die minimale Anzahl von Einfüge-, Änderungs- oder Löschoperationen von Zeichen, um String σ_1 in σ_2 umzuwandeln. Bei jeder Ausführung einer der drei Operationen wird

ein Zähler um eins inkrementiert. In Formel 2.4 ist die Levenshtein-Distanz dargestellt.

$$\text{Levenshtein}\left(\sigma_1, \sigma_2\right) = 1 - \left(\frac{\text{dist}\left(\sigma_1, \sigma_2\right)}{\max\left(\left|\sigma_1\right|, \left|\sigma_2\right|\right)}\right) \tag{2.4}$$

Als Beispiel soll die Levenshtein-Distanz zwischen den beiden Strings σ_1 = „Volswagon“ und σ_2 = „Volkswagen“ berechnet werden. Die $\text{dist}\left(\sigma_1, \sigma_2\right)$ zwischen „Volswagon“ und „Volkswagen“ beträgt zwei, da ein 'k' eingefügt und ein 'o' zu einem 'e' verändert werden muss. Der String mit der maximalen Länge $\max\left(\left|\sigma_1\right|, \left|\sigma_2\right|\right)$ ist σ_2 mit zehn, sodass sich für das Beispiel eine Levenshtein Ähnlichkeit von 80% ergibt, siehe Formel 2.5.

$$\text{Levenshtein}\left(Volswagon, Volkswagen\right) = 1 - \left(\frac{2}{10}\right) = 0.80 \tag{2.5}$$

Jaro-Winkler: Das Jaro-Winkler-Distanzmaß besteht zunächst aus der Jaro Berechnung, die in Formel 2.6 dargestellt ist.

$$\text{Jaro}\left(\sigma_1, \sigma_2\right) = \frac{1}{3}\left(\frac{c}{\left|\sigma_1\right|} + \frac{c}{\left|\sigma_2\right|} + \frac{c - t/2}{c}\right) \tag{2.6}$$

In der Formel 2.6 stellt c die Anzahl der gemeinsamen Zeichen und t die Anzahl der Transpositionen der gemeinsamen Zeichen dar. Die gemeinsamen Zeichen werden über die Zeichen der zu vergleichenden Strings σ_1 und σ_2 berechnet, für die gilt, $\sigma_1[i] = \sigma_2[j]$ and $|i - j| \leq \frac{1}{2}\min\left\{\left|\sigma_1\right|, \left|\sigma_2\right|\right\}$. Wenn die Zeichen an Position i in den beiden Strings σ_1 und σ_2 nicht übereinstimmen, wird eine Transposition erfasst.

Die Jaro-Winkler-Distanz, dargestellt in Formel 2.7 nutzt noch einen Prefix Faktor p, über die die Anzahl der übereinstimmenden Zeichen l am Anfang des Strings höher gewichtet werden kann.

$$\text{JaroWinkler}\left(\sigma_1, \sigma_2\right) = \text{Jaro}\left(\sigma_1, \sigma_2\right) + \left(\ell \cdot p\left(1 - \text{Jaro}\left(\sigma_1, \sigma_2\right)\right)\right) \tag{2.7}$$

Als Beispiel wird die Jaro-Winkler Ähnlichkeit zwischen den Strings σ_1 = „Hewlett Packard“ und σ_2 = „Hewlet Peckard“ berechnet werden, siehe Formel 2.8.

$$\text{JaroWinkler}\left(HewlettPackard, HewletPeckard\right) = 0.932 + 6 \cdot 0.1 \cdot (1 - 0.932) = 0.9728 \tag{2.8}$$

Die beiden Strings besitzen 13 gemeinsame Zeichen und das gemeinsame Prefix „Hewlet“ hat die Länge l von sechs Zeichen. Da keine Transpositionen vorhanden sind, beträgt die Jaro Ähnlichkeit 0.932. Mit einem Prefix Faktor p von 0,1 ergibt sich eine Jaro-Winkler Ähnlichkeit von 0,9728.

Jaccard: Die Jaccard-Distanz berechnet die Ähnlichkeit zwischen den Token, die sich in den zu vergleichenden Strings befinden. Die Jaccard-Distanz ist definiert als die Anzahl der überschneidenden Token der beiden Strings dividiert durch die Anzahl der Vereinigung der Token der beiden Strings, siehe Formel 2.9.

$$\text{Jaccard}\left(\sigma_1, \sigma_2\right) = \frac{|T_{\sigma_1} \cap T_{\sigma_2}|}{|T_{\sigma_1} \cup T_{\sigma_2}|} \tag{2.9}$$

Als Beispiel soll die Jaccard Ähnlichkeit zwischen den Strings σ_1 = „Volkswagen AG“ und σ_2 = „Volkswagen Aktiengesellschaft“berechnet werden, siehe Formel 2.10.

$$\text{Jaccard}\left(Volkswagen\ AG, Volkswagen\ Aktiengesellschaft\right) = \frac{1}{3} = 0.333 \tag{2.10}$$

Die beiden Strings werden über das Leerzeichen tokenisiert, sodass die TokenmengenT_{σ_1} = {Volkswagen, AG } and T_{σ_2} = {Volkswagen, Aktiengesellschaft} erzeugt werden. Für die Tokenmengen gibt es eine Überschneidung $T_{\sigma_1} \cap T_{\sigma_2}$ = {Volkswagen}. Insgesamt besteht die Vereinigung der beiden Tokenmengen aus drei eindeutigen Token $T_{\sigma_1} \cup T_{\sigma_2}$ = {Volkswagen, AG, Aktiengesellschaft}, sodass sich eine Jaccard Ähnlichkeit von 0,333 ergibt.

Monge-Elkan: Die Monge-Elkan-Distanz wurde nach ihren Autoren Monge und Elkan (1996) benannt und speziell für den Vergleich von Strings mit mehreren Worten entwickelt (vgl. Christen, 2012a, S. 111). Die Berechnung der Monge-Elkan-Distanz ist in Formel 2.11 dargestellt.

$$\text{MongeElkan}\left(\sigma_1, \sigma_2\right) = \frac{1}{|T_{\sigma_1}|} \sum_{i=1}^{|T_{\sigma_1}|} \max_{j=1,\ldots,|T_{\sigma_2}|} s'\left(t_{i\sigma_1}, t_{j_{\sigma_2}}\right) \tag{2.11}$$

Die Monge-Elkan-Distanz besitzt eine innere Ähnlichkeitsfunktion $\text{sim}'\left(t_1, t_2\right)$, die bspw. die Levenshtein-Distanz sein kann, um die Ähnlichkeit der Token der zu vergleichenden Strings σ_1 und σ_2 zu berechnen.

Als Beispiel soll die Monge-Elkan Ähnlichkeit zwischen den Strings σ_1 = „Hewlett Packard“ und σ_2 = „Hewlet Peckard“ berechnet werden, siehe folgende Berechnung:

$$\begin{aligned} &\sigma_1 = \text{Hewlett Packard}; \quad t_{1\sigma_1} = \text{Hewlett}; \quad t_{2\sigma_1} = \text{Packard} \\ &\sigma_2 = \text{Hewlet Peckard}; \quad t_{1\sigma_2} = \text{Hewlet}; \quad t_{2\sigma_2} = \text{Peckard} \end{aligned}$$

$$\begin{aligned} &s'\left(t_{1\sigma_1}, t_{1\sigma_2}\right) \approx 0.8571; \quad s'\left(t_{1\sigma_1}, t_{2\sigma_2}\right) = 0 \\ &s'\left(t_{2\sigma_1}, t_{1\sigma_2}\right) = 0; \qquad\quad s'\left(t_{2\sigma_1}, t_{2\sigma_2}\right) \approx 0.8571 \end{aligned}$$

$$\begin{aligned}\text{MongeElkan}(\sigma_1, \sigma_2) =& \frac{1}{2}\left(\max\left(s'\left(t_{1\sigma_1}, t_{1\sigma_2}\right), s'\left(t_{1\sigma_1}, t_{2\sigma_2}\right)\right) + \right. \\ & \left. \max\left(s'\left(t_{2\sigma_1}, t_{1\sigma_2}\right), s'\left(t_{2\sigma_1}, t_{2\sigma_2}\right)\right)\right) \\ =& \frac{1}{2}(0.8571 + 0.8571) = 0.8571\end{aligned}$$

Die Strings σ_1 und σ_2 wurden über das Leerzeichen tokenisiert. Anschließend wurde für alle möglichen Token Kombinationen die Levenshtein Ähnlichkeit berechnet. Daraus ergibt sich, dass die Token Kombinationen $t_{1\sigma_1}, t_{1\sigma_2}$ und $t_{2\sigma_1}, t_{2\sigma_2}$ mit jeweils 0,8571 die höchste Ähnlichkeit aufweisen. Mit diesen Werten wird dann die gesamte Ähnlichkeit der beiden Strings berechnet und insgesamt ist die Monge-Elkan Ähnlichkeit des Beispiels 0,8571.

Haversine: Die Haversine-Distanz kann genutzt werden, um die Distanz zwischen geografischen Angaben über die Longitude und Latitude in Kilometern zu berechnen. Ein Python Framework, welches die Haversine-Distanz implementiert hat, ist bspw. das Haversine Framework[4]. Adressdaten können ebenso über die klassischen String Ähnlichkeitsmaße verglichen werden. Die Haversine-Distanz stellt eine Alternative dar. Der Vergleich der Adressdaten über die Longitude und Latitude erfordert besondere Berücksichtigung der vorliegenden Datenqualität, da fehlende Adressangaben eine große Auswirkung auf die Longitude und Latitude haben und durch fehlende Angaben fälschlicherweise große Kilometer-Distanzen entstehen können (vgl. Christen, 2012a, S. 124).

Nach dem RL-Prozessschritt Comparison sollte für jedes zu vergleichende Attribut eine Ähnlichkeit über ein Ähnlichkeitsverfahren berechnet worden sein, sodass für jedes Datensatzpaar ein Vektor aus Ähnlichkeitswerten gebildet worden ist. Dieser Vektor wird an den nächsten Prozessschritt Classification übergeben, um zu bestimmen, welches Datensatzpaar ein Match und Kein-Match ist (vgl. Christen, 2012a, S. 30).

2.2.1.4 Classification

Der RL-Prozessschritt Classification ist für die Klassifikation der Datensatzpaare anhand der vorhandenen Vektoren in zwei Klassen oder drei Klassen zuständig. Beim zwei Klassen Klassifikationsproblem werden die Datensatzpaare in die Klassen Match und Kein-Match unterteilt. Ein Match repräsentiert dieselbe Realwelt-Entität, während ein Datensatzpaar der Kein-Match Klasse Datensätze beinhält, die nicht zur selben Realwelt-Entität gehören. Die Datensatzpaare, die während des RL-Prozessschrittes Blocking entfernt worden sind, werden ebenfalls als Kein-Match klassifiziert (vgl. Christen, 2012a, S. 32).

Neben dem zwei Klassen Klassifikationsproblem existiert das drei Klassen Klassifikationspro-

[4] `https://pypi.org/project/haversine/`

blem. Das drei Klassen Klassifikationsproblem besteht aus den Klassen Match, potenzielles-Match und kein-Match. Die Datensätze, die der Kategorie potenzielles-Match zugeordnet werden, müssen manuell überprüft werden, ob sie ein Match oder kein-Match sind (vgl. Christen, 2012a, S. 32).

Viele Forschungsarbeiten, die im Bereich RL angesiedelt sind, fokussieren die Optimierung der Algorithmen des Prozessschrittes Classification. Über die Jahre sind eine Vielzahl verschiedener Verfahren entstanden, die im Folgenden beschrieben werden (vgl. Christen, 2012a, S. 129):

Deterministisch: Das deterministische RL-Verfahren ist der einfachste Ansatz die Datensatzpaare in die Kategorien Match und Kein-Match zu unterteilen. Bei diesem Verfahren werden ein oder mehrere Attribute definiert, die exakt übereinstimmen müssen, damit ein Datensatzpaar als Match klassifiziert wird. Die übrigen Datensatzpaare werden als Kein-Match klassifiziert. Im einfachsten Fall existiert eine gemeinsame Identifikationsnummer (ID), um die Datensätze in Match und Kein-Match zu klassifizieren. Liegt keine gemeinsame ID vor, können auch Attribute, wie bspw. Vorname und Nachname, genutzt werden. Der Nachteil des deterministischen RL-Verfahren besteht darin, dass die einfachsten Unterschiede zwischen den Attributen, wie bspw. vertauschte Buchstaben „Kruse„und „Kurse“, dazu führen, dass Datensatzpaare als Kein-Match klassifziert werden (vgl. Sayers, Ben-Shlomo, Blom & Steele, 2016).

Schwellwert-basiert: Ein weiterer Ansatz für den Prozessschritt Classification ist die Schwellwert-basierte Klassifikation. Dieses Verfahren addiert die zuvor berechneten Ähnlichkeitswerte und anschließend wird ein Schwellwert festgelegt, der bestimmt welche Datensatzpaare ein Match und welche Kein-Match bilden (vgl. Christen, 2012a, S. 131). Das Festlegen des Schwellwertes kann sowohl manuell als auch über ein supervised Machine Learning-Verfahren erfolgen. Die Schwellwertbasierte Klassifikation besitzt zwei Nachteile. Der erste Nachteil besteht darin, dass die Ähnlichkeitswerte alle zwischen 0 und 1 normalisiert sind und alle Attribute, die zum Vergleich herangezogen werden, mit dem gleichen Gewicht zur Ähnlichkeitsberechnung beitragen. Dadurch ist es nicht möglich, einzelnen Attributen ein höheres Gewicht zu geben, die eine höhere Bedeutung für die Unterscheidung der Datensatzpaare in Match und Kein-Match besitzen. Ein weiterer Nachteil besteht darin, dass die Detailinformationen der einzelnen Ähnlichkeitswerte je Attribut durch die Summierung der Ähnlichkeitswerte verloren gehen. So kann ein Datensatzpaar welches ein Match darstellt denselben summierten Ähnlichkeitswert besitzen, wie ein Datensatzpaar welches ein Kein-Match darstellt (vgl. Christen, 2012a, S. 131-132).

Probabilistisch: Der probabilistische RL-Ansatz geht auf Fellegi und Sunter (1969) zurück. Die Idee hinter dem probabilistischen Verfahren besteht darin, dass die Attribute der

Datensätze, wie bspw. Name, Vorname und Adressangaben, genutzt werden für den RL-Prozess, wenn keine gemeinsame ID vorliegt. Da die Werte der Attribute aufgrund von Datenqualitätsproblemen nicht immer exakt übereinstimmen, sollte den unterschiedlichen Attributen für die Ähnlichkeitsberechnung unterschiedliche Gewichte verliehen werden. Dabei sollten die Gewichte nicht nur in Abhängigkeit vom Attribut sondern auch von den zu vergleichenden Werten abhängen, wie folgendes Beispiel zeigt: Unter der Annahme, dass mehr Personen den Nachnamen „Meier“ als „Stahl-Holtmann“ besitzen, sollte einem Datensatzpaar mit übereinstimmenden Nachnamen „Meier“ ein geringeres Gewicht als dem Datensatzpaar mit dem gleichen Nachnamen „Stahl-Holtmann“ verliehen werden (vgl. Christen, 2012a, S. 133-134).

Regelbasiert: Der regelbasierte-Ansatz unterscheidet sich vom zuvor vorgestellten probabilistischen Ansatz. Der regelbasierte-Ansatz umfasst Regeln, die die Datensatzpaare in die Kategorien Match und Kein-Match unterteilen. Das Regelwerk wird auf die berechneten Ähnlichkeitswerte im Prozessschritt Comparison angewendet. Für eine Regel werden die einzelnen Ähnlichkeitswerte der Datensatzpaare auf die in den Regeln festgelegten Bedingungen geprüft. Die einzelnen Bedingungen für die Ähnlichkeitswerte werden über Konjunktionen, Disjunktionen und Negationen kombiniert. Ein Beispiel für eine Regel ist in Formel 2.12 dargestellt (vgl. Christen, 2012a, S. 139-140).

$$\begin{aligned}&(s(\text{ Nachname })[r_i, r_j] \geq 0.8) \wedge (s(\text{ Vorname })[r_i, r_j] \geq 0.9) \\ \wedge\, &(s(\text{ Wohnort })[r_i, r_j] = 1.0) \wedge (s(\text{ Geburtstag })[r_i, r_j] = 1.0) \Rightarrow [r_i, r_j] \rightarrow \text{ Match}\end{aligned} \tag{2.12}$$

Die Regel in Formel 2.12 prüft, ob der Nachname mindestens zu 80%, der Vorname mindestens zu 90% sowie der Wohnort und das Geburtsdatum zu 100% übereinstimmen. Christen (2012a) empfiehlt das Regelwerk möglichst klein zu halten, da Regelwerke mit wenigen Regeln einfacher zu warten und zu pflegen sind. Besonders vor dem Hintergrund, dass die Regelwerke für jede Datenintegration neu entwickelt oder angepasst werden müssen (vgl. Christen, 2012a, S. 140). Die Regelwerke können entweder manuell oder mit Hilfe von Trainingsdaten erstellt werden (vgl. Christen, 2012a, S. 141).

Die manuelle Entwicklung eines Regelwerkes basiert auf dem Fachwissen des Entwicklers über die jeweils zu integrierenden Datenquellen. Dabei ist die zuvor getroffene Auswahl der Algorithmen und Verfahren für die RL-Prozessschritte Blocking und Comparison zu berücksichtigen, da beide Prozessschritte auf die generierten Datensatzpaare und die berechneten Vektoren mit den entsprechenden Ähnlichkeitswerten Auswirkung haben. Das manuelle Erstellen eines Regelwerkes ist ein aufwändiger Prozess, da viele Bedingungen innerhalb einer Regel und die Kombination der Regeln miteinander iterativ entwickelt

werden müssen, um zum bestmöglichen Ergebnis zu gelangen. Die Regeln werden mit Hilfe von Trainingsdaten, die Datensatzpaare enthalten von denen der Status Match oder Kein-Match bekannt ist, getestet und manuell geprüft. Sollten keine Trainingsdaten zur Verfügung stehen, sollten die Ergebnisse jeder Regel manuell geprüft werden. Bei der Prüfung wird für jedes selektierte Datensatzpaar der Regel entschieden, ob es sich um einen Match oder Kein-Match handelt. Auch diese Bewertung ist ein aufwändiger Prozess (vgl. Christen, 2012a, S. 141).

Ein alternativer Ansatz ist, die Regeln anhand von Trainingsdaten zu erstellen. Da die Trainingsdaten die Datensatzpaare mit den Informationen Match und Kein-Match enthalten, können die Regeln über den Sequential Covering Algorithmus entwickelt werden (vgl. Han, Kamber & Pei, 2012, S. 359). Die Idee des Algorithmus ist es, eine Regel auf Basis der Trainingsdaten zu erstellen. Anschließend werden die durch die Regel selektierten Datensätze aus der Trainingsdatenmenge entfernt und die nächste Regel wird erstellt. Dieses sequentielle Vorgehen wird wiederholt, bis eine zu definierende Endbedingung erfüllt ist. Im Falle des RL-Prozessschrittes Classification könnten sequentiell Regeln erstellt werden, die Datensatzpaare mit der Klasse Match selektieren (vgl. Christen, 2012a, S. 141-142).

Supervised Machine Learning: Die supervised Verfahren benötigen Trainingsdaten, um ein supervised Machine Learning Modell zu trainieren. Das trainierte Modell wird genutzt, um Datensatzpaare, dessen Zuordnung zur Kategorie Match oder Kein-Match unbekannt ist, einer der Kategorien zuzuordnen. Vielfach verwendete supervised Machine Learning-Verfahren sind der Decision Tree oder die Support Vector Machine. Das Erstellen von supervised Machine Learning Modellen zur Klassifikation der Datensatzpaare ist nicht trivial. Denn das Erstellen eines ausbalancierten Trainingsdatensatzes ist schwierig, da beim RL generell mehr Datensatzpaare existieren, die Kein-Match sind, als Datensätze, die ein Match sind. Das zweite Problem besteht darin, dass das Erstellen eines repräsentativen und somit ausbalancierten Trainingsdatensatzes für die zu integrierenden Datenquellen schwierig ist. Für ein robustes und qualitatives supervised Machine Learning Modell wird ein ausbalancierter Trainingsdatensatz benötigt, der oftmals mit hohem manuellen Aufwand generiert werden muss. Daher sind die erstellten Trainingsdatensätze oftmals ein kleiner Auszug der zu integrierenden Datenquellen, der gleichzeitig alle Ähnlichkeitswert Kombination der Matches und Kein-Matches aus der gesamten Datenmenge repräsentieren muss (vgl. Christen, 2012a, S. 142-147).

Active Learning: Auch Active Learning-Verfahren wurden für den RL-Prozessschritt Classification erforscht. Beim Active Learning werden möglichst wenige Datensatzpaare manuell mit der Kategorie Match und Kein-Match versehen, um Trainingsdaten für das Training eines supervised Machine Learning-Verfahrens zu erstellen. Beim Active Learning werden die Trainingsdaten in einem iterative Prozess solange erweitert, bis das trainierte supervi-

sed Machine-Learning-Verfahren die gewünschte Güte erreicht hat. Durch dieses Vorgehen soll der manuelle Aufwand reduziert werden, da weniger Trainingsdaten erstellt werden als für klassische supervised Machine Learning-Verfahren (vgl. Christen, 2012a, S. 32).

Unsupervised Machine Learning: Die bisher vorgestellten Verfahren für den RL-Prozessschritt Classification haben das Klassifizieren der Datensatzpaare in die Kategorien Match und Kein-Match als klassisches Klassifikationsproblem betrachtet. Die Zuordnung der Datensatzpaare in die Kategorien Match und Kein-Match kann auch über Clustering-Verfahren erfolgen. Beim Clustering werden die Datensatzpaare, die sich nach den ausgewählten Kriterien ähnlich sind, in ein Cluster zusammengefasst. Das Ziel eines Cluster-Verfahrens besteht darin, Cluster zu erzeugen, deren Datensatzpaare in einem Cluster eine sehr hohe Ähnlichkeit aufweisen und die Datensatzpaare aus unterschiedlichen Clustern möglichst unähnlich sind. Da Clustering-Verfahren zu den unsupervised Machine Learning-Verfahren gehören, werden keine Trainingsdaten für das Clustering benötigt (vgl. Christen, 2012a, S. 150-151).

Für den RL-Prozessschritt Classification existiert eine Vielzahl von Algorithmen und Verfahren, aus denen ausgewählt werden kann. Dabei hängt die Wahl des Algorithmus oder Verfahrens von den folgenden Faktoren ab (vgl. Christen, 2012a, S. 161):

- die in der verwendeten Software zur Verfügung stehenden Algorithmen
- die Eigenschaften der zu integrierenden Datenquellen
- das Vorhandensein von Trainingsdaten

2.2.1.5 Evaluation

Nachdem die Datensatzpaare im RL-Prozessschritt Classification den Kategorien Match und Kein-Match zugeordnet worden sind, erfolgt im Prozessschritt Evaluation die Bewertung der Ergebnisqualität. Die Ergebnisqualität setzt sich aus der *Accuracy* und der *Completeness* zusammen. Die *Accuracy* gibt an, wie viele der klassifizierten Matches korrekt sind und somit zur selben Realwelt-Entität gehören. Die *Completeness* gibt an, wie viele der in den Datenquellen tatsächlich existierenden korrekten Matches identifiziert und klassifiziert wurden (vgl. Christen, 2012a, S. 34).

Sowohl die *Accuracy* als auch die *Completeness* werden durch jeden RL-Prozessschritt beeinflusst. Die *Accuracy* wird hauptsächlich von den RL-Prozessschritten Data Preparation, Comparison und Classification beeinflusst. Die *Completeness* wird hauptsächlich vom RL-Prozessschritt Blocking beeinflusst, da alle Datensatzpaare, die durch das Blocking eliminiert

werden, als Kein-Match klassifiziert werden und dies oftmals auch korrekte Matches sein können (vgl. Christen, 2012a, S. 34).

Damit die *Accuracy* und *Completeness* in der Evaluation bewertet werden kann, wird die sogenannte Ground-Truth Datenmenge benötigt. Die Ground-Truth Datenmenge muss alle korrekten Matches der zu integrierenden Datenquellen enthalten. Ground-Truth Daten sind in der Realität selten vorhanden und das Erstellen von Ground-Truth Daten würde das manuelle Verknüpfen der zu integrierenden Datenquellen bedeuten. Dies würde den RL-Prozess obsolet machen. Das manuelle Verknüpfen der zu integrierenden Datenquellen bedeutet allerdings einen hohen manuellen Aufwand, der selten betrieben wird (vgl. Christen, 2012a, S. 34-35).

Wenn ein RL-Prozess für zwei zu integrierende Datenquellen durchgeführt worden ist und für die beiden Datenquellen Ground-Truth Daten vorliegen, werden die Datensatzpaare in die folgenden Kategorien eingeteilt (vgl. Christen, 2012a, S. 165):

True Positive (TP): Die TP Datensatzpaare sind durch den RL-Prozess als Match klassifiziert worden und entsprechen einem korrekten Match. Die Datensatzpaare repräsentieren dieselbe Realwelt-Entität.

False Positive (FP): Die FP Datensatzpaare sind durch den RL-Prozess als Match klassifiziert worden, sind aber kein korrekter Match, da die Datensätze unterschiedliche Realwelt-Entitäten repräsentieren.

True Negative (TN): Die TN Datensatzpaare sind durch den RL-Prozess als Kein-Match klassifiziert worden und die Datensätze repräsentieren unterschiedliche Realwelt-Entitäten, sodass diese tatsächlich Kein-Match sind.

False Negative (FN): Die FN Datensatzpaare sind durch den RL-Prozess als Kein-Match klassifiziert worden, die Datensatzpaare repräsentieren allerdings dieselbe Realwelt-Entität, sodass das Datensatzpaar einen Match darstellt und fälschlicherweise als Kein-Match klassifiziert wurde.

Mit den Ergebnissen eines RL-Prozesses und den dazugehörigen Ground-Truth Daten können die TP, FP, TN und FN berechnet werden. Dabei dienen die Werte für die Berechnung unterschiedlicher Metriken zur Bewertung der Ergebnisqualität, die im Folgenden vorgestellt werden (vgl. Christen, 2012a, S. 166):

Accuracy: Die Accuracy ist eine Kennzahl, die zur Bewertung von Machine Learning Modellen häufig verwendet wird. Dabei ist die Accuracy eine sinnvolle Metrik, wenn die zu klassifizierenden Kategorien in der Ergebnismenge gleichverteilt sind. Im RL sind die klassifizierenden Kategorien in der Ergebnismenge selten gleichverteilt, da es meist mehr

Datensatzpaare aus der Kategorie TN gibt. Daher ist die Accuracy für RL nicht geeignet, wie in Formel 2.13 zu sehen, da eine hohe Anzahl von TN Datensatzpaaren das Ergebnis dominiert (vgl. Christen, 2012a, S. 167).

$$\text{acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{2.13}$$

Precision Die Precision ist eine Kennzahl, die zur Bewertung von Machine Learning Modellen verwendet wird. Da die Precision nicht die Anzahl der TN Datensatzpaare berücksichtigt, siehe Formel 2.14, wirken sich die ungleich verteilten Kategorien nicht auf das Ergebnis der Kennzahl aus (vgl. Christen, 2012a, S. 167).

$$\text{prec} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.14}$$

Die Precision gibt an, wieviele durch den RL-Prozess als Match klassifizierte Datensatzpaare, TP + FP, tatsächlich korrekte Matches, TP, sind. Die Precision gibt an, wie präzise der RL-Prozess korrekte Matches klassifiziert hat (vgl. Christen, 2012a, S. 167).

Recall Der Recall ist eine Kennzahl, die häufig für die Bewertung von Machine Learning Modellen verwendet wird. Auch die Recall Kennzahl ist nicht vom Problem der nicht gleich verteilten Kategorien des Ergebnisses betroffen, siehe Formel 2.15.

$$rec = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.15}$$

Der Recall wird ohne die TN Datensatzpaare berechnet. Mit dem Recall wird berechnet, wie viele der insgesamt vorhanden korrekten Matches, TP + FN, durch den RL-Prozess als TP identifiziert wurden. Daher kann der Recall auch als Trefferquote bezeichnet werden. In den meisten Fällen entsteht ein Trade-off zwischen der Precision und dem Recall bei der Entwicklung eines RL-Prozesses. Entweder ist es wichtig eine hohe Precision zu erzielen, was allerdings für einen niedrigen Recall sorgt oder es wird ein hoher Recall gefordert, was aber eine niedrige Precision bedeutet (vgl. Christen, 2012a, S. 167).

F-measure Die F-measure ist eine Kennzahl, die häufig für die Bewertung von Machine Learning Modellen verwendet wird. Sie bildet das harmonische Mittel aus der Precision und dem Recall, siehe Formel (vgl. Christen, 2012a, S. 168).

$$\text{fmeas} = 2 \times \left(\frac{\text{prec} \times \text{rec}}{\text{prec} + \text{rec}} \right) \tag{2.16}$$

Die F-measure Kennzahl ist am höchsten, wenn sowohl Precision als auch Recall hoch sind. Daher kann mit der F-measure ein Kompromiss zwischen Precision und Recall evaluiert werden (vgl. Christen, 2012a, S. 168).

Specificity Die Specificity Kennzahl wird häufig in der Medizin verwendet. Da auch die Specificity in ihrer Berechnung, siehe Formel 2.17, die TN Datensatzpaare berücksichtigt, leidet auch diese Kennzahl unter dem Problem der nicht gleich verteilten Ergebnis Kategorien (vgl. Christen, 2012a, S. 168).

$$\text{spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.17}$$

Auch die Specificity Kennzahl wird durch die TN Datensatzpaare dominiert, sodass diese Kennzahl nicht für die Evaluation von RL-Prozessen geeignet ist (vgl. Christen, 2012a, S. 168).

Für die Evaluation von RL-Prozessen eignen sich die Kennzahlen Precision, Recall und F-measure, währenddessen die Kennzahlen Accuracy und Specificity eher ungeeignet sind.

Bei der Evaluation von RL-Prozessen müssen die berechneten Ergebnisse bewertet werden. Hierzu sind Ground Truth Daten zwingend erforderlich. Wenn keine Ground Truth Daten verfügbar sind, muss für die Entwicklung eines RL-Prozesses geprüft werden, ob es eine praktische Möglichkeit gibt, qualitative Ground Truth Daten mit einem vertretbaren Aufwand zu generieren. Da dies in der Praxis meistens nicht möglich ist, werden RL-Prozesse in der Regel mit öffentlich oder synthetischen Datensätzen, zu denen Ground Truth Daten existieren, entwickelt (vgl. Christen, 2012a, S. 183). Laut Barlaug und Atle Gulla (2020) ist dies einer der Hauptgründe, weshalb der aktuelle Stand der RL-Forschung bisher wenig in unternehmerische Lösungen berücksichtigt worden ist.

2.2.2 Record Linkage-System

In dieser Arbeit wird ein prototypisches RL-System entwickelt. Daher sollen in diesem Kapitel die unterschiedlichen Definitionen für den Begriff RL-System diskutiert werden, um anschließend eine für diese Arbeit gültige Definition aufzustellen.

Während die Begriffe RL und RL-Prozess über viele Publikationen hinweg konkret definiert sind, finden sich für den Begriff RL-System keine einheitlichen Definitionen. In Tabelle 2.3 sind einige Definition aufgeführt, die im Folgenden diskutiert werden.

Tabelle 2.3: Definitionen Record Linkage-System

Autor	**Definition**
Köpcke und Rahm (2010)	„Entity matching frameworks provide several methods and their combination to effectively solve different match tasks. “
Christen (2012a)	„Today, there are dozens, if not hundreds, of commercial data matching and deduplication products and solutions on the market. Many of these are either stand-alone packages that are specialised for a certain type of application, such as the deduplication of mailing lists or the matching of health databases; or data matching is one component of a much larger business intelligence, data integration, data quality or customer relationship management system.“
Konda et al. (2016b)	„In contrast to current EM systems, which mostly provide a set of implemented matchers/blockers, these new systems are far more advanced. First and foremost, they seek to handle a wide variety of EM scenarios. These scenarios can use very different EM workflows. So it is difficult to build a single system to handle all EM scenarios. Instead, we should build a set of systems, each handling a well-defined set of similar EM scenarios.“
Govind et al. (2018)	„While much progress has been made, current solutions are still limited in that they often require a developer to be involved in the matching process. For example, several recent solutions require a developer to write heuristic rules, called blocking rules, to reduce the number of candidate tuple pairs to be matched, then train and apply a matcher to the remaining pairs to predict matches. The developer must know how to code (e.g., to write rules in Python) and match entities (e.g., to select learning models and features). [...] onsequently, it is increasingly critical that we develop EM solutions that are very easy for lay users to use.“
Doan et al. (2020)	„Thus, the notion of “system” in Magellan has changed. It is no longer a stand-alone monolithic system such as RDBMSs or most current EM systems. Instead, this new “system” spans multiple EEs. Within each EE, it provides a growing ecosystem of interoperable EM tools, situated in a larger ecosystem of DS tools. Finally, it provides detailed guides that tell users how to use these tools to perform EM. “

Papadakis et al. (2021)	„In more detail, ER systems should: (1) be extensible, open-source tools; (2) cover the entire end-to-end pipeline; (3) process data of any structuredness; (4) require no coding from their users; (5) provide guidelines for creating effective solutions; (6) exploit a wide range of techniques.“

Köpcke und Rahm (2010) haben in ihrer Publikation einen Benchmark für RL-Systeme durchgeführt. In ihrer Publikation verwenden sie den synonym zu betrachtenden Begriff „entity matching framework“. Sie definieren das RL-System als ein System, das verschiedene Methoden und deren Kombinationen bereitstellt, um die Aufgaben entlang des RL-Prozsses zu unterstützen.

Während Köpcke und Rahm (2010) das RL-System als Unterstützung für den RL-Prozess definieren, beschreibt Christen (2012a) das RL-System, wie es technisch eigenständig oder als Teillösung eines größeren IT-Systems gesehen werden kann. Christen (2012a) verwendet in seiner Veröffentlichung den Begriff „data matching and deduplication products and solutions“. Er definiert RL-Systeme als Stand-alone-Software, die für spezielle Anwendungsbereiche entwickelt worden ist oder als Softwarekomponente einer größeren Business Intelligence-, Data Integration-, Data Quality- oder Customer Relationship Management-Softwarelösung.

Konda et al. (2016b) fokussieren in ihrer Definition oder Beschreibung das ein RL-System Verfahren und Algorithmen für die RL-Prozessschritte Blocking und Classification bereitstellt. Dabei berücksichtigt das RL-System die Vielfalt der vorkommenden RL-Szenarien, die sich bspw. über die Domäne wie Medizin oder E-Commerce unterscheiden können. Daher definiert Konda et al. (2016b) ein RL-System bestehend aus mehreren RL-Sytemen, die jeweils Algorithmen und Verfahren bereitstellen, um ähnliche RL-Szenarien zu bearbeiten.

Die Definition von Govind et al. (2018) fokussiert bei der Beschreibung eines RL-Systems die Anwenderperspektive. Auch Govind et al. (2018) sieht die Prozessschritte Blocking und Classification von einem RL-System unterstützt. Ein RL-System erfordere einen Entwickler, der am RL-Prozess beteiligt ist. Der Entwickler, der das RL-System nutzt, muss programmieren können und Fachwissen über die zu integrierenden Datenquellen besitzen. Govind et al. (2018) sieht den Bedarf nach RL-Systemen, die auch durch Personal ohne Programmierkenntnisse bedient werden können.

Die Definition von Doan et al. (2020) geht zunächst darauf ein, dass bisherige RL-Systeme als monolithische Systeme implementiert worden sind. Doan et al. (2020) sieht die neuen RL-Systeme als Ökosystem von interoperablen Tools, das in ein größeres Ökosystem von

Data Science Tools integriert ist. Zudem soll ein RL-System Anleitungen bieten, wie die interoperablen Tools zu nutzen sind für den RL-Prozess.

Die Definition von Papadakis et al. (2021) ist eine der umfassendsten, da sie die Anforderungen an RL-Systeme aus verschiedenen Publikationen umfasst. Demnach sollten RL-Systeme Open-Source sein, den gesamten RL-Prozess unterstützen, jegliche Datenstrukturen verarbeiten können, keine Programmierkenntnisse der Anwender erfordern, How-to-Guides für die Nutzung des Systems bereitstellen und eine Vielzahl von Algorithmen und Verfahren zur Verfügung stellen.

Die verschiedenen Definitionen zeigen, dass es unterschiedliche Auffassungen für den Begriff RL-System gibt, obwohl eine Vielzahl von nicht-kommerziellen RL-Systemen, wie bswp. D-Dupe, Febrl oder Dedoop und kommerziellen RL-Systemen, wie bspw. Tamr, Informatica, and IBM InfoSphere, existieren (vgl. Doan et al., 2020).

Zusammenfassend existieren Definitionen die RL-System als Unterstützung des RL-Prozesses sehen, ohne näher ins Detail zu gehen, wie die Definition von Köpcke und Rahm (2010). Weiterhin existieren Definitionen, wie die von Christen (2012a) und Doan et al. (2020), die ein RL-System aus technischer Perspektive als Softwarekomponente oder Ökosystem von Softwaretools definieren. Andere Definitionen, wie die von Konda et al. (2016b), definieren RL-Systeme mit dem Fokus auf die Unterstützung der ähnlichen RL-Szenarien. Definitionen, wie die von Govind et al. (2018), fokussieren in der RL-System Definition die Anwenderperspektive. Die Definition von Papadakis et al. (2021) umfasst alle genannten Perspektiven, da alle Anforderungen zusammenfassend beschrieben werden.

In dieser Arbeit wird versucht eine erste allgemeingültige Definition für den Begriff RL-System zu erstellen. Die folgenden Definition lehnt sich an die gerade beschriebenen Definitionen an und berücksichtigt die Problemstellung dieser Arbeit, die Automatisierung des RL-Prozesses:

Defintion 2.3

Ein **Record Linkage-System** unterstützt alle fünf RL-Prozessschritte, indem es Verfahren, Algorithmen oder auch vortrainierte Machine Learning Modelle bereitstellt, um die Durchführung und Entwicklung des RL-Prozess mit beliebigen Datenquellen für die Anwender möglichst vollständig zu automatisieren. Ein RL-System sollte als Ökosystem aus interoperablen Tools bestehen, um die rasant wachsende Menge an Tools aus dem Data Science Umfeld effizient einbeziehen zu können.

3 Aktueller Stand der Forschung und verwandte Arbeiten

In diesem Kapitel wird die erste Teilforschungsfrage des Forschungsvorhabens adressiert (siehe Abb. 3.1). Die erste Teilforschungsfrage dient zur Identifikation der existierenden Verfahren und Vorgehensweisen im RL. Zur Beantwortung der Teilforschungsfrage wurde zu Beginn des Forschungsvorhabens eine qualitative Inhaltsanalyse durchgeführt, dessen Ergebnisse im Folgenden beschrieben werden. Abschließend werden die relevanten Forschungsarbeiten vorgestellt und die Abgrenzung zu dieser Arbeit beschrieben.

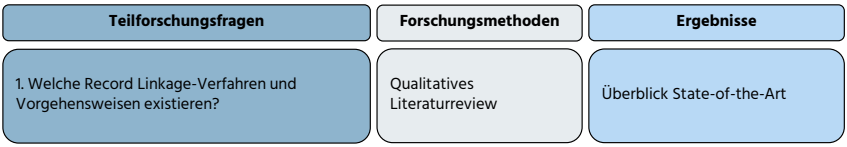


Abbildung 3.1: Einordnung des Literaturreviews in das gesamte Forschungsvorgehen

3.1 Qualitative Inhaltsanalyse Record Linkage

Die qualitative Inhaltsanalyse wird von Wilde und Hess (2007) als qualitative oder quantitative Querschnittsanalyse definiert (vgl. Wilde & Hess, 2007). Die qualitative Inhaltsanalyse ist eine empirische Methode, die es ermöglicht größere Texte unter Berücksichtigung des Kontextes auszuwerten. Im Gegensatz zur qualitativen Inhaltsanalyse beachtet die quantitative Inhaltsanalyse nicht den inhaltlichen Kontext, sondern zählt beispielsweise die Häufigkeit des Auftretens von bestimmten Wörtern in Texten (vgl. Mayring, 2000). Qualitative Inhaltsanalysen helfen Forschungsprobleme- und Forschungslücken zu identifizieren und die Relevanz und Aktualität ihrer Bearbeitung zu begründen, daher eignen sie sich zu Beginn eines Forschungsvorhabens wie einer Promotion (vgl. vom Brocke et al., 2015, S. 9). Sie können als Grundlage und Rahmen für Forschungsprojekte dienen, um das Verständnis für eine Domäne zu entwickeln, das untersuchte Thema zu erklären und die Entwicklung von Theorien zu identifizieren und zu überprüfen (vgl. vom Brocke et al., 2015, S. 9). Laut vom Brocke et al. (2015) können qualitative Inhaltsanalysen das Ergebnis von Forschungsprojekten maßgeblich verbessern[5].

Ausgangspunkt einer qualitativen Inhaltsanalyse ist die zu beantwortende Forschungsfrage,

[5] „Done correctly, they can significantly improve the outcome of research projects.“(vgl. vom Brocke et al., 2015, S. 9)

die wie folgt lautet (vgl. Mayring, 2014, S. 58):

Welche Record Linkage-Verfahren und -Vorgehensweisen existieren?

Webster und Watson (2002) und Cato (2016) beschreiben einen vierstufigen Prozess zur Durchführung einer qualitativen Inhaltsanalyse. Die vier Prozessschritte werden im Folgenden beschrieben (vgl. Webster & Watson, 2002; Cato, 2016):

1. **Definition der Suchstrategie- und Parameter:** Der erste Prozessschritt definiert, in welchen Datenbanken, Journals oder Konferenzen die Publikationen gesucht werden und welche Anforderungen an die Publikationen gestellt werden. Anforderungen können beispielsweise die Länge oder den Veröffentlichungszeitraum der Publikation betreffen. Die wichtigste Anforderung ist die Liste der Suchbegriffe, die in der Publikation enthalten sein sollen, damit diese selektiert wird (vgl. Webster & Watson, 2002; Cato, 2016).

   Die ausgewählten Datenbanken und Suchbegriffe für die qualitative Inhaltsanalyse dieser Arbeit sind in Tabelle 3.1 aufgeführt.

   Tabelle 3.1: Suchstrategie der qualitativen Inhaltsanalyse

Datenbanken	**Suchbegriffe**
• IEEE • Science Direct • ACM Digital Library	• Record Linkage • Entity Matching • Duplicate Detection • Entity Resolution • Entity Reconciliation

   Die Auswahl der Suchbegriffe erfolgte über eine erste Literaturrecherche zum Begriff RL und auf Basis eines vorangegangen Literaturreviews von Enríquez, Domínguez-Mayo, Escalona, Ross und Staples (2017). Auf dieser Basis wurden die Synonyme *Entity Matching*, *Duplicate Detection*, *Entity Resolution* und *Entity Reconciliation* identifiziert. Alle Suchbegriffe wurden in der Suchanfrage über einen ODER-Operator verknüpft. Die Suchanfragen wurden in den Datenbanken *IEEE*[6], *Science Direct*[7] und *ACM Digital Library*[8] ausgeführt. Der Suchzeitraum erstreckt sich von 2000 bis 2018, da das Literaturreview zu Beginn des Forschungsvorhabens im Jahr 2018 durchgeführt worden ist. Zudem müssen die

[6] https://ieeexplore.ieee.org/
[7] https://www.sciencedirect.com/
[8] https://dl.acm.org/

Suchbegriffe mindestens im Titel, Abstract oder den Keywords der Publikation enthalten sein.

2. **Identifikation und Selektion relevanter Publikationen:** In diesem Prozessschritt wird zuerst die Literatursuche mit den definierten Suchparametern durchgeführt. Daraufhin erfolgt ein erstes Screenings der Publikationen anhand des Titels und der Zusammenfassung. Anschließend werden die gefilterten Publikationen durch ein Volltextscreening weiter gefiltert (vgl. Webster & Watson, 2002; Cato, 2016).

   Die Suchanfragen wurden in den ausgewählten und genannten Datenbanken durchgeführt und durch ein Screening des Titels und der Zusammenfassung wurden die relevanten Publikationen selektiert. Die Ergebnisse der Suchanfrage und des Screenings sind in Tabelle 3.2 zusammengefasst. Die Suchanfrage in der IEEE Datenbank ergab 588 Treffer, von denen 21 als relevant klassifiziert wurden. Die Suchanfrage in der Science Direct Datenbank ergab 637 Treffer, von denen 17 als relevant klassifiziert wurden. Die Suchanfrage in der ACM Digital Library ergab 127 Treffer, von denen 23 als relevant klassifiziert wurden.

3. **Vorwärts- und Rückwärtssuche:** Mithilfe der Rückwärtssuche wird überprüft, ob in den relevanten Publikationen weitere relevante ältere Publikationen referenziert werden. Mit der Vorwärtssuche wird überprüft, ob die relevanten Publikationen in jüngeren Publikationen referenziert werden, die für die qualitative Inhaltsanalyse relevant sind (vgl. Webster & Watson, 2002; Cato, 2016).

   Die Durchführung der Vorwärts- und Rückwärtssuche hat 7 weitere relevante Publikationen ergeben (siehe Tabelle 3.2). Damit wurden insgesamt 68 relevante Publikationen identifiziert. Die Übersicht der 68 Publikationen ist in Anhang A Tabelle A.1 zu finden. Die relevanten Publikationen werden im Folgenden weiter analysiert.

Tabelle 3.2: Ergebnisse der durchgeführten Suchanfragen

Datenbank	**Ergebnis**	**Relevante Publikationen**
IEEE	588	21
ScienceDirect	637	17
ACM Digital Library	127	23
Vorwärts- und Rückwartssuche	/	7
Summe	1.352	68

4. **Analyse der relevanten Publikationen:** Im letzten Prozessschritt werden die relevanten Publikationen analysiert. Mayring unterscheidet zwischen drei folgenden Grundformen, um Texte zu interpretieren (vgl. Mayring, 2015, S. 67-68; Mayring, 2014, S. 64):

   Zusammenfassung: Diese Grundform hat das Ziel, das Material zu reduzieren, sodass

die wesentlichen Inhalte erhalten bleiben. Am Ende soll durch die zusammenfassende Abstraktion des Grundmaterials ein repräsentativer und überschaubarer Korpus geschaffen werden.

Explikation: Mit der Grundform Explikation wird zu fraglichen Textteilen, wie bspw. Begriffen oder Sätzen, weiteres Material zusammengetragen, um den entsprechenden Textteil zu erklären, zu erläutern oder besser deuten zu können.

Strukturierung: Mit der Strukturierung werden relevante Aspekte, anhand zuvor definierter Kriterien, aus dem Material herausgearbeitet. Dadurch wird es ermöglicht einen Querschnitt durch das Material zu legen oder dieses nach bestimmten Kriterien zu beurteilen.

Diesen drei Grundformen ordnet Mayring konkrete Analyseverfahren zu, die in Tabelle 3.3 aufgeführt sind. Es ist möglich die Analyseverfahren der drei Grundformen miteinander zu kombinieren. Hierdurch ergibt sich die vierte Grundform KOMBINATION, die in in Tabelle 3.3 aufgeführt ist (vgl. Mayring, 2014, S. 104).

Tabelle 3.3: Analyseverfahren der qualitativen Inhaltsanalyse (vgl. Mayring, 2015, S. 68; Mayring, 2014, S. 65)

Grundform	**Analyseverfahren**
Zusammenfassung	Zusammenfassung
	Induktive Kategoriebildung
Explikation	Enge Kontextanalyse
	Weite Kontextanalyse
Deduktive Strukturierung	Formale Kategorien
	Inhaltliche Kategorien
	Typisierende Kategorien
	Skalierende Kategorien
Kombination	Inhaltliche Strukturierung und Themenanalyse

In dieser Arbeit wird das Analyseverfahren INHALTLICHE STRUKTURIERUNG UND THEMENANALYSE verwendet. Mayring beschreibt dieses Verfahren als einen zweistufigen Prozess, der zunächst die deduktive Strukturierung vorsieht. Mit den Ergebnissen der deduktiven Strukturierung wird dann die induktive Kategoriebildung durchgeführt, um die vorangestellte Forschungsfrage zu beantworten (vgl. Mayring, 2014, S. 104). Die Analyseform wurde gewählt, um die Ergebnisse nicht zu restriktiv zu limitieren. Der durchgeführte Prozess sieht wie folgt aus Mayring (2014):

1. Deduktiver Teil des Prozesses
   (a) Definition der Kategorien (Haupt- und Unterkategorien) aus der Theorie

(b) Definition des Kodierleitfadens (Definitionen, Ankerbeispiele und Kodierregeln)

(c) Analyse der Publikationen, Sätze und Absätze den Kategorien zuordnen, Ankerbeispiele dokumentieren

(d) Überprüfung der Kategorien und Kodierregeln nach 10-50% der zu analysierenden Publikationen

2. Induktiver Teil des Prozesses

(a) Durcharbeiten der zuvor kategorisierten Sätze und Absätze, um induktive Kategorien zu bilden

(b) Überprüfung der Kategorien und Kodierregeln nach 10-50% der zu analysierenden Publikationen

(c) Finales Überprüfen der Publikationen und Bilden von Hauptkategorien, falls sinnvoll

(d) Analyse der Haupt- und Unterkategorien

Die Kategorien für den deduktiven Teil des Prozesses basieren auf den RL-Prozessschritten Data Preparation, Blocking, Comparison und Classification (siehe Tabelle 3.4). Um die relevanten Vorgehensweisen und Verfahren im RL zur Beantwortung der Forschungsfrage zu identifizieren, wurden zusätzlich die Kategorien Datenset, Record Linkage-System und Forschungsziel hinzugefügt. Alle 68 Publikationen wurden analysiert und die Sätze und Absätze wurden den definierten Kategorien zugeordnet. Anschließend wurden aus diesen Sätzen und Absätzen die induktiven Kategorien gebildet. Die induktiv gebildeten Kategorien mit der Beziehung zu den deduktiven Kategorien ist in Tabelle 3.4 aufgeführt.

Tabelle 3.4: Deduktive Kategorien und abgeleitete induktive Kategorien

Deduktive Kategorie	**Induktive Kategorie**
Forschungsziel	Fokussiertes Record Linkage-Ziel
Datenset	Verwendete Datensets, Datenstruktur, Realwelt-Entität
Data Preparation	Verwendete Methoden
Blocking	Verwendete Algorithmen
Comparison	Verwendete Ähnlichkeitsmaße
Classification	Verwendete Algorithmen, Kategorie Algorithmus
Record Linkage-System	Verwendetes System

3.2 Deskriptive Analyse der durch die Inhaltsanalyse generierten Daten

Durch die Suchstrategie wurden 68 relevante Publikationen identifiziert. In Abbildung 3.2 ist die Verteilung der relevanten Publikationen über die Jahre dargestellt. Ab dem Jahr 2014 ist ein Anstieg der relevanten Publikationen gegenüber den vorherigen Jahren zu erkennen. Die beiden Jahre 2016 und 2018 weisen mit 15 und 11 relevanten Publikationen die höchste Anzahl aus (siehe Abb. 3.2).

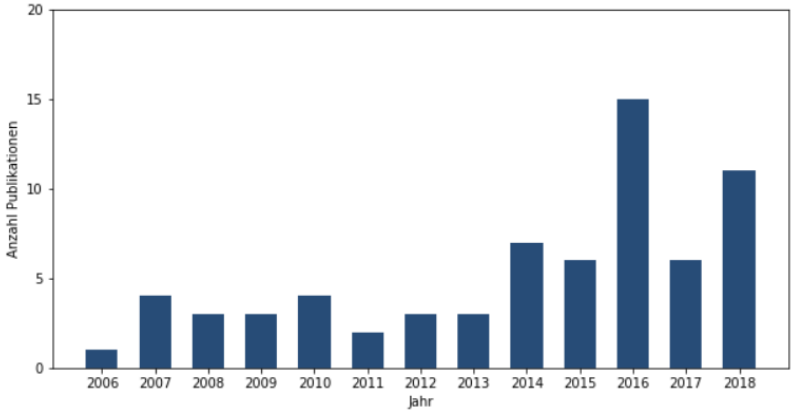


Abbildung 3.2: Anzahl der Publikationen pro Jahr

3.2.1 Auswertung der Kategorie Fokussiertes Record Linkage-Ziel

In Tabelle 3.5 ist aufgeführt, welches RL-Ziel die Publikationen fokussieren. 24 Publikationen fokussieren ihre Forschung auf den RL-Prozessschritt Classification. Kooli et al. (2018) wenden verschiedene Klassifikationsalgorithmen an, um eine RL-Aufgabe zu lösen. So wird diese Publikation mit dem Forschungsschwerpunkt auf den Prozessschritt Classification kategorisiert. 23 Publikationen fokussieren den Record Linkage-Prozess. Shu et al. (2012) führen den gesamten RL-Prozess mit einem entwickelten Framework durch. So wird diese Publikation mit dem Forschungsfokus auf den Record Linkage-Prozess eingeordnet. Koudas, Sarawagi und Srivastava (2006), Elmagarmid, Ipeirotis und Verykios (2007), Wandelt et al. (2014), Enríquez et al. (2017), El-Ghafar et al. (2017) und Fier, Augsten, Bouros, Leser und Freytag (2018) liefern ein Literatur Review über RL. Zwei der Literatur Reviews fokussieren auf Similarity Join Algorithmen (vgl. Fier et al., 2018; Wandelt et al., 2014). Zwei der Literatur Reviews geben einen Überblick über RL, sind aber älter als zehn Jahre (vgl. Elmagarmid et al., 2007; Koudas et al., 2006). Die letzten beiden Literatur

Reviews konzentrieren sich auf die Beziehung zwischen Big Data und RL (vgl. El-Ghafar et al., 2017; Enríquez et al., 2017). Peled, Fire, Rokach und Elovici (2016) wenden verschiedene Ähnlichkeitsmaße an, um Entitäten über soziale Online-Netzwerke hinweg abzugleichen. Diese Publikationen werden dem RL-Ziel COMPARISON zugeordnet (vgl. Peled et al., 2016).

Tabelle 3.5: Qualitative Inhaltsanalyse Kategorie Fokussiertes Record Linkage-Ziel

Ziel der Arbeit	**Anzahl**	**ID**
Classification	24	ID_2, ID_5, ID_7, ID_8, ID_10, ID_12, ID_14, ID_21, ID_22, ID_23, ID_34, ID_35, ID_36, ID_37, ID_47, ID_49, ID_55, ID_60, ID_61, ID_63, ID_64, ID_65, ID_67, ID_68
Record Linkage-Prozess	23	ID_3, ID_15, ID_16, ID_17, ID_18, ID_19, ID_20, ID_24, ID_25, ID_26, ID_29, ID_31, ID_32, ID_38, ID_40, ID_41, ID_42, ID_43, ID_46, ID_51, ID_52, ID_56, ID_59
Literatur Review	6	ID_1, ID_4, ID_30, ID_53, ID_54, ID_62
Blocking	5	ID_9, ID_11, ID_45, ID_48, ID_50
Comparison	5	ID_21, ID_27, ID_28, ID_47, ID_66
Data Preparation	3	ID_33, ID_57, ID_66
Record Linkage-Software Testing	2	ID_39, ID_58
Benchmark von RL-Systemen	1	ID_13

Fünf Publikationen fokussieren ihre Forschung auf den RL-Prozessschritt BLOCKING und versuchen diesen zu verbessern (vgl. Gómez-Bao, Larriba-Pey & Ribes Puig, 2009,; Mishra, Saha & Mondal, 2016; Simonini, Bergamaschi & Jagadish, 2016; van Dam et al., 2016; de Vries, Ke, Chawla & Christen, 2009). Drei Publikationen fokussieren den Prozessschritt DATA PREPARATION (vgl. Liu, Kumar & Thomas, 2015; Marple, Desmarais & Young, 2017; Prabhu & Gnana Dhas, 2018). So versuchen beispielsweise Marple et al. (2017), die Datenintegration mit externen Datenquellen zu verbessern und legen dabei ihren Fokus auf den Prozessschritt DATA PREPARATION (vgl. Marple et al., 2017). Die Autoren Blanco et al. (2018) und Enríquez, Blanco, Domínguez-Mayo, Tuya und Escalona (2016) fokussieren ihre Forschung auf das RECORD LINKAGE-SOFTWARE TESTING. Eine Publikation führt einen BENCHMARK VON RL-SYSTEMEN durch (vgl. Köpcke & Rahm, 2010).

3.2.2 Auswertung der Kategorie Verwendete Datensets

In Tabelle 3.6 sind die acht meist genutzten Datensets in den Publikationen aufgeführt. Zu den am häufigsten verwendeten Datensets zählt das DBLP Datenset. Das Datenset wird in Kombination mit dem Datenset ACM, dem Datenset SCHOLAR oder allein verwendet, um RL zu erforschen. Die Datensets beinhalten bibliografische Daten und enthalten die Attribute Autor, Titel, Jahr und Venue. In der Publikation von Köpcke, Thor und Rahm (2010) werden

die Datensets DBLP-ACM und DBLP-Scholar als Benchmark Datensets[9] veröffentlicht. Zu den veröffentlichten Benchmark Datensets gehört auch das Datenset Abt-Buy, welches in drei Publikationen verwendet wurde. Das Datenset Abt-Buy enthält Produktdaten mit den Attributen Name, Beschreibung und Preis.

Tabelle 3.6: Qualitative Inhaltsanalyse Kategorie Verwendete Datensets

Datenset	Anzahl	ID
DBLP-ACM	8	ID_6, ID_13, ID_14, ID_22, ID_27, ID_48, ID_55, ID_65
DBLP	6	ID_10, ID_41, ID_44, ID_52, ID_60, ID_62
DBLP-Scholar	6	ID_6, ID_14, ID_22, ID_27, ID_55, ID_65
Restaurant	6	ID_3, ID_6, ID_13, ID_55, ID_65, ID_68
CORA	5	ID_25, ID_52, ID_55, ID_60, ID_68
FEBRL	4	ID_3, ID_9, ID_11, ID_25
Abt-Buy	3	ID_14, ID_27, ID_65
Census	3	ID_3, ID_13, ID_68

Das Datenset Restaurant wird in sechs Publikationen verwendet und die Attribute Name, Adresse, Stadt, Telefonnummer, Typ und Klasse. Mudgal et al. (2018) verwenden das Restaurant Datenset und stellen weitere Datensets in einem GIT-Repository[10] zur Verfügung. Das CORA[11] Datenset wird in fünf Publikationen verwendet und das Census[12] Datenset in drei Publikationen. Simonini, Papadakis, Palpanas und Bergamaschi (2018) verwenden beide Datensets in ihrer Publikation. Das CORA Datenset enthält bibliografische Daten mit den Attributen authors, booktitle, date, editor, id, institution, journal, month, note, pages, publisher, tech, title, type, volume and year. Das Census Datenset enthält Personendaten mit den Attributen first_name, last_name, middle_name, street_address und zip_code. Das Datenset FEBRL wird in vier Publikationen genutzt. Unter dem Begriff FEBRL werden Datensets beschrieben, die durch den FEBRL Datenset Generator, der von Christen (2005) entwickelt wurde, entstanden sind. Das FEBRL Datenset besteht aus Personendaten mit den Attributen Vorname, Nachname und Adressattributen. Zudem können mit dem FEBRL Datenset Generator zufällige Duplikate erzeugt werden (vgl. Christen, 2005).

3.2.3 Auswertung der Kategorie Realwelt-Entität

In Tabelle 3.7 sind die acht häufigsten Realwelt-Entitäten in den Datensets der Publikationen aufgeführt. In 42 Publikationen ist die Realwelt-Entität Person in den Datensätzen

[9] https://dbs.uni-leipzig.de/research/projects/object_matching/benchmark_datasets_for_entity_resolution
[10] https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md
[11] https://hpi.de/naumann/projects/repeatability/datasets/cora-dataset.html
[12] https://hpi.de/naumann/projects/repeatability/datasets/census-dataset.html

vorhanden. Publikationen, die das Census Datenset nutzen, integrieren die Daten über die Realwelt-Entität PERSON und fallen in dieses Kategorie. Publikationen die bibliographische Daten verwenden wie bspw. DBLP-ACM, in denen ebenfalls Personendaten enthalten sind, werden auch der Kategorie Realwelt-Entität PERSON zugeordnet. Die Realwelt-Entität BIBLIOGRAPHIE ist in 22 Publikationen vorhanden und umfasst alle Datensets, die Publikationen, Musiktitel oder Filmtitel beinhalten. In elf Publikationen werden Datensets verwendet, die die Realwelt-Entität PRODUKT repräsentieren.

Tabelle 3.7: Qualitative Inhaltsanalyse Kategorie Realwelt-Entität

Realwelt-Entität	Anzahl	ID
Person	42	ID_2, ID_3, ID_5, ID_6, ID_8, ID_9, ID_10, ID_11, ID_12, ID_14, ID_15, ID_16, ID_17, ID_20, ID_22, ID_23, ID_25, ID_26, ID_27, ID_28, ID_29, ID_33, ID_34, ID_35, ID_38, ID_41, ID_45, ID_47, ID_48, ID_51, ID_52, ID_55, ID_56, ID_57, ID_58, ID_59, ID_60, ID_61, ID_63, ID_64, ID_65, ID_68
Bibliographie	22	ID_5, ID_6, ID_8, ID_10, ID_14, ID_20, ID_22, ID_23, ID_27, ID_28, ID_34, ID_41, ID_48, ID_51, ID_52, ID_55, ID_58, ID_59, ID_60, ID_64, ID_65, ID_68
Produkt	11	ID_14, ID_19, ID_21, ID_27, ID_33, ID_36, ID_40, ID_48, ID_50, ID_65, ID_67
Restaurant	4	ID_3, ID_6, ID_13, ID_55, ID_65, ID_68
Unternehmen	4	ID_22, ID_42, ID_57, ID_65
Geografische Location	3	ID_24, ID_30, ID_46
Gene	2	ID_7, ID_66
Hotel	2	ID_60, ID_67

In sechs Publikationen werden Datensets verwendet, die die Realwelt-Entität RESTAURANT repräsentieren. In vier Publikationen werden Datensets verwendet, die die Realwelt-Entität UNTERNEHMEN repräsentieren. Beispielsweise nutzen Mudgal et al. (2018) ein Datenset mit der Realwelt-Entität UNTERNEHMEN, das aus dem Attribut Unternehmensbeschreibung besteht. In drei Publikationen werden Datensets verwendet, die die Realwelt-Entität GEOGRAFISCHE LOCATION repräsentieren. In der Publikation von Dalvi, Olteanu, Raghavan und Bohannon (2014) wird das WikiMapia Datenset mit 20 Millionen Datensätzen verwendet, das die Attribute place_name, latitude und longitude besitzt. In zwei Publikationen werden Datensets verwendet, die die Realwelt-Entität GENE repräsentieren, wie bspw. die Datensets NCBI GenBank und Gene Ontology. Ebenfalls in zwei Publikationen werden Datensets verwendet, die die Realwelt-Entität HOTEL repräsentieren, wie bspw. in der Publikation von Schneider, Mukherjee und Dragut (2018), die selbst gesammelte Hoteldaten[13] verwendet.

[13] u.a. von `Hotels.com`

3.2.4 Auswertung der Kategorie Datenstruktur

In Tabelle 3.8 ist aufgeführt, wie viele Publikationen STRUKTURIERTE, SEMI-STRUKTURIERTE und UNSTRUKTURIERTE Datenquellen für ihre RL-Forschung verwenden. 51 Publikationen verwenden STRUKTURIERTE Datenquellen. Zehn Publikationen verwenden UNSTRUKTURIERTE Datenquellen. Ein Beispiel für die Kategorie UNSTRUKTURIERTE Datenquelle liefert Mudgal et al. (2018), da sie eine Datenquelle verwenden, die eine unstrukturierte Unternehmensbeschreibung enthält. Lediglich eine Publikation, die von Leitão, Calado und Weis (2007), verwendet eine SEMI-STRUKTURIERTE Datenquelle im XML-Format für ihre Forschung.

Tabelle 3.8: Qualitative Inhaltsanalyse Kategorie Datenstruktur

Datenstruktur	Anzahl	ID
strukturiert	51	ID_2, ID_6, ID_7, ID_8, ID_9, ID_10, ID_11, ID_12, ID_13, ID_14, ID_16, ID_17, ID_18, ID_19, ID_20, ID_22, ID_23, ID_24, ID_25, ID_26, ID_27, ID_29, ID_31, ID_33, ID_34, ID_35, ID_37, ID_38, ID_40, ID_41, ID_42, ID_43, ID_44, ID_45, ID_46, ID_47, ID_48, ID_51, ID_52, ID_55, ID_56, ID_57, ID_59, ID_60, ID_61, ID_62, ID_63, ID_64, ID_65, ID_67, ID_68
unstrukturiert	10	ID_14, ID_19, ID_27, ID_33, ID_36, ID_40, ID_47, ID_65, ID_67, ID_68
semi-strukturiert	1	ID_5

3.2.5 Auswertung der Kategorie Data Preparation

In Tabelle 3.9 sind die am häufigsten genannten Data Preparation Methoden in den Publikationen aufgeführt, die häufiger als ein Mal genannt worden sind. Am häufigsten beschreiben die Publikationen DATA CLEANING Methoden.

Tabelle 3.9: Qualitative Inhaltsanalyse Kategorie Data Preparation

Data Preparation	Anzahl	ID
Data Cleaning	10	ID_19, ID_29, ID_35, ID_40, ID_42, ID_46, ID_51, ID_57, ID_59, ID_64
Word Embedding	4	ID_40, ID_64, ID_65, ID_67
Feature Engineering	4	ID_33, ID_47, ID_51, ID_57
n. a.	39	ID_1, ID_2, ID_3, ID_5, ID_8, ID_9, ID_10, ID_11, ID_12, ID_13, ID_14, ID_15, ID_17, ID_18, ID_20, ID_21, ID_23, ID_25, ID_26, ID_27, ID_28, ID_30, ID_31, ID_32, ID_34, ID_36, ID_37, ID_39, ID_41, ID_44, ID_45, ID_49, ID_52, ID_54, ID_58, ID_60, ID_62, ID_63, ID_68

Unter Data Cleaning werden allgemeine Data Preparation Ansätze zusammengefasst. Beispielsweise verwenden Medhat, Hassan und Salama (2015) Regeln, um Interpunktionen wie Komma, Semikolon oder Doppelpunkt zu entfernen, alle Buchstaben in Klein- oder Großbuchstaben umzuwandeln oder Leerzeichen zu entfernen (vgl. Medhat et al., 2015). Außerdem wird auch das Anwenden von Synonymlisten unter der Methode Data Cleaning zusammengefasst, wie es bspw. Köpcke et al. (2012) beschreiben. In vier Publikationen wird Feature Engineering als Data Preparation Methode beschrieben. Die Publikation von F. Wang und Wang (2016) beschreibt, wie aus bestehenden Attributen weitere Attribute abgeleitet und gebildet werden, um den RL-Prozess zu optimieren. Ein solches Vorgehen fällt unter die Methode Feature Engineering. In vier Publikationen werden Word Embeddings als Data Preparation Methode eingesetzt, um Worte in Vektoren umzuwandeln, wie in den Publikationen von Mudgal et al. (2018) und Schneider et al. (2018). In 39 Publikationen wird nicht beschrieben, welche Methoden für den RL-Prozessschritt Data Preparation verwendet worden sind.

3.2.6 Auswertung der Kategorie Blocking

In Tabelle 3.10 sind die am häufigsten genannten Blockingverfahren in den Publikationen aufgeführt. Es sind nur Blockingverfahren aufgeführt, die in mehr als einer Publikation genannt worden sind. Dreizehn Publikationen verwenden Standard Blocking. Standard Blocking bedeutet, dass die am besten geeigneten Attribute ausgewählt werden und diese als Ganzes oder in Teilen zu einem Blocking Key verkettet werden. Beispielsweise nutzen Conrad, Dozier, Molina-Salgado, Thomas und Veeramachaneni (2011) die Attribute Nachname und den ersten Buchstaben des Vornamens als Blocking Key. Kooli et al. (2018) nutzen verschiedene Blocking Keys, wie die ersten Buchstaben des Unternehmensnamens kombiniert mit der Postleitzahl oder die ersten Worte des Titels einer Publikation. In drei Publikationen wurde Locality Sensitive Hashing verwendet (vgl. Karapiperis, Gkoulalas-Divanis & Verykios, 2016; Kong et al., 2016; van Dam et al., 2016). Ebenfalls in drei Publikationen wird Sorted Neighborhood für das Blocking angewendet (vgl. Elmagarmid et al., 2007; El-Ghafar et al., 2017; Simonini et al., 2018). Köpcke und Rahm (2008) und El-Ghafar et al. (2017) verwenden die Methode Q-Gram in ihren Publikationen für das Blocking. In den Publikationen von de Vries et al. (2009) und Simonini et al. (2018) wird die Suffix Array Blocking Methode verwendet.

3.2.7 Auswertung der Kategorie Comparison

Tabelle 3.11 listet die verwendeten Ähnlichkeitsmaße im RL-Prozessschritt Comparison auf, die in den Publikationen verwendet wurden. Die am häufigsten genannten Ähnlichkeitsmaße

Tabelle 3.10: Qualitative Inhaltsanalyse Kategorie Blocking

Kategorie_Blocking	Anzahl	ID
Standard Blocking	12	ID_3, ID_14, ID_16, ID_19, ID_26, ID_38, ID_42, ID_52, ID_53, ID_61, ID_64, ID_68
Locality Sensitive Hashing	3	ID_41, ID_43, ID_50
Sorted Neighborhood	3	ID_4, ID_53, ID_68
Q-Gram	3	ID_6, ID_25, ID_53
Suffix Array Blocking	2	ID_11, ID_68

sind Levenshtein, Jaro Winkler, Jaccard, Soundex, Kosinus-Distanz, Jaro, Smith-Waterman, NYSIIS und Euklidische-Distanz. Metaphone und Soundex sind phonetische Ähnlichkeitsmaße. Vier Publikationen nutzen Word Embeddings, wie Word2Vec, Glove, FastText oder die Word Movers Distance, mit denen semantische Ähnlichkeiten berücksichtigt werden können. Kooli et al. (2018) verwenden das Word Embedding Word2Vec, um semantische Ähnlichkeiten von Zeichenketten zu berücksichtigen (vgl. Kooli et al., 2018). Die Word Embeddings GloVe und FastText werden ebenfalls verwendet, aber nur in der Publikation von Mudgal et al. (2018).

Tabelle 3.11: Qualitative Inhaltsanalyse Kategorie Comparison

Comparison	Anzahl	ID
Levenshtein	15	ID_1, ID_4, ID_6, ID_15, ID_16, ID_21, ID_24, ID_27, ID_29, ID_30, ID_35, ID_38, ID_51, ID_55, ID_59
Jaccard	10	ID_6, ID_19, ID_22, ID_27, ID_43, ID_44, ID_45, ID_47, ID_53, ID_62
Jaro Winkler	10	ID_1, ID_3, ID_6, ID_16, ID_29, ID_35, ID_47, ID_51, ID_55, ID_59
Soundex	6	ID_1, ID_4, ID_29, ID_35, ID_38, ID_47
Kosinus-Distanz	5	ID_6, ID_21, ID_24, ID_44, ID_65
Jaro	4	ID_4, ID_29, ID_51, ID_55
Word Embedding	4	ID_56, ID_64, ID_65, ID_67
Smith-Waterman	3	ID_4, ID_16, ID_55
NYSIIS	3	ID_4, ID_29, ID_35
Euklidische-Distanz	3	ID_43, ID_56, ID_65
Metaphone	2	ID_4, ID_35
Longest Common Sub-string	2	ID_43, ID_47

3.2.8 Auswertung der Kategorie Classification

Tabelle 3.12 zeigt die Anzahl der verwendeten Algorithmen im RL-Prozessschritt Classification in den Publikationen. Aufgeführt sind die Algorithmen, die in mehr als einer Publikationen genutzt worden sind. Regelbasierte Ansätze werden in 18 Publikationen verwendet, wie zum Beispiel in Ferguson, Hannigan und Stack (2018), Jupin und Shi (2014) und Kobayashi, Eram und Talburt (2018). Am zweit- und dritthäufigsten verwendet, folgen die Supervised Learning Ansätze Support Vektor Machine und Decision Tree. Ein Beispiel für die Verwendung von graphenbasierten Verfahren, die in vier Publikationen verwendet werden, liefern Liu et al. (2015). Unter den Algorithmen Neuronales Netz, mit vier Nennungen, werden verschiedene Neuronale Netz Architekturen zusammengefasst, bspw. die verwendete Recurrent Neural Network Architektur in Mudgal et al. (2018) und die verwendete Convolutional Neural Network Architektur in Gottapu, Dagli und Ali (2016). Das Paper von Nentwig, GroB und Rahm (2016) ist eines der drei, die Clustering Algorithmen verwenden. Die letzten drei Algorithmen mit drei Erwähnungen sind die Latent Dirichlet Allocation, die logistische Regression und Naive Bayes.

Tabelle 3.12: Qualitative Inhaltsanalyse Kategorie Classification

Classification Algorithmus	Anzahl	ID
Regelbasierter Ansatz	16	ID_7, ID_9, ID_12, ID_14, ID_15, ID_17, ID_21, ID_24, ID_26, ID_29, ID_38, ID_50, ID_56, ID_61, ID_63, ID_66
Support Vector Machine	13	ID_1, ID_3, ID_4, ID_6, ID_14, ID_16, ID_19, ID_20, ID_36, ID_51, ID_52, ID_55, ID_64
Decision Tree	7	ID_6, ID_8, ID_14, ID_20, ID_42, ID_47, ID_64
Graphbasierte Verfahren	4	ID_5, ID_20, ID_23, ID_33
Neuronales Netz	4	ID_40, ID_47, ID_64, ID_65
Logistische Regression	3	ID_6, ID_8, ID_63
Latent Dirichlet Allocation	3	ID_4, ID_10, ID_36
Naive Bayes	3	ID_1, ID_20, ID_64
Clustering	3	ID_21, ID_25, ID_46

3.2.9 Auswertung der Kategorie Record Linkage-System

Tabelle 3.13 zeigt, welche RL-Systeme in den Publikationen verwendet worden sind. In drei Publikationen wird das RL-System FEBRL genutzt, dass von Christen, Churches und Hegland (2004) entwickelt worden ist. TAILOR, STEM, MARLIN, MOMA und SERF werden im RL-System Benchmark von Köpcke und Rahm (2010) aufgeführt. Die in den jüngsten Publikationen verwendeten RL-Systeme sind Magellan und Deepmatcher, die beide an der Universität Madison-Wisconsin entwickelt worden sind (vgl. Mudgal et al., 2018).

Tabelle 3.13: Qualitative Inhaltsanalyse Kategorie RL-System

RL-System	Anzahl	ID
FEBRL	3	ID_4, ID_13, ID_14
TAILOR	2	ID_4, ID_13
STEM	2	ID_6, ID_13
MARLIN	2	ID_13, ID_14
MOMA	1	ID_13
SERF	1	ID_13
OYSTER	1	ID_63
Magellan	1	ID_65
DeepMatcher	1	ID_65

3.3 Bewertung der deskriptiven Analyse und Ableitung des weiteren Forschungsbedarfs

Ziel der qualitativen Inhaltsanalyse ist es, aufzuzeigen, welche RL-Verfahren und Vorgehensweisen existieren. Die Ergebnisse zeigen, dass die Bedeutung von RL in den letzten Jahren stetig zugenommen hat. Im RL-Prozessschritt Classification ist die evolutionäre Entwicklung von regelbasierten Algorithmen zu Supervised und von Unsupervised Learning hin zu Deep Learning Algorithmen wahrzunehmen (siehe Tabelle 3.11). In RL-Publikationen werden neben strukturierten Daten seit 2010 auch unstrukturierte Datenquellen (siehe Tabelle 3.8) verwendet, wie bspw. Produktbeschreibungen. Die meisten RL-Publikationen haben das Forschungsziel den Prozessschritt Classification zu optimieren oder betrachten den RL-Prozess für ein spezielles RL-Problem, das aus zwei Datenquellen besteht (siehe Tabelle 3.5). Um diese beiden Forschungsziele zu erreichen, werden am häufigsten die bestehenden Benchmark Datensets von Köpcke und Rahm (2010) und die erweiterten Benchmark Datensets aus dem Magellan Projekt[14] verwendet (siehe Tabelle 3.6). Am häufigsten werden dabei die Realwelt-Entitäten Person, Bibliographie (Buch, Publikation, Musik- oder Filmtitel) oder Produkt integriert (siehe Tabelle 3.7). Dies steht in Zusammenhang mit den verwendeten Benchmark Datensets, die diese Realwelt-Entitäten beinhalten. In den Publikationen werden wenige Datenquellen genutzt und selten wird eine Realwelt-Entität fokussiert. Lediglich Köpcke et al. (2012) fokussieren das Produkt als Realwelt-Entität (siehe Tabelle 3.5). Die Auswertung der Kategorie Data Preparation zeigt, dass in diesem Prozessschritt nur drei Algorithmen und Verfahren mehrfach verwendet worden sind und in einem Großteil der Publikationen, 39 von 68, keine näheren Angaben zum RL-Prozessschritt Data Preparation gemacht wurden (siehe Tabelle 3.9). In den übrigen Publikationen sind die verwendeten Data Preparation Vorgehensweisen sehr heterogen. Allerdings ist für eine Rekonstruktion der For-

[14] https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md

schungsergebnisse durch Dritte eine Beschreibung der Vorgehensweise im RL-Prozessschritt Data Preparation notwendig. Die Auswertung der Kategorie Blocking zeigt, dass in diesem RL-Prozessschritt einige Algorithmen und Verfahren in mehreren Publikationen verwendet werden (siehe Tabelle 3.10). Da in 43 von 68 Publikationen kein Blockingverfahren oder Algorithmus genannt wird, scheinen die aufgeführten Blocking Algorithmen und Verfahren für den RL-Prozess favorisiert zu werden. Diese Erkenntnis bestätigen Papadakis, Mandilaras et al. (2020), die ebenfalls feststellen, dass der attribut-basierte Blocking Key und Sorted Neighbourhood die populärsten Blockingverfahren sind (vgl. Papadakis, Mandilaras et al., 2020, S. 32). Die Auswertung der Kategorie Comparison (siehe Tabelle 3.11) zeigt, dass viele verschiedene Ähnlichkeitsmaße in den Publikationen verwendet werden. Dabei werden ausschließlich String Ähnlichkeitsmaße mehrfach verwendet. Die klassischen zeichenbasierten Ähnlichkeitsmaße, wie Levenshtein und Jaro Winkler, token-basierten Ähnlichkeitsmaße, wie Jaccard, und die phonetischen Ähnlichkeitsmaße, wie Soundex, sind die am häufigsten verwendeten. Seltener finden hybride Ähnlichkeitsmaße wie Monge-Elkan oder Soft TF-IDF Verwendung. Diese sind lediglich in einer Publikation verwendet worden. Einen modernen Ähnlichkeitsmaß-Ansatz stellen Word Embeddings dar, die neben Schreibfehlern auch semantische Ähnlichkeiten berücksichtigen sollen und meist zusammen mit Neuronalen Netzen eingesetzt werden. Die Auswertung der Kategorie der verwendeten Algorithmen im Prozessschritt Classification (siehe Tabelle 3.12) zeigt, dass die regelbasierten Ansätze auch in den jüngsten Publikationen am häufigsten verwendet werden. Am zweithäufigsten werden Supervised Learning Verfahren wie Support Vector Machines oder Decision Trees verwendet. Regelbasierte Ansätze sind nach wie vor populär, da kein zusätzlicher Aufwand für das Erstellen von Trainingsdaten anfällt. Lediglich in drei Publikationen werden Unsupervised Learning Algorithmen angewendet, die ebenfalls keine Trainingsdaten benötigen weshalb das Erstellen eines Regelwerkes entfällt. Tabelle 3.12 zeigt bereits eine Vielzahl von mehrfach verwendeten Algorithmen. 34 weitere Algorithmen werden einfach genannt, was die Vielzahl an Möglichkeiten für den Prozessschritt Classification unterstreicht. Die Auswertung der Kategorie RL-System (siehe Tabelle 3.13) zeigt, dass es sehr wenige RL-Systeme gibt, die in Forschungsarbeiten verwendet werden. Das am häufigsten verwendete FEBRL System wird in nur drei Publikationen verwendet und stammt aus dem Jahr 2004. Zuletzt erwähnt wurden die meisten RL-Systeme 2010 in der Publikation von Köpcke und Rahm (2010), die diese einem Benchmark unterzogen haben. Die jüngsten RL-Systeme sind Magellan und Deepmatcher, die an der Universität Madison-Wisconsin entwickelt worden sind. Die RL-Systeme basieren auf Python und sind Open-Source zugänglich, weshalb sie für die weitere Forschung von RL einen wertvollen Beitrag liefern können.

Die Ergebnisse der qualitativen Inhaltsanalyse zeigen, dass es für jeden RL-Prozessschritt eine Vielzahl von verschiedenen Algorithmen und Verfahren gibt, aus denen für jeden zu entwickelnden RL-Prozess ausgewählt werden muss. Für jede neue zu integrierende Datenquelle

und jedes neue zu integrierende Datenquellenpaar muss eine neue Auswahl von Algorithmen und Verfahren für jeden RL-Prozessschritt durchgeführt werden. Dies erfordert einen hohen manuellen Aufwand und viel Know-how, um geeignete RL-Prozesse zu implementieren. Die aktuelle Forschung fokussiert oftmals die Optimierung der RL-Ergebnisqualität auf den bestehenden Benchmark Datensets und verwendet dabei wenige Datenquellen, die verschiedene Realwelt-Entitäten repräsentieren. Keine Publikation fokussiert die Unterstützung bei der Auswahl von Algorithmen und Verfahren in den einzelnen RL-Prozessschritten, um den manuellen Aufwand und das erforderliche Know-how zu reduzieren. Ebenfalls fokussiert keine Publikation die Forschung eines Vorgehens hinzu einem generischen RL-Prozess der für neue zu integrierende Datenquellen oder Datenquellenpaare wiederverwendet werden kann und dadurch den manuellen Aufwand und das erforderliche Know-how reduziert.

Dieses Forschungslücke wird im Rahmen dieser Arbeit adressiert, um den manuellen Aufwand bei der Entwicklung von RL-Prozessen zu reduzieren. Viele Publikationen erwähnen in ihrem Ausblick, dass in der zukünftigen Forschung die Ansätze auf weitere Datensätze oder Domänen angewendet werden sollten. Darüber hinaus wurde bereits von Rahm (2016) ein Ansatz zur Entwicklung einer ganzheitlichen Datenintegration als zukünftiges Forschungsthema genannt. Es sollte mehr Forschung betrieben werden, um Datenquellen-unabhängige generische RL-Prozesse zu entwickeln, die eine Mindest-Ergebnisqualität liefern. Weitere entscheidende Aspekte für die Forschung hinzu generischen RL-Prozessen sind (1) die Reduktion des Aufwandes für das Erstellen von Trainingsdaten für jede neue Datenquelle oder jedes neue Datenquellenpaar und (2) die Beurteilung der Ergebnisqualität, ohne die gesamten Ergebnisse manuell überprüfen zu müssen.

3.4 Verwandte Arbeiten

Die Ergebnisse der qualitativen Inhaltsanalyse zeigen, dass in keiner der Publikationen versucht wird, einen generischen RL-Prozess zu entwickeln, durch den der manuelle Aufwand reduziert werden könnte. Durch die qualitative Inhaltsanalyse wurden die relevanten Publikationen, Forschenden, Institutionen und Projekte identifiziert, deren weitere Forschungsarbeiten während des Fortgangs dieser Arbeit weiter beobachtet wurden. Die identifizierten weiteren Forschungsarbeiten werden im Folgenden vorgestellt und eine Abgrenzung zu dieser Arbeit getroffen.

3.4.1 Magellan - Record Linkage-Ecosystem

Das Magellan Projekt der Universität Wisconsin hat das Ziel ein RL-Ecosystem zu entwickeln und ist damit relevant für diese Forschungsarbeit. Das Projekt wurde im Jahr 2015 gestartet.

Dabei kollaboriert die Universität Wisconsin mit verschiedenen Industriepartnern, um das Ziel, ein RL-Ecosystem zu entwickeln, zu erreichen (vgl. Doan et al., 2020; Govind et al., 2019). Im Magellan Projekt sind in den letzten Jahre zehn Publikationen entstanden, die die Forschungsergebnisse des Projektes beschreiben und für diese Arbeit relevant sind (siehe Tab. 3.14).

Tabelle 3.14: Publikationen aus dem Magellan Projekt

Titel	Jahr	Autor
Magellan: Toward Building Entity Matching Management Systems	2016	Konda et al.
Human-in-the-Loop Challenges for Entity Matching	2017	Doan et al.
Magellan: Toward Building Entity Matching Management Systems	2018	Konda et al.
Cloudmatcher: a hands-off cloud/crowd service for entity matching	2018	Govind et al.
Deep Learning for Entity Matching	2018	Mudgal et al.
Toward a System Building Agenda for Data Integration (and Data Science)	2018	Doan et al.
Executing Entity Matching End to End: A Case Study	2019	Konda et al.
Entity Matching Meets Data Science: A Progress Report from the Magellan Project	2019	Govind et al.
Magellan: Toward Building Ecosystems of Entity Matching Solutions	2020	Doan et al.
Deep Entity Matching with Pre-Trained Language Models	2020	Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan und Wang-Chiew Tan

Die Publikationen von Govind et al. (2019) und Doan et al. (2020) beschreiben den aktuellen Stand des Forschungsprojektes und die bisher erzielten Ergebnisse basierend auf allen zuvor veröffentlichten Publikationen. Daher werden diese beiden Publikationen genutzt, um das Magellan Projekt zusammenfassend zu beschreiben.

Um das RL-Ecosystem zu entwickeln, wurden im Projekt die folgenden Ziele definiert: Zunächst sollen häufig vorkommende RL-Prozesse identifiziert werden. Weiterhin sollen How-to-Guides für die häufig vorkommenden RL-Prozesse entwickelt werden. Durch die How-to-Guides sollen die Schwachstellen innerhalb der RL-Prozesses identifiziert werden, für die wiederum Tools zur Automatisierung entwickelt werden sollen. Bei der Entwicklung der Tools soll ML eingesetzt werden, wenn geeignete Einsatzszenarien identifiziert werden (vgl. Govind et al., 2019).

Im Magellan Projekt werden die RL-Prozesse prozessual und algorithmisch definiert. In Abbildung 3.3 sind die Magellan RL-Prozesse A und B abgebildet und in den Standard RL-Prozess aus Abbildung 2.4 eingeordnet. Prozess A und B bestehen aus dem Prozessschritt Blocker, der dem Blocking entspricht und dem Prozessschritt Matcher, der den Prozessschrit-

ten Comparison und Classification entspricht. Im RL-Prozess A wendet der Benutzer einen Blocking-Algorithmus an, um eine Menge von Tupeln zu erhalten, und wendet dann einen Klassifikations-Algorithmus an, um die Tupel in Match oder No-Match einzuteilen. Im RL-Prozess B wird berücksichtigt, dass der Anwender mehrere Blocking-Algorithmen im Prozessschritt Blocking verwenden kann. In beiden RL-Prozessen werden für den Prozessschritt Classification nur überwachte Lernverfahren zur Verfügung gestellt. Der Einsatz von regelbasierten Verfahren für den Prozessschritt Classification wird erwähnt, aber es wird nicht weiter darauf eingegangen (vgl. Govind et al., 2019).

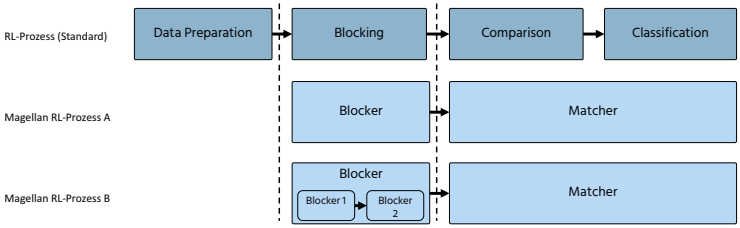


Abbildung 3.3: Record Linkage-Prozesse in Magellan

Für die Durchführung der zwei RL-Prozesse wurde das Magellan RL-Ecosystem entwickelt, welches in den Publikationen des Projektes auch als PyMatcher bezeichnet wird. Das Magellan RL-Ecosystem setzt sich aus den folgenden Python Tools zusammen (vgl. Govind et al., 2019):

py_entitymatching [15] Mit diesem Open-Source Python Tool können Datenquellen mithilfe von supervised Learning Algorithmen integriert werden. Dieses Tool bietet Algorithmen für die Prozessschritte Blocking und Classification. Darüber hinaus bietet es auch Algorithmen und Verfahren für die Data Exploration, das Data Profiling, das Debugging, das Sampling oder auch das Labeling an.

py_stringmatching [16] Dieses Open-Source Python Tool stellt String Tokenizer, wie bspw. alphabetische Tokenizer und Leerzeichen Tokenizer, und String Similarity Measures wie bspw. Levenshtein, Jaccard und TF/IDF zur Verfügung. Das Tool ist Bestandteil des py_entitymatching Tools.

py_stringsimjoin [17] Dieses Open-Source Python Tool stellt Implementierungen von String Similarity Joins zur Verfügung, die auf den String Similarity Measures wie Jaccard, Cosi-

[15] `https://github.com/anhaidgroup/py_entitymatching`
[16] `https://github.com/anhaidgroup/py_stringmatching`
[17] `https://github.com/anhaidgroup/py_stringsimjoin`

nus, Overlap oder Levenshtein basieren. Das Tool ist ein Bestandteil vom py_entitymatching Tools.

DeepMatcher [18] Dieses Open-Source Python Tool stellt Funktionen bereit, um mit neuronalen Netzen strukturierte Datensätze oder unstrukturierte Texte in Match und No-Match klassifizieren zu können. Es umfasst Funktionen zum Training und Anwenden von neuronalen Netzen für das RL. DeepMatcher kann ergänzend zum py_entitymachting Tool eingesetzt werden, indem es für den RL-Prozessschritt Classification eingesetzt wird.

Konda et al. (2016b) grenzen das Magellan RL-Ecosystem durch seine Systemarchitektur von 18 nicht kommerziellen und 15 kommerziellen existierenden RL-Systemen ab (vgl. Konda et al., 2016b). Die existierenden Systeme sind als monolithische System konzeptioniert und implementiert worden (vgl. Konda et al., 2016b). Im Gegensatz dazu orientiert sich die Systemarchitektur des Magellan RL-Ecosystem am Data Science Ecosystem Pydata. Der Einfluss der Konzepte und Verfahren aus der Data Science auf das RL liegt darin begründet, dass RL im Magellan Projekt als eine Teilaufgabe der Data Science gesehen wird. Daher ist die Magellan Systemarchitektur ein Ecosystem aus vielen interoperablen Tools, die entlang des RL-Prozesses unterstützen sollen. Für die Entwicklung der Tools innerhalb des Ecosystems wurden die folgenden Grundsätze definiert (vgl. Govind et al., 2019; Doan et al., 2020):

1. Sie sollten untereinander und mit bestehenden Python Tools interoperabel sein.
2. Sie sollten atomar sein, d.h. jedes Tool erledigt nur eine Aufgabe.
3. Sie sollten unabhängig von anderen Tools sein, d.h. die Tools können alleinstehend verwendet werden.
4. Sie sollten anpassbar sein.
5. Sie sollten sowohl für Menschen als auch für Maschinen effizient sein.

Zusätzlich sind im Projekt How-to-Guides für die Anwendung der Tools des Ecosystems entstanden (vgl. Govind et al., 2019; Doan, 2017). Die How-to-Guides (vgl. Doan, 2017) werden durch Quellcodebeispiele[19] ergänzt, die Anwendungsbeispiele für die einzelnen Tools liefern. Die How-to-Guides unterscheiden die Entwicklungs- und Produktionsphase des RL-Prozesses. Die Entwicklungsphase repräsentiert den iterativen Entwicklungsprozess des RL-Prozesses bis die zuvor definierten Anforderungen wie bspw. die definierte Qualität der Ergebnisse erfüllt sind. In der Produktionsphase wird der in der Entwicklungsphase implementierte RL-Prozess produktiv geschaltet und kontinuierlich auf dem gesamten Datenbestand ausgeführt. Die

[18] `https://github.com/anhaidgroup/deepmatcher`
[19] `http://anhaidgroup.github.io/py_entitymatching/v0.3.x/user_manual/guides.html`

How-to-Guides beschreiben und unterstützen ausschließlich die in Abb. 3.3 beschriebenen RL-Prozesse mit dem Einsatz von supervised Learning Algorithmen für den Prozessschritt Classification. Zudem wird nur die Entwicklungsphase des RL-Prozesses beschrieben. Die Produktionsphase wird nicht beschrieben (vgl. Doan, 2017).

Das Magellan RL-Ecosystem, bestehend aus den Python Tools und den How-to-Guides, wurde in acht verschiedenen Realwelt RL-Problemstellungen der Industrie und Wissenschaft eingesetzt (siehe Tab. 3.15). In der ersten Spalte sind die vier Unternehmen bzw. die vier Fachbereiche der Universität Wisoncsin aufgeführt, die die RL-Probleme für die Fallstudien stellen. In der Spalte RL-Problem sind die Realwelt-Entitäten aufgeführt, die in den Fallstudien integriert werden. Alle acht Fallstudien haben unterschiedliche RL-Entitäten. In Spalte Team ist aufgeführt, wer an der Entwicklung der Lösung beteiligt gewesen ist (vgl. Govind et al., 2019).

Tabelle 3.15: Realwelt Anwendungen von Magellan
UW = University Wisconsin

Unternehmen/Fachbereich	RL-Problem	Team
Walmart	Produkt Matching	1 Student, 1 Mitarbeiter
Recruit Holdings	Matching von Geschäften, Firmen, und Immobilien	Mehrere Mitarbeiter
Johnson Controls	Zulieferer Matching	1 Student
Marshfield Clinic	Medikamenten Matching	1 Student, 1 Mitarbeiter
Economics (UW)	Finanzhilfen Matching	2 Student
Land Use (UW)	Rinderfarmen Matching	1 Student, 1 Programmierer, 2 Mitarbeiter
Biomedicine (UW)	Ontologie-Term Matching	1 Student
Limnology (UW)	Tabellenattribute Matching	2 Student

Durch das Anwenden des Magellan RL-Ecosystems in Fallstudien wurden viele Erkenntnisse gewonnen. Generell wurde festgestellt, dass RL-Projekte in der Realität sehr chaotisch sind, da die Entwickler der RL-Prozesse oftmals alle möglichen Verfahren on-the-fly ausprobieren. In einem RL-Projekt sind Entwickler und Domänenexperten involviert, die in einem regelmäßigen Austausch stehen sollten. Im Folgenden werden die im Magellan Projekt generierten Erkenntnisse beschrieben (vgl. Govind et al., 2019):

RL-Ecosystem: Der RL-Ecosystem Ansatz ist geeignet, da verschiedene RL-Tools entwickelt worden sind, mit denen eine breite Anzahl von RL-Problemen gelöst wurde. Durch die vielen interoperablen Tools können die Anwender schnell neue RL-Prozesse implementieren und diese schnell verändern. Zudem ist es im Gegensatz zu monolithischen Systemen einfacher Tools zu entwickeln, anzupassen und diese bereitzustellen. Dennoch wurde im Magellan Projekt festgestellt, dass es schwierig ist, alle fünf Grundsätze zur Entwicklung

der einzelnen RL-Tools einzuhalten. Für jedes neue Szenario können einzelne ausgewählte Tools genutzt und für dieses spezielle Szenario Mini-How-to-Guides erstellt werden (vgl. Govind et al., 2019).

How-to-Guides: Durch die Entwicklung eines RL-Ecosystems und das Anwenden in Realwelt Fallstudien haben die Forscher festgestellt, dass die How-to-Guides eine sehr wichtige Rolle einnehmen, um die Entwickler und Domänenexperten beim Lösen eines RL-Problems zu unterstützen. Die vollständige Automatisierung des RL-Prozesses ist sehr schwierig (vgl. Doan et al., 2020). Zudem wurde bei der Anwendung der How-to-Guides festgestellt, dass diese in ihrer aktuellen Form nicht geeignet sind, um Anwender bei der Entwicklung eines RL-Prozesses zu unterstützen. Dies liegt darin begründet, dass die betrachteten RL-Probleme heterogener als gedacht sind. Dies führte dazu, dass die Anwender in einen „trial and error“ Prozess übergangegen sind, bis sie die beste RL-Lösung identifiziert haben. In den durchgeführten RL-Fallstudien wurde festgestellt, dass die Kommunikation zwischen RL-Entwicklern und den Domänenexperten entscheidend ist. Für eine zielgerichtete und effektive Kommunikation zwischen RL-Entwicklern und Domänenexperten ist ein entsprechender How-to-Guide wichtig. Govind et al. (2019) schlagen vor, dass zukünftig Mini-How-to-Guides gemeinsam mit und für spezielle RL-Prozesse entwickelt werden sollten (vgl. Govind et al., 2019).

Einsatz von ML im RL: ML spielt eine Schlüsselrolle für RL-Tools. Allerdings müssen - entgegen der Erwartungen der Forscher im Magellan Projekt - signifikante Herausforderungen gemeistert werden, um ML effektiv im RL einsetzen zu können. ML sollte mit manuell erstellten Regeln, einer effektiven Nutzerinteraktion und Big Data Skalierung eingesetzt werden, um das volle Potenzial für den RL-Prozess auszuschöpfen. Der Einsatz von supervised Learning Algorithmen im Magellan RL-Ecosystem erfordert einen hohen manuellen Aufwand für die Erstellung eines Trainingsdatensatzes, der aus Tupeln mit dem Label Match und No-Match besteht. Das manuelle Erstellen des Trainingsdatensatzes ist für jedes neue zu integrierende Datenquellenpaar erforderlich, da das Übertragen der entwickelten Lösungen auf neue Datenquellen schwierig ist. Es wurde festgestellt, dass eine Kombination aus ML und Regeln die besten RL-Prozesse ergeben. (vgl. Govind et al., 2019).

RL-Know-how, Self-service RL und manueller Aufwand: Weiterhin wurde festgestellt, dass die Anwender viel RL und ML Know-how benötigen, um die Tools des Magellan RL-Ecosystems anzuwenden. Um Anwendern ohne Programmierkenntnisse zu ermöglichen, zwei Datenquellen zu integrieren, wurde im Magellan Projekt das Self-Service RL-System namens Cloudmatcher entwickelt (vgl. Govind et al., 2019). Cloudmatcher erfordert ebenfalls das Erstellen von Trainingsdaten, um einen Klassifikator für den Prozessschritt Classification zu trainieren. Cloudmatcher stellt die im Magellan Projekt entwickelten Tools für

die verschiedenen RL-Prozesse in einer cloudbasierten Webanwendung zur Verfügung. Um diese Webanwendung zu nutzen, sind keine Programmierkenntnisse erforderlich (vgl. Govind et al., 2018). Allerdings müssen die Anwender die Trainingsdatensätze erstellen und zudem die möglichen Datenprobleme für jede Datenquelle identifizieren, verstehen und geeignete Algorithmen und Verfahren aus den Tools auswählen, um diese zu lösen. Die Forscher stellten in der Anwendung von Cloudmatcher in Fallstudien fest, dass die Data Preparation ein kritischer RL-Prozessschritt ist. Diesen Prozessschritt in einem Self-Service RL-System umzusetzen ist schwierig und aktuell nicht gelöst (vgl. Govind et al., 2018). Zudem entwickelt sich Cloudmatcher mehr zu einem monolithischen RL-System, was den Grundsätzen des Magellan Projektes widerspricht. Um diese Entwicklung zu verhindern, soll die Systemarchitektur von Cloudmatcher geändert werden, sodass Cloudmatcher in viele interoperable Microservices überführt wird (vgl. Govind et al., 2019).

3.4.2 JedAI - Record Linkage-System

Ein weiteres relevantes Forschungsprojekt, mit dem Ziel ein RL-System zu entwickeln, befindet sich an der griechischen National und Kapodistrian Universität in Athen. Dort wird das RL-System „Java gEneric DAta Integration Toolkit “ kurz JedAI entwickelt (vgl. Papadakis, Mandilaras et al., 2020). Rund um JedAI sind neun Publikationen entstanden, die in Tabelle 3.16 aufgeführt sind. Die Publikationen Papadakis, Mandilaras et al. (2020) und Papadakis et al. (2021) geben einen umfassenden Überblick der gesamten Forschung und stellen das RL-System JedAI vor, weshalb diese beiden Publikationen im Folgenden als Hauptquellen dienen.

Tabelle 3.16: Publikationen aus dem JedAI Projekt

Titel	Jahr	Autor
Schema-agnostic vs schema-based configurations for blocking methods on homogeneous data	2015	Papadakis, Alexiou, Papastefanatos und Koutrika
Schema-agnostic Progressive Entity Resolution	2018	Simonini et al.
The return of JedAI: End-to-End Entity Resolution for Structured and Semi-Structured Data	2018	Papadakis et al.
Entity Resolution: Past, Present and Yet-to-Come	2020	Papadakis, Tsekouras et al.
Three-dimensional Entity Resolution with JedAI	2020	Papadakis, Ioannou und Palpanas
Blocking and Filtering Techniques for Entity Resolution: A Survey	2020	Papadakis, Mandilaras et al.
Domain- and Structure-Agnostic End-to-End Entity Resolution with JedAI	2020	Papadakis, Skoutas et al.
An Overview of End-to-End Entity Resolution for Big Data	2021	Christophides et al.
The Four Generations of Entity Resolution	2021	Papadakis et al.

JedAI ist ein monolithisches Open-Source RL-System. JedAI soll domänenunabhängig sein, da es nicht auf Hintergrundwissen von Experten angewiesen sein soll und mit geringem manuellen Aufwand mit Daten jeder Domäne angewendet werden soll. Mit JedAI können strukturierte, semi-strukturierte und unstrukturierten Datenquellen verarbeitet werden. JedAI besteht aus zwei Komponenten: (1) JedAI-core, eine Bibliothek mit zahlreichen State-of-the-Art RL-Algorithmen und Verfahren[20] für die einzelnen RL-Prozessschritte, die zu RL-Prozessen zusammengefügt werden können und ein einfaches Benchmarking ihrer relativen Performance ermöglicht. (2) JedAI-gui, eine Desktop-Anwendung, die die Zusammenstellung der RL-Prozesse über eine assistentenähnliche Oberfläche erleichtert. Diese soll sowohl für Domänenexperten als auch für Entwickler geeignet sein und bietet neben einer Dokumentation für jede Funktionalität (eine Art How-to-Guide) auch eine Default-Konfiguration der Algorithmen. Die Classification der Tupel in Matches und No-Matches basiert auf Clustering Algorithmen, die keine gelabelten Trainingsdaten benötigen (vgl. Papadakis, Skoutas et al., 2020).

Mit JedAI können RL-Prozesse implementiert werden, die über die folgenden drei Dimensionen beschrieben werden (vgl. Papadakis, Ioannou & Palpanas, 2020):

Schemafähigkeit: JedAI unterstützt RL-Prozesse die schemabasierte und schemalose Datenquellen integrieren sollen. Dies grenzt JedAI von den meisten RL-Systemen wie bspw. Magellan ab, die lediglich schemabasierte Datenquellen verarbeiten können.

Budget-Bewusstsein: JedAI unterscheidet in budgetunabhängige und budgetabhängige RL-Prozesse. Budgetunabhängige RL-Prozesse werden im Batch-Prozess ausgeführt und haben keine Restriktionen bezüglich Zeit- und Rechenkapazität. Budgetabhängige RL-Prozesse arbeiten nach dem Pay-as-you-go Prinzip, da nach und nach Ergebnisse produziert werden, bis die zur Verfügung stehende Zeit- oder Rechenkapazität erreicht ist.

Ausführungsmodus: JedAI unterscheidet und unterstützt die serielle und die parallele Ausführung der RL-Prozesse. Die parallele Ausführung wird durch Apache Spark realisiert.

Auf Basis der Dimensionen (1) Schemafähigkeit, (2) Budget-Bewusstsein und (3) Ausführungsmodus werden in JedAI vier RL-Prozesse definiert. Diese sind ebenso wie im Magellan RL-Ecosystem über die prozessuale und algorithmische Perspektive entstanden und werden im Folgenden beschrieben (vgl. Papadakis, Ioannou & Palpanas, 2020):

1. **Budgetunabhängiger und schemaunabhängiger RL-Prozess:** Der budgetunabhängige und schemaunabhängige RL-Prozess (siehe Abb. 3.4) beginnt mit dem Prozessschritt

[20] `https://github.com/scify/JedAIToolkit`

Schema Clustering, der optional ausgeführt werden kann. Optionale Prozessschritte sind in Abbildung 3.4 grau. Der Prozessschritt unterstützt bei der Identifizierungen von Attributen, die dieselben Inhalte repräsentieren, was dem Big Data Integration Prozessschritt Schema Matching gleichzusetzen ist. Das Schema Clustering von JedAI kann über den Attributnamen, dem Inhalt der Attribute oder aus einer hybriden Variante auf Basis des Attributnamens und der Inhalte der Attribute durchgeführt werden. Der nächste Prozessschritt Block Building entspricht dem Standard RL-Prozessschritt Blocking, in dem die Anzahl der zu vergleichenden Tupel reduziert wird. JedAI stellt Hash-basierte Algorithmen wie bspw. Token Blocking oder Q-Gram Blocking und Ähnlichkeits-basierte Algorithmen wie bspw. Sorted Neighbourhood oder LSH MinHash zur Verfügung. Der nächste Prozessschritt Block Cleaning ist optional und stellt Algorithmen zur Verfügung, mit denen redundante und überflüssige Tupel entfernt werden können, um die Anzahl an Vergleichen zu reduzieren. Der darauf folgende Prozessschritt Comparison Cleaning ist ebenfalls optional. Auch dieser Prozessschritt hat das Ziel, redundante und überflüssige Tupel zu entfernen, die nicht verglichen werden müssen. Im Gegensatz zum Block Cleaning werden für das Comparison Cleaning Algorithmen zur Verfügung gestellt, die auf Tupel-Ebene die Notwendigkeit des Verlgeichs prüfen. Damit erreichen diese Algorithmen eine höhere Genauigkeit, die auf Kosten der Ausführungsdauer geht. Der Prozessschritt Entity Matching entspricht dem Standard RL-Prozessschritt Comparison. JedAI bietet Ähnlichkeitsmetriken und ein graph-basiertes Verfahren, um die Ähnlichkeit der Tupel berechnen zu können. Nachdem die Ähnlichkeiten berechnet worden sind folgt der letzte Prozessschritt Entity Clustering, der den Standard RL-Prozessschritt Classification darstellt. JedAI setzt auf Clustering Algorithmen für die Klassifikation der Tupel in Match und No-Match, sodass keine Trainingsdaten erstellt werden müssen. Für die Clean-Clean-ER-Probleme stellt JedAI bspw. das Unique Mapping Clustering zur Verfügung und für die Dirty-ER-Probleme bspw. das Connected Components Verfahren.

2. **Budgetunabhängig und schemabasierter RL-Prozess:** Der budgetunabhängig und schemabasierte RL-Prozess ist in Abbildung 3.4 abgebildet. Er besteht aus den Prozessschritten Similarity Join, der dem Standard RL-Prozessschritt Comparison entspricht und dem Prozessschritt Entity Clustering. JedAI stellt Similarity Join Algorithmen wie AllPairs oder FastSS zur Verfügung, um die Ähnlichkeit der Tupel über die zuvor definierten Attribute und den festgelegten Schwellwerte zu berechnen. Im Prozessschritt Entity Clustering werden dann auf Basis der zuvor berechneten Ähnlichkeiten die Matches und No-Matches bestimmt. Dieser RL-Prozess soll sich eignen, wenn Domänenwissen über das Datenset vorhanden ist und das Datenset Attribute enthält, die unterscheidbar genug sind, um Match und No-Match zu klassifizieren.

3. **Budgetabhängiger und schemaunabhängiger RL-Prozess:** Der Budgetabhängige

und schemaunabhängige RL-Prozess (siehe Abb. 3.4) eignet sich, wenn die Zeit- oder Rechenkapazität begrenzt ist. Der Prozess ähnelt dem budgetunabhängigen und schemaunabhängigen RL-Prozess (siehe Abb. 3.4), unterscheidet sich allerdings in einigen Aspekten. Dem RL-Prozess werden die Parameter maximale Rechenzeit oder maximale Anzahl an zu vergleichenden Tupeln übergeben. Das Block Building wird zu einem optionalen Prozessschritt. Im Prozessschritt Entity Matching wird ein Vergleich zur selben Zeit ausgeführt. Der Prozessschritt Comparison Prioritization ist neu hinzugekommen. Dieser Prozessschritt stellt Algorithmen zur Verfügung um einen Verarbeitungsablauf der Datensätze und Tupel zu erstellen. Dieser Verarbeitungsablauf wird abgearbeitet, bis das Budget aus Zeit- oder Rechenkapazität ausgeschöpft ist.

4. **Budgetabhängiger und schemabasierter RL-Prozess:** Der budgetabhängige und schemabasierte RL-Prozess besteht aus denselben Prozessschritten wie der budgetunabhängige RL-Prozess in Abbildung 3.4. Der einzige Unterschied besteht darin, dass ein Top-k Similarity Join im Prozessschritt Similarity Join angewendet wird, um die Reihenfolge der Vergleiche bis zum Limit der Zeit- und Rechenkapazität zu bestimmen.

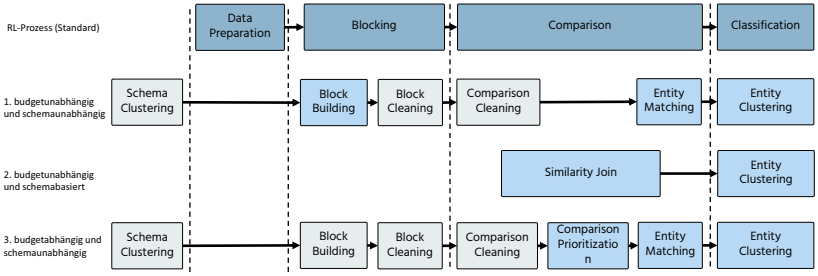


Abbildung 3.4: Record Linkage-Prozesse in JedAI

Papadakis, Ioannou und Palpanas (2020) sagen über JedAI, dass es vier wichtige Herausforderungen beim Aufbau von RL-Systemen berücksichtigen würde: (1) es erfordere keine Programmierkenntnisse der Anwender, (2) es biete integrierte How-to-Guides für die Erstellung von RL-Prozessen, (3) es decke den gesamten RL-Prozess ab und (4) es stelle eine breite Auswahl von Algorithmen und Verfahren zur Verfügung. Der Vergleich von Magellan und JedAI im Paper von Papadakis, Ioannou und Palpanas (2020) zeigte, dass beide Systeme auf unterschiedlichen Datensets eine ähnliche Ergebnisqualität erreichen. Hinsichtlich der Zeiteffizienz bei der Durchführung des RL-Prozesses ist JedAI schneller als Magellan, da es keine zusätzlich Zeit für das Training der Klassifikationsmodelle benötigt (vgl. Papadakis, Ioannou & Palpanas, 2020).

3.4.3 Weitere Arbeiten

Neben den relevanten RL-Systemen Magellan und JedAI existieren noch weitere relevante RL-Forschungsbereiche für diese Arbeit. Die Publikationen zu den relevanten Forschungsbereichen wurden durch die kontinuierliche Vorwärtssuche anhand der Publikationen aus dem Magellan und JedAI Projekt identifiziert. Die Vorwärtssuche wurde durch das Tool Connected Papers[21] unterstützt. Einige Publikationen fokussieren eine Realwelt-Entität wie das Produkt oder das Unternehmen und einige Publikationen fokussieren den RL-Prozessschritt Data Preparation. Diese Forschungsbereiche werden im Folgenden dargestellt.

Realwelt-Entität Produkt: Einen wesentlich Beitrag zur RL-Forschung bieten die Publikationen von Köpcke (vgl. Köpcke & Rahm, 2008 Köpcke & Rahm, 2010; Köpcke et al., 2010; Köpcke et al., 2012). Die Publikationen werden in der Dissertation (vgl. Köpcke, 2014) zusammengefasst. In der Forschung wird das Integrieren von Datenquellen mit der Realwelt-Entität Produkt fokussiert und als ein Spezialfall des RL betrachtet. Köpcke (2014) liefert einen maßgeschneiderten Gesamtansatz für das RL der Realwelt-Entität Produkt. Der Ansatz unterstützt einen RL-Prozess der aus den Prozessschritten Data Preparation und Classification mit supervised Learning Algorithmen besteht. In der Data Preparation werden neue Attribute extrahiert und bereinigt, die für die Klassifikation genutzt werden können. Insbesondere extrahiert und verwendet der Ansatz sogenannte Produktcodes, um Produkte zu identifizieren und von ähnlichen Produktvarianten zu unterscheiden. Nach der Data Preparation wird mit supervised Learning Algorithmen die Klassifikation der Tupel in Match und No-Match durchgeführt.

Zecchini, Simonini und Bergamaschi (2020) beschreiben einen implementierten RL-Prozess für die Realwelt-Entität Produkt, am Beispiel von Kameras, ohne den Einsatz von Machine Learning (ML). Auf der Suche nach einer guten Lösung für das RL-Problem der Kameradaten haben die Autoren die aktuellen RL-Methoden des maschinellen Lernens (Magellan) und des Deep Learning (DeepMatcher) untersucht. Zecchini et al. (2020) stellen fest, dass in vielen realen Szenarien das Matching auf kleinen Variationen basiert, was eine Generalisierung der trainierten ML-Modelle auf den gesamten Datensatz erschwert. In den verwendeten Kameradaten existieren markenabhängige Variationen, um eine Kamera (Realwelt-Entität Produkt) von einer anderen zu unterscheiden. Es wurden manuell die benötigten Regeln und Listen (Verwaltung von Präfixen und Suffixen, Ausnahmen usw.) entworfen, die durch ML-Methoden nicht synthetisiert werden können. Die Autoren weisen darauf hin, dass in Ihrem Fall das Erstellen von Regeln und Listen nicht aufwändiger ist, als das Erstellen eines gelabelten Datensatzes für den Einsatz von ML-Methoden.

[21] `https://www.connectedpapers.com/`

Realwelt-Entität Unternehmen: Die Publikationen von Schild und Schultz (2017), Cuffe und Goldschlag (2018) und Gschwind, Miksovic, Minder, Mirylenka und Scotton (2019) wurden als Forschungsarbeiten identifiziert, die sich speziell mit dem RL der Realwelt-Entität Unternehmen beschäftigen.

Schild und Schultz (2017) stellen in ihrer Arbeit einen selbstentwickelten RL-Prozess vor, der speziell für sieben Datenquellen, die die Realwelt-Entität Unternehmen beinhalten, entwickelt worden ist. Zwei der Datenquellen stammen von den Datenprovidern Bureau van Dijk und Bisnode Hoppenstedt. Fünf Datenquellen sind interne Datenquellen der Deutschen Bundesbank. Im entwickelten RL-Prozess werden zwölf Attribute verwendet, wie bspw. der Unternehmensname, die Rechtsform, die Postleitzahl, der Ort, die Straße und der Unternehmensumsatz. Schild und Schultz (2017) beschreiben den Unternehmensnamen als wichtigstes Attribut zur Unterscheidung von Unternehmen. Die Unterscheidbarkeit des Unternehmensnamens kann durch geographische Zusätze oder die Rechtsform im Unternehmensnamen erhöht werden. Für Schild und Schultz (2017) ist das wichtigste Attribut zum Vergleich von Unternehmen deren Rechtsform. Um die Rechtsform zu nutzen, haben sie reguläre Ausdrücke zur Klassifizierung der Rechtsform entwickelt. Mit den regulären Ausdrücken können nur deutsche Rechtsformen klassifiziert werden. Das Blocking wurde spezifisch für die Datenquellen entwickelt. Das Blocking besteht aus acht Filtern die auf den Attributen Unternehmensname, Postleitzahl, gemeinsamen IDs und Telefonnummer basieren. Für die Ähnlichkeitsberechnung der Attribute wurde die Levenshtein-Distanz und eine Namenstoken basierte Soft-IDF Metrik verwendet. Für den Prozessschritt Classification wurde ein dreistufiges Verfahren entwickelt. Zuerst werden alle Tupel, die eine gemeinsame externe ID besitzen, als Match klassifiziert. Im zweiten Schritt werden alle Tupel als Match klassifiziert, die eine exakte Übereinstimmung von Name und Adresse besitzen. Im dritten Schritt wird auf den übrig gebliebenen Tupeln das trainierte ML-Verfahren angewandt, um einen Match Score zu berechnen.

Cuffe und Goldschlag (2018) gehen auf die Problematik ein, dass es viele einzelne Algorithmen und Methoden für die einzelnen RL-Prozessschritte gibt. Es wird ein RL-Prozess entwickelt, der sich für Census Business Microdata Datenquellen eignet. Cuffe und Goldschlag (2018) nennen ihren entwickelten RL-Prozess „Multiple Algorithm Matching for Better Analytics“ kurz MAMBA. Der MAMBA RL-Prozess beginnt mit der Data Preparation. In diesem Prozessschritt werden die Daten durch das Ersetzen von Abkürzungen und Suffixen standardisiert. Hierfür wurden 500 Korrekturen, wie bspw. Corporation zu Corp zu vereinheitlichen, definiert und implementiert. Der Prozessschritt Blocking basiert auf den Adressangaben wie bspw. der Postleitzahl, der Stadt oder dem Staat. Für den Prozessschritt Comparison werden die String Similarity Measures Levenshtein, Jaro oder Jaro-Winkler auf den vorhandenen Unternehmensnamens und Adressattributen berechnet. Auf Basis dieser Tupel wird ein Entscheidungsbaum für den RL-Prozessschritt Classifica-

tion trainiert, der die besten Ergebnisse für die Klassifikation der Tupel in Match und No-Match erzielte.

Gschwind et al. (2019) konzentrieren sich auf das RL von Datenquellen, die die Realwelt-Entität Unternehmen beinhalten. Dabei werden die Attribute Unternehmensname, Standort und Branche verwendet. Das Ergebnis der Arbeit ist ein für die in der Publikation vorhandenen Datenquellen spezifischer RL-Prozess. Für den Prozessschritt Blocking wird der LSH MinHash Algorithmus verwendet. Für den RL-Prozessschritt Classification wird ein regelbasierter Ansatz verwendet. Der regelbasierte Ansatz besteht aus manuell definierten Schwellwerten für die berechneten Ähnlichkeiten auf den Attributen Unternehmensname, den Adressdaten wie Straße, Postleitzahl, Stadt und Ländercode und der Branche. Für den Unternehmensnamen werden die String Similarity Measures Jaccard, Levenshtein und die selbst entwickelte RLS Scoring-Funtkion verwendet. Die Postleitzahl wird anhand der übereinstimmenden Zeichen verglichen. Die Stadt wird über die Haversine-Distanz verglichen, sofern GPS Koordinaten vorliegen, die aus der GeoNames[22] Datenquelle stammen. Wenn keine GPS Koordinaten vorliegen, wird die Stadt über die Levenshtein-Distanz verglichen. Der Ländercode wird auf Gleichheit geprüft. Die Branche wird wie die Postleitzahl über die übereinstimmenden Zeichen verglichen, da die Branche kodiert als Branchencode vorliegt. Der Fokus des Papers liegt auf dem Prozessschritt Data Preparation. Für diesen Prozessschritt wurde ein ML-Verfahren entwickelt, das aus dem Unternehmensnamen eine Unternehmenskurzbezeichnung generiert. Zum Beispiel extrahiert das ML-Verfahren die Unternehmenskurzbezeichnung „Aston Martin“aus dem Unternehmensnamen „Aston Martin Lagonda Limited“. Die Autoren definieren dieses Problem als ein Sequenz-Labeling-Problem und trainieren einen Conditional Random Field Algorithmus zur Identifikation und Extraktion des Firmenkurznamens. Gschwind et al. (2019) gehen in ihrer ML Methode zur Extraktion des Firmenkurznamens nicht auf die Rechtsform des Unternehmens ein. Die Rechtsform des Unternehmens kommt in ihrem Fall nur selten im Unternehmensnamen vor. Ihr erster Ansatz für den RL-Prozess löschte die Rechtsform wie „inc.“ oder „ltd.“. Sie stellten jedoch fest, dass sich einige Unternehmen nur durch ihre Rechtsform unterscheiden. Als Lösung haben sie der Rechtsform in ihrem Scoringprozess zur Berechnung der Ähnlichkeit der Tupel weniger Gewicht gegeben. In der Publikation wird nicht beschrieben, wie die Rechtsform identifiziert wird, um ihr ein geringeres Gewicht zu geben. Ebenso wird die Rechtsform nicht extrahiert, um sie als zusätzliches Attribut im RL-Prozess für den Ähnlichkeitsbestimmung der Tupel zu verwenden. Eine wesentlich Erkenntnis die Gschwind et al. (2019) schildern ist die Limitation des Einsatzes von ML im RL, die darin besteht, ausreichend Trainingsdaten für die ML Verfahren zu erzeugen. Für jede neue Datenquelle, die integriert werden soll, müssen neue Trainingsdaten erstellt werden. Für den Prozessschritt Classification sehen Gschwind et al. (2019) den Einsatz

[22] `https://www.geonames.org/`

von ML als sehr limitiert an. Allerdings sehen sie das Potenzial von ML in Teilaufgaben des RL-Prozesses, wie bspw. dem Identifizieren und Extrahieren des Unternehmenskurznamens im Prozessschritt Data Preparation.

RL-Prozesschritt Data Preparation: Lediglich die Publikationen von Randall, Ferrante, Boyd und Semmens (2013) und l. Koumarelas, Papenbrock und Naumann (2020) fokussieren den RL-Prozessschritt Data Preparation.

Randall et al. (2013) untersuchen den Einfluss der Data Preparation auf die RL-Ergebnisqualität. Basierend auf einem Review von RL-Software identifizierten sie eine Reihe von verschiedenen Data Preparation Verfahren. Sie wendeten diese auf einen synthetischen Datensatz und einen realen Datensatz aus der Verwaltung an, um den Einfluss der Data Preparation auf die RL-Ergebnisqualität zu messen. Die Ergebnisse zeigen, dass die Data Preparation in ihren Experimenten wenig Einfluss auf die RL-Ergebnisqualität hat. Die Arbeit berücksichtigt nicht die Realwelt-Entität Unternehmen und die verwendeten Data Preparation Verfahren sind sehr allgemein. Die Autoren führen an, dass weitere Datensätze ausgewertet werden müssten, um eine endgültige Aussage über den negativen oder positiven Einfluss der Data Preparation auf die RL-Qualität zu treffen. Randall et al. (2013) fordern in ihrem Ausblick, dass weitere Forschung zum Einfluss von speziellen Data Preparation Verfahren auf Namens- und Adressattributen durchgeführt werden sollte.

l. Koumarelas et al. (2020) haben das Ziel den Einfluss des RL-Prozessschrittes Data Preparation auf den gesamten RL-Prozess zu untersuchen. l. Koumarelas et al. (2020) sehen noch Verbesserungspotenzial in der standardisierten Beschreibung des RL-Prozessschrittes Data Preparation in den Publikationen, um Ergebnisse reproduzierbar zu machen. Eine geeignete Data Preparation soll die nachfolgenden RL-Prozessschritte einfacher und erfolgreicher machen. Um diese Aussagen zu überprüfen wird ein RL-Prozess implementiert der auf den Datenquellen CDDB, Census, Cora, Hotels, Movies und Restaurants angewendet wird, wobei lediglich im RL-Prozessschritt Data Preparation Änderungen vorgenommen werden, um deren Auswirkungen zu messen. Die übrigen Prozessschritte werden nicht verändert und mit klassischen Algorithmen implementiert, die die Autoren nicht näher beschreiben. Für den RL-Prozessschritt Blocking werden einzelne Attribute auf Gleichheit oder Q-Gramme von Attributen auf Übereinstimmung geprüft. Für den RL-Prozessschritt Comparison wurde die Monge-Elkan-Distanz mit dem inneren String Similarity Measure Levenshtein-Distanz und die Haversine-Distanz ausgewählt. Für den Prozessschritt Classification wurde ein regelbasierter Ansatz mit manuell festgelegten Schwellwerten ausgewählt. Für den RL-Prozessschritt Data Preparation haben l. Koumarelas et al. (2020) elf verschiedene Verfahren ausgewählt. Zu den elf Verfahren gehörten Split attribute, Normalize address, Geocode, Remove special characters, Transliterate UTF-8 to ASCII, Merge attributes, Acronymize, Capitalize characters, Syllabify, Phonetic encode und Stem. In

den Experimenten wurde untersucht, ob die Data Preparation Verfahren einen positiven oder negativen Effekt auf die Ergebnisse des RL-Prozesses haben. Positiver Effekt meint, dass die Matches ähnlicher und die Non-Matches unähnlicher werden. Negativer Effekt meint, dass die Matches unähnlicher werden und die Non-Matches ähnlicher werden. Die Ergebnisse der Publikationen zeigen, dass eine gezielte Auswahl von Data Preparation Verfahren in Abhängigkeit von der vorliegenden Datenquelle einen positiven Effekt auf die RL-Ergebnisse hat.

3.4.4 Zusammenfassung und Bewertung

In diesem Abschnitt sollen die zuvor vorgestellten RL-Forschungsarbeiten einander gegenüber gestellt werden. Zusätzlich wird diese Arbeit in die Gegenüberstellung miteinbezogen, um die Abgrenzung zur existierenden Forschung aufzuzeigen. Für die Bewertung werden, vor dem Hintergrund der Ziele dieser Arbeit, subjektive Kriterien herangezogen, die im Folgenden beschrieben werden:

Data Preparation: Dieses Kriterium beschreibt, ob der RL-Prozessschritt Data Preparation berücksichtigt wird.

Ziel: Dieses Kriterium beschreibt das Ziel der jeweiligen Forschung, das in die zwei Kategorien RL-System oder RL-Problem unterteilt werden kann. Die Kategorie RL-Problem meint, dass für eine vorliegende Menge an Datenquellen ein spezifischer RL-Prozess entwickelt wird, der diese Datenquellen integriert. Die Kategorie RL-System meint, dass ein System entwickelt wird, dass bei für Implementierung von RL-Prozessen Algorithmen und Verfahren bereitstellt.

Manueller Aufwand: Dieses Kriterium beschreibt, ob der manuelle Aufwand während der Entwicklung des RL-Prozesses oder der Bewertung der RL-Ergebnisse berücksichtigt wird.

Generischer RL-Prozess: Diese Kriterium beschreibt, ob ein RL-Prozess entwickelt wird, der auf neue noch unbekannte Datenquellen übertragen werden kann.

In Tabelle 3.17 sind die beschriebenen Forschungsarbeiten mit den vier Kriterien bewertet worden. Zudem ist diese Arbeit in die vier Kriterien eingeordnet worden. Die Forschungsarbeiten Magellan und JedAI berücksichtigen den RL-Prozessschritt Data Preparation nicht. Beide haben als Ziel die Entwicklung eines RL-Systems, das Algorithmen und einen How-to-Guide für die RL-Prozessschritte Blocking, Comparison und Classification bereitstellt, um RL-Prozesse implementieren zu können. Durch die How-to-Guides betrachten beide RL-Systeme den manuellen Aufwand entlang des RL-Prozesses, da dieser reduziert werden soll.

Zudem setzt JedAI auf unsupervised Learning Algorithmen, um den Aufwand der Trainingsdatenerstellung, der bei Magellan existiert, zu vermeiden. Die drei Publikationen, die sich auf die Realwelt-Entität Unternehmen beziehen, und die zwei Publikationen, die sie sich auf die Realwelt-Entität Produkt beziehen, behandeln den RL-Prozessschritt Data Preparation. Als Ziel fokussieren sie ein RL-Problem, da sie einen spezifischen RL-Prozess für die jeweils vorhanden Datenquellen entwickeln. Lediglich die Publikationen zur Realwelt-Entität Produkt gehen auf den manuellen Aufwand ein, indem sie den Aufwand für das Erstellen eines Trainingsdatensatzes mit dem Aufwand für das Erstellen eines Regelwerkes gegenüberstellen. Die zwei Forschungsarbeiten zum RL-Prozessschritt Data Preparation fokussieren ausschließlich die Data Preparation und können daher nicht in die übrigen Kriterien eingeordnet werden.

Tabelle 3.17: Forschungsarbeiten zum Record Linkage

Forschungsarbeit	**Data Preparation**	**Ziel**		**Manueller Aufwand**	**Generischer RL-Prozess**
		RL-System	RL-Problem		
Magellan		x		x	
JedAI		x		x	
Entität Unternehmen	x		x		
Entität Produkt	x		x	x	
Data Preparation	x				
Diese Arbeit	x	x		x	x

Diese Arbeit unterscheidet sich von den vorgestellten Forschungsarbeiten, da ein generischer RL-Prozess entwickelt wird, der auf unbekannte neue Datenquellen übertragen werden kann. Lediglich Magellan hat dies ansatzweise im Zusammenspiel zwischen ihrem RL-System und den How-to-Guides versucht, sind daran allerdings, auch nach eigenen Aussagen (Govind et al. (2019)), gescheitert. Dies könnte darin begründet sein, dass sie den RL-Prozessschritt Data Preparation nicht berücksichtigen. Diese Arbeit wird den RL-Prozessschritt Data Preparation bei der Entwicklung eines RL-Systems berücksichtigen, da dieser mit gezielt eingesetzten Algorithmen und Verfahren, wie l. Koumarelas et al. (2020) festgestellt haben, zur RL-Ergebnisqualität beiträgt. Gleichzeitig soll der manuelle Aufwand im Fokus stehen, da dieser für den Einsatz eines RL-Systems in der Praxis relevant ist. Um den manuellen Aufwand zu reduzieren, soll ein generischer RL-Prozess entwickelt werden, der auf neue unbekannte Datenquellen angewendet werden kann, ohne die erneute Auswahl von Algorithmen und Verfahren entlang des gesamten RL-Prozesses durchführen zu müssen. Dies soll durch die Betrachtung der RL-Prozesse je nach Realwelt-Entität ermöglicht werden, da Datenquellen, die dieselbe Realwelt-Entität behandeln oftmals dieselben Attribute und diese Attribute oftmals dieselben Datenprobleme besitzen, die im RL-Prozess gelöst werden müssen.

4 Datenquellenauswahl im Datenintegrationsprozess

In diesem Kapitel wird die zweite Teilforschungsfrage des Forschungsvorhabens adressiert (siehe Abb. 4.1). Durch die zweite Teilforschungsfrage soll erforscht werden, wie bei der Auswahl der zu integrierenden Datenquellen unterstützt werden kann. Die Datenquellenauswahl ist entscheidend für die Effizienz und Effektivität von Data Science Projekten (vgl. Lin et al., 2016). Oftmals sind die benötigten Datenquellen zu Beginn des Projektes nicht bekannt, da meist eine Vielzahl von internen und externen Datenquellen zur Verfügung stehen (vgl. Kruse, Schröer & Marx Gómez, 2021). Um die Teilforschungsfrage zu beantworten wird die Forschungsmethode zur Erstellung einer Taxonomie von Nickerson et al. (2013) verwendet, die als Ergebnis eine Taxonomie liefert. Im Folgenden wird das Vorgehen zur Beantwortung der zweiten Teilforschungsfrage beschrieben und die Ergebnisse werden präsentiert.

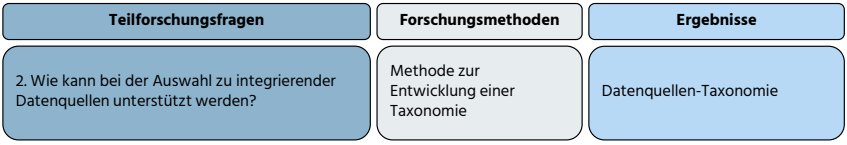


Abbildung 4.1: Einordnung der Datenquellen-Taxonomie in das gesamte Forschungsvorgehen

4.1 Erweiterung des Datenintegrationsprozesses um die Datenquellenauswahl

Bevor der Datenintegrationsprozess entwickelt wird, muss entschieden werden, welche Datenquellen integriert werden sollen. Die oftmals hohe Anzahl an zur Verfügung stehenden Datenquellen macht es Data Scientisten und Fachexperten schwer, diese im Überblick zu behalten. Zumal Data Scientisten und Fachexperten die Datenquellen hinsichtlich einer Vielzahl von technischen und fachlichen Anforderungen des jeweiligen Anwendungsfalls bewerten und auswählen müssen. Fachexperten müssen bspw. beurteilen, ob die Datenquellen die relevanten Informationen für den jeweiligen Use Case beinhalten. Data Scientisten müssen bspw. die technische Möglichkeit und die Aufwände der Datenintegration abschätzen. Die Auswahl geeigneter Datenquellen ist eine entscheidende Aufgabe im Datenintegrationsprozess, da sie die Effizienz und Effektivität beeinflusst (vgl. Lin et al., 2016). Im Datenintegrationsprozess nach Dong und Srivastava (2015) (siehe Abb. 2.2) wird die Datenquellenauswahl nicht berücksichtigt. Keine Publikation mit dem Fokus auf den Datenintegrationsprozess adressiert das vorgelagerte Problem der Datenquellenauswahl. In dieser Arbeit wird der Datenintegrationsprozess um den Schritt der Datenquellenauswahl erweitert (siehe Abb. 4.2), um die End-to-End-Datenintegration zu betrachten (vgl. Kruse, Schröer & Marx Gómez, 2021).

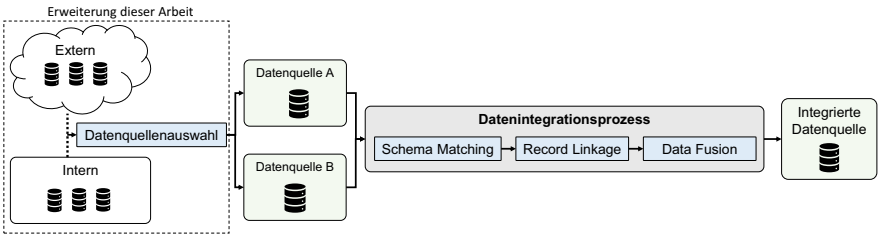


Abbildung 4.2: Erweiterter Datenintegrationsprozess (vgl. Kruse, Schröer & Marx Gómez, 2021)

Es existieren Forschungsarbeiten in den Bereichen Datenquellenauswahl und der Taxonomie-Erstellung zur Beschreibung von Datenquellen. Diese beiden Bereiche werden bisher isoliert betrachtet und nicht gemeinsam mit dem Datenintegrationsprozess in Verbindung gebracht. Die relevanten Forschungsarbeiten aus den beiden Bereichen sollen im Folgenden beschrieben werden.

Für den Bereich Datenquellenauswahl im Kontext von Big Data existieren Publikationen wie die von Safhi, Frikh und Ouhbi (2019). In dieser Publikation wird ein Algorithmus entwickelt, um aus einer vorhandenen Menge von Datenquellen, die Teilmenge relevanter und zuverlässiger Quellen mit den geringsten Kosten zu identifizieren (vgl. Safhi et al., 2019). Voraussetzung für das Verfahren ist, dass alle Datenquelle vorliegen und zugänglich sind, um die entwickelten Metriken zu berechnen. Dabei fassen Safhi et al. (2019) das Problem der Datenquellenauswahl als einen Kompromiss zwischen dem Beitrag der Quelle, ihrer Qualität und den damit verbundenen Kosten zusammen. In der Publikation von Assaf, Senart und Troncy (2016) wird ein Rahmen zur Bewertung der Qualität von offenen Datenquellen - Linked Open Data - entwickelt. Dabei wird ein Werkzeug präsentiert, dass die Profile der Datenquellen erstellt und diese auf Basis von objektiv messbarer Indikatoren bewertet. Auch in dieser Publikation fehlt der Bezug zum Datenintegrationsprozess. In der Publikation von Lin et al. (2016) wird ebenfalls ein Algorithmus zur Bewertung der Datenqualität entwickelt. Der Algorithmus berechnet die Anzahl an zu erwartenden richtigen Werten je Attribut für eine Datenquelle. Mit diesem einzelnen Kriterium kann die Datenquelle mit den meisten wahren Attributen ausgewählt werden (vgl. Lin et al., 2016). Das Verfahren benötigt die vollständige Datenquelle im Zugriff, um den Algorithmus auszuführen. Ebenfalls zielt es lediglich auf das Kriterium Wahrheitsgehalt der Daten ab und hilft die Datenquellen mit dem höchsten Wahrheitsgehalt auszuwählen. Ein Bezug zum Datenintegrationsprozess fehlt. Die Publikation von Dong, Saha und Srivastava (2012) hat das Ziel die Datenquellenauswahl zu unterstützten, sodass die Qualität der Daten und der Datenintegrationsaufwand vor dem Beginn des Datenintegrationsprozesses abgewogen werden können. Die Publikation fokussiert zunächst den

letzten Schritt des Datenintegrationsprozesses, die Data Fusion, in dem die Konflikte der bereits integrierten Datenquellen gelöst werden müssen (vgl. Dong et al., 2012). Aufbauend darauf, erweitern die Autoren in Rekatsinas, Dong und Srivastava (2014) diesen Ansatz für sich verändernde Datenquellen.

Es existieren ebenfalls Forschungsarbeiten, um Datenquellen mit Hilfe einer Taxonomie klassifizieren zu können. In der Publikation von Zrenner, Hassan, Otto und Marx Gómez (2017) wird eine Datenquellen-Taxonomie für die Lieferketten-Transparenz entwickelt. Ziel der Taxonomie ist es, Praktikern und Forschern das Wissen über Datenquellen für Lieferketten zu erweitern. Mit Hilfe der Taxonomie sollen die initiale Datenquellenauswahl unterstützt werden. Doch die Taxonomie beschränkt sich auf Datenquellen zu Lieferketten und bietet keinen Bezug zum Datenintegrationsprozess. Die Publikation von Li, Li, Wang und Gao (2011) präsentiert eine regelbasierte Taxonomie von fehlerhaften und unsauberen Daten. Dabei soll die Taxonomie in Unternehmen helfen, die fehlerhaften und unsauberen Daten besser zu überwachen, zu analysieren und zu bereinigen. Dabei wird in der Publikation eine Methode vorgestellt, um das fehlerhafte und unsaubere Datenauswahlproblem zu lösen, da oftmals nicht alle Datenbereinigungsverfahren aufgrund von Rechenkapazität durchgeführt werden können (vgl. Li et al., 2011). Diese Taxonomie fokussiert die Unterstützung der generellen Data Preparation und nicht des Datenintegrationsprozesses. Die Publikation von Roeder, Muntermann und Kneib (2020) präsentiert eine Taxonomie, um die Heterogenität von Datenquellen zu klassifizieren und damit Forschern und Praktikern bei der Exploration von Datenquellen zu helfen (vgl. Roeder et al., 2020). Bei der Entwicklung der Taxonomie wird der Mehrwehrt und der Wahrheitsgehalt der Datenquellen nicht berücksichtigt. Aus Sicht von Roeder et al. (2020) ist der Mehrwert der Daten schwer objektiv zu messen. Die Evaluation der Taxonomie erfolgt durch das Anwenden dieser auf fünf weiteren Datenquellen. Als Kritikpunkt ist aufzuführen, dass eine Evaluation mit Praktikern oder Forschern, die nicht in den Taxonomie-Erstellungsprozess eingebunden waren, nicht erfolgt (vgl. Roeder et al., 2020).

In dieser Arbeit werden die Forschungsbereiche Datenquellenauswahl und Taxonomie-Erstellung zur Beschreibung von Datenquellen im Kontext des Datenintegrationsprozesses zusammengeführt. Data Scientisten und Fachexperten müssen bei der Datenquellenauswahl die Datenquellen hinsichtlich einer Vielzahl an technischen und fachlichen Kriterien überprüfen und vergleichen. Um das relevante Wissen über eine Datenquelle zu erfassen und den Datenquellenauswahlprozess zu unterstützen, kann eine Taxonomie nützlich sein. Denn in vielen Disziplinen, wie auch der Wirtschaftsinformatik, helfen Taxonomien bei der Klassifikation von Objekten einer Domäne, um bei der Strukturierung und Organisation von Wissen zu unterstützen. In diesem Fall sind die Objekte die Datenquellen. Taxonomien spielen als struktur-gebende Artefakte eine Schlüsselrolle in der Erforschung neuer Forschungsfelder der

Informationssysteme (IS). Um das benötigte Wissen aus den Perspektiven der Data Scientisten und der Fachexperten zu strukturieren und zu organisieren, eignet sich eine Taxonomie (vgl. Nickerson et al., 2013, S. 336; Szopinski, Schoormann & Kundisch, 2019, S. 2).

4.2 Entwicklungsprozess der Datenquellen-Taxonomie

Eine Taxonomie T ist definiert als eine Menge von n Dimensionen $D_i (i = 1, \ldots, n)$. Jede dieser Dimensionen enthält k_i ($k_i \geq 2$) sich gegenseitig ausschließende und insgesamt vollständige Charakteristiken C_{ij} ($j = 1, \ldots, k_i$). Nickerson et al. (2013) definieren, dass ausschließlich ein Charakteristikum aus jeder Dimension einem Objekt zugeordnet werden darf (vgl. Nickerson et al., 2013; Roeder et al., 2020). In der in dieser Arbeit erstellten Taxonomie ist eine Mehrfachauswahl von Charakteristika möglich, um die Nützlichkeit der Taxonomie zu erhöhen.

Für die Entwicklung einer Taxonomie haben Nickerson et al. (2013) einen vielfach verwendeten Prozess entwickelt, der Forscher bei der Entwicklung einer Taxonomie unterstützt. Der Prozess ist in Abbildung 4.3 abgebildet und wird im Folgenden beschrieben.

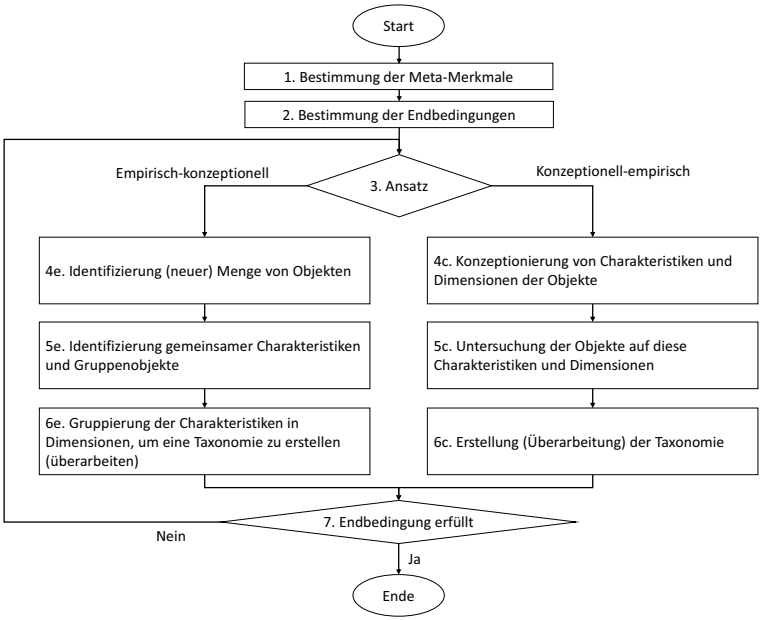


Abbildung 4.3: Methode zur Entwicklung einer Taxonomie (vgl. Nickerson et al., 2013, S. 345)

Im ersten Prozessschritt Bestimmung der Meta-Merkmale soll das Ziel der Taxonomie formuliert werden. Anhand des definierten Ziels kann die Auswahl der Dimensionen und Charakteristiken zielgerichtet erfolgen. Nickerson et al. (2013) empfehlen das Ziel über die potenziellen Nutzer und die damit verbundenen Anwendungsfälle der Taxonomie herzuleiten (vgl. Nickerson et al., 2013). In dieser Dissertation wird das Meta-Merkmal zudem über die definierte Teilforschungsfrage abgeleitet. Das Meta-Merkmal der Datenquellen-Taxonomie lautet:

Die Taxonomie soll Data Scientisten und Fachexperten bei der Auswahl von Datenquellen im Datenintegrationsprozess unterstützen. Es sollen inhaltliche Charakteristika der Datenquelle beschrieben werden. Zusätzlich sollen Charakteristika beschrieben werden, um die technische Möglichkeiten und den technischen Aufwand der Datenintegration abschätzen zu können.

Im zweiten Prozessschritt Bestimmung der Endbedingungen werden objektive und subjektive Endbedingungen für den Prozess definiert. Diese Endbindungen sind notwendig, um den iterativ ablaufenden Prozess zu beenden (vgl. Nickerson et al., 2013). In dieser Publikation werden die von Nickerson et al. (2013) vorgeschlagenen acht objektiven und fünf subjektiven Endbedingungen übernommen und verwendet (siehe Tabelle 4.1).

Nach jeder Iteration werden die Endbedingungen überprüft (Prozessschritt 7 Endbedingung erfüllt). Der Prozess endet, wenn alle Endbedingungen erfüllt sind. Der iterative Prozess beginnt in Schritt drei Ansatz. In diesem Prozessschritt wird entschieden, ob der Ansatz Empirisch-konzeptionell oder Konzeptionell-empirisch in der Iteration verfolgt wird. Die Auswahl des Ansatzes wird durch das Domänenwissen und die verfügbaren Objekte bestimmt. In unserem Taxonomie-Entwicklungsprozess sind die Datenquellen die Objekte. Liegen wenige Datenquellen aber ein signifikantes Domänenwissen vor, wird der Ansatz Konzeptionell-empirisch empfohlen. Wenn kaum Domänenwissen aber viele Datenquellen zur Verfügung stehen, wird der Ansatz Empirisch-konzeptionell empfohlen (vgl. Nickerson et al., 2013; Roeder et al., 2020).

Für die erste Iteration zur Entwicklung der Datenquellen-Taxonomie wurde der Ansatz Konzeptionell-empirisch ausgewählt, da Dömanenwissen über Datenquellen vorhanden ist. Dafür wurden zunächst aus vorhandenen Theorien die Dimensionen und Charakteristika abgeleitet (Prozessschritt 4c Konzeptionierung von Charakteristiken und Dimensionen der Objekte). Für die initiale Erstellung der Taxonomie wurde die Publikation von Zrenner et al. (2017) verwendet, da in dieser eine spezifische Datenquellen-Taxonomie für den Anwendungsbereich Supply Chain Management entwickelt wurde, die zunächst generalisiert werden soll, sodass die Taxonomie für beliebige Anwendungsbereiche verwendet werden kann.

Zusätzlich werden die 15 Information Quality (IQ) Dimensionen von R. Y. Wang und Strong (1996) berücksichtigt, da diese einen bis heute geltenden Überblick über relevante Bewer-

Tabelle 4.1: Iterationen der Taxonomie-Entwicklung und die Endbedingungen - nach Nickerson et al. (2013)

Iteration			**Endbedingungen**
1	*2*	*3*	*Objektiv*
		x	Alle Objekte oder eine repräsentative Teilmenge an Objekten wurde geprüft.
x	x	x	Kein Objekt wurde in der letzten Iteration mit einem ähnlichen Objekt zusammengeführt oder in mehrere Objekte aufgeteilt.
	x	x	Mindestens ein Objekt wird unter jedem Charakteristika jeder Dimension klassifiziert.
		x	Keine neue Dimension oder neues Charakteristikum ist hinzugefügt worden.
		x	Keine Dimension für zusammengeführt oder aufgeteilt.
x	x	x	Jede Dimension ist einzigartig und wird nicht wiederholt.
x	x	x	Jedes Charakteristikum ist innerhalb seiner Dimension einzigartig.
x	x	x	Keine Duplikate in der Kombination von Charakteristika vorhanden.
			Subjektiv
	x	x	Prägnant: Aussagekräftig, ohne unübersichtlich oder überwältigend zu sein
	x	x	Robust: Signifikante und informative Merkmale
		x	Umfassend: Alle Objekte oder eine Stichprobe von Objekten können klassifiziert werden
x	x	x	Erweiterungsfähig: Dimensionen und Charakteristika können leicht hinzugefügt werden
		x	Erläuternd: Die Dimensionen und Merkmale erklären die Objekte

tungsdimensionen von Datenquellen darstellen (vgl. R. Y. Wang & Strong, 1996; Hildebrand, Gebauer & Mielke, 2021, S. 26-29). Diese werden in die IQ-Kategorien (1) Systemunterstützt, (2) Inhärent, (3) Darstellungsbezogen und (4) Zweckabhängig eingeteilt.

Weiterhin wird die Taxonomie von Roeder et al. (2020) in die Entwicklung einbezogen, da diese Taxonomie relevante Eigenschaften von Datenquellen rund um die 5Vs von Big Data betrachtet, die allerdings zu allgemein gehalten werden. Im Gegensatz zu Roeder et al. (2020) werden in dieser Arbeit der Wert (Value) einer Datenquelle und die Big Data Eigenschaft Vertrauenswürdikgeit (Veracity) objektiv beschrieben. Die Vertrauenswürdigkeit einer Datenquelle ist ein schwer messbares Kriterium. Es wird angenommen, dass dies nur über das Arbeiten mit den Daten und das eigenständige Prüfen der Daten verlässlich abgeschätzt werden kann und nur eine Grobe Tendenz bei der Auswahl von unbekannten Datenquellen erfolgen kann.

Im nächsten Prozesschritt 5c Untersuchung der Objekte auf diese Charakteristiken und Dimensionen wurde die Datenquelle des Unitied States Patent and Trademark

Office (USPTO)[23] auf die Taxonomie angewandt. Danach wurde der Prozesschritt 6c Erstellung (Überarbeitung) der Taxonomie durchgeführt und im Anschluss festgestellt, dass nicht alle Endbedingungen erfüllt waren (siehe Tabelle 4.1). Für die zweite Iteration wurde der Ansatz Empirisch-konzeptionell verfolgt. Zunächst wurden in Schritt 4e Identifizierung (neuer) Menge von Objekten die Datenquellen OpenCorporates[24], Crunchbase Open Data Map[25], Crunchbase 2013 Snapshot[26] und Level-1 Daten der Global Legal Entity Identifier Foundation[27] verwendet. Mit diesen Datenquellen wurde die Taxonomie in den Prozessschritten 5e Identifizierung gemeinsamer Charakteristiken und gruppieren von Objekten und 6e Gruppierung der Charakteristiken in Dimensionen, um eine Taxonomie zu erstellen (überarbeiten) weiterentwickelt. Auch nach der zweiten Iteration waren nicht alle Endbedingungen erfüllt (siehe Tabelle 4.1). Die dritte Iteration wurde mit dem Ansatz Empirisch-konzeptionell durchgeführt. Dazu wurden die Datenquellen upcitemdb[28] und ein Datenset der Enigma-Plattform[29] für die Weiterentwicklung der Taxonomie benutzt. Nach der dritten Iteration waren alle Endbedingungen erfüllt (siehe Tabelle 4.1) und der Prozess zur Entwicklung der Taxonomie konnte beendet werden.

4.3 Evaluation der Datenquellen-Taxonomie

Bereits durch den iterativen Taxonomie-Entwicklungsprozess nach Nickerson et al. (2013) ist eine ex-ante Evaluation durchgeführt worden (vgl. Szopinski et al., 2019). In der Publikation von Szopinski et al. (2019) wurde ein Framework zur ex-post Evaluation einer Taxonomie vorgestellt, das für die Evaluation der Datenquellen-Taxonomie genutzt werden soll. Das Framework unterteilt die Evaluation in die Bereiche (1) Teilnehmer der Evaluation, (2) Objekt der Evaluation und (3) Methode zur Evaluation. Szopinski et al. (2019) stellen einige Methoden zur Evaluation einer Taxonomie in ihrer Publikation vor. Für diese Arbeit sind die Methoden Experteninterview, Fokusgruppe und illustratives Szenario relevant und werden im Folgenden näher beschrieben (vgl. Szopinski et al., 2019):

Experteninterview In dieser Evaluationsmethode werden Experten der Taxonomie-relevanten Domäne befragt. Typische Fragen sind bspw. „ist die Taxonomie vollständig?“, „sind alle relevanten Objekte berücksichtigt?“, „sollte die Taxonomie modifiziert werden?“ und „welche Dimensionen oder Charakteristika sollten entfernt oder hinzugefügt wer-

[23] https://developer.uspto.gov/product/patent-grant-bibliographic-dataxml
[24] https://opencorporates.com/
[25] Powered by Crunchbase: https://data.crunchbase.com/docs/open-data-map
[26] https://data.crunchbase.com/docs/2013-snapshot, ©2013 CC-BY
[27] https://www.gleif.org/de/lei-data/gleif-concatenated-file/download-the-concatenated-file
[28] https://www.upcitemdb.com/
[29] Zum Zeitpunkt des Zugriffs noch frei verfügbar: https://public.enigma.com/browse/collection/stock-exchanges-company-listings/50a2457d-6407-4581-8f14-5d37a9410fa9

den?“. Mit Hilfe des Experteninterviews soll die Verständlichkeit, Vollständigkeit, wahrgenommene Nützlichkeit sowie der Abstraktionsgrad der Charakteristika und Dimensionen der Taxonomie ermittelt werden (vgl. Szopinski et al., 2019, S. 9).

Fokusgruppe Die Fokusgruppe ist eine Alternative zum Experteninterview. In dieser Methode wird eine Gruppe von bis zu zehn Personen gemeinsam interviewt, um die Taxonomie zu evaluieren. Mit Hilfe einer Fokusgruppe soll die Vollständigkeit, Robustheit, Verständlichkeit, Ausführlichkeit und Formulierung der Charakteristika und Dimensionen der Taxonomie ermittelt werden. Um die Vollständigkeit zu ermitteln, sollten die Experten ein weiteres Objekt in die Taxonomie einordnen, das nicht Bestandteil der Entwicklung gewesen ist (vgl. Szopinski et al., 2019, S. 9-10).

Illustratives Szenario Die am meisten verwendete Evaluationsmethode ist die Methode illustratives Szenario. Mit dieser Methode wird die Taxonomie in realen Szenarien oder in künstlichen erzeugten Szenarien angewendet, um die Eignung und Nützlichkeit der Taxonomie zu zeigen. Dabei können Realwelt-Objekte oder Literatur über die Realwelt-Objekte genutzt werden, um die Taxonomie anzuwenden (vgl. Szopinski et al., 2019, S. 11-12):

Anwendung auf realen Objekten: Ein illustratives Szenario kann mit Realwelt-Objekten evaluiert werden, die zunächst identifiziert werden müssen. Die identifizierten Objekten sollen dann mit der Taxonomie in die Dimensionen und Charakteristika eingeordnet werden. Dies kann durch Personen geschehen, die an der Entwicklung der Taxonomie beteiligt oder nicht beteiligt gewesen sind. Die Anwendung der Taxonomie auf Realwelt-Objekte ermöglicht es, die praktische Anwendbarkeit für die Klassifizierung, Differenzierung und den Vergleich der Objekte zu bewerten. Weiterhin soll die Robustheit, Nützlichkeit, Effizienz, Stabilität und Vollständigkeit überprüft werden.

Anwendung auf Literatur zu realen Objekten: Innerhalb eines illustrativen Szenarios kann eine Taxonomie mit Forschungsergebnissen über die zu klassifizierenden Objekte evaluiert werden. Dies ist hilfreich, wenn die Realwelt-Objekte für die Evaluation nicht ausreichen. Zusätzlich wird die Reflektion des aktuellen Stands der Forschung über die zu klassifizierenden Realwelt-Objekte gefördert, um Gemeinsamkeiten und Unterschiede in der Literatur zu identifizieren sowie mögliche Forschungslücken aufzudecken. Mit Hilfe der Literatur über Realwelt-Objekte wird die Nützlichkeit, Wirksamkeit, Eindeutigkeit, Verständlichkeit und Vollständigkeit der Taxonomie evaluiert.

Die Evaluation der Taxonomie erfolgte in zehn Evaluationsdurchläufen (siehe Tabelle 4.2). Jeder Evaluationsdurchlauf wurde mit dem Framework von Szopinski et al. (2019) klassifiziert und beschreibt die Teilnehmer, die Objekte und die Methode des jeweiligen Evaluationsdurchlaufs (siehe Tabelle 4.2). Im Folgenden werden die Evaluationsdurchläufe beschrieben:

Teilnehmer der Evaluation: Die Evaluation einer Taxonomie können Forscher und Praktiker durchführen, die bereits bei der Entwicklung der Taxonomie beteiligt waren oder neu in das Forschungsprojekt für die Evaluation aufgenommen worden sind (vgl. Szopinski et al., 2019). Die Evaluation der Datenquellen-Taxonomie wurde mit Forschern der Universitäten Oldenburg und Göttingen durchgeführt (siehe Tabelle 4.2), die sowohl Erfahrung bei der Entwicklung einer Taxonomie als auch mit der Auswahl von Datenquellen besitzen. Weiterhin wurde die Evaluation mit Data Scientisten und Fachexperten aus den Branchen Automobilindustrie, Softwareentwicklung, Fotodienstleistungen, Energieversorgung, Energievertrieb und Finanzdienstleistungen durchgeführt. Damit ist die Zielgruppe der Taxonomie abgedeckt und durch die branchenübergreifende Expertise kann die Allgemeingültigkeit der Taxonomie sichergestellt werden.

Objekt der Evaluation: Die Evaluation kann mit verschiedenen Objekten durchgeführt werden. Zum einen können die Objekte selbst ausgewertet werden, wie in diesem Fall die Datenquellen. Zum anderen kann die Forschung über die Objekte bzw. Datenquellen ausgewertet werden, wie bspw. Publikationen oder Expertenmeinungen über Datenquellen. Für die Evaluation können Datenquellen verwendet werden, die bereits für die Entwicklung der Taxonomie benutzt worden sind oder völlig neue Datenquellen (vgl. Szopinski et al., 2019). Die Evaluation dieser Taxonomie erfolgte in einem Fall (ID 1) mit Datenquellen, die bereits für die Entwicklung der Taxonomie genutzt worden sind (siehe Tabelle 4.2). Die Evaluationsdurchgänge mit den IDs 3, 6, 7 und 9 basierten auf dem Wissen der Teilnehmer über Datenquellen. Die restlichen Evaluationsdurchgänge wurden mit Datenquellen durchgeführt, die im Entwicklungsprozess nicht genutzt worden sind.

Methode zur Evaluation: Es wurden die Methoden Fokusgruppe, Experteninterview und illustratives Szenario (mit realen Datenquellen) verwendet. Das Experteninterview wurde verwendet, wenn eine Person für die Evaluation zur Verfügung stand (siehe Tabelle 4.2, ID's 3, 4, 6, 7 und 9). Wenn mehr als eine Person zur Verfügung stand, wurde die Fokusgruppe verwendet (siehe Tabelle 4.2, ID's 2, 5, 8 und 10). In beiden Methoden wurde den Teilnehmern zunächst die Taxonomie vorgestellt und dann die folgenden offenen Fragen gestellt, die von Szopinski et al. (2019) empfohlen werden: (1) Ist die Taxonomie verständlich und vollständig (2) Sind alle relevanten Objekte in der Taxonomie berücksichtigt worden (3) Welche Dimensionen oder Charakteristiken sollten verändert, hinzugefügt oder gelöscht werden.

In der Evaluationsmethode illustratives Szenario (siehe Tabelle 4.2, ID's 1, 2, 4, 5, 8, 10) wurde die Taxonomie von den jeweiligen Evaluationspartnern auf Datenquellen angewendet wie bspw. Wetterdaten (ID 5), Covid-Daten (ID 4) oder auf unternehmensinternen Daten (ID 2 und 10).

Das Feedback aus den zehn Evaluationsdurchläufen hat zu Anpassungen an der Taxonomie geführt, sodass diese iterativ weiterentwickelt wurde (siehe Tabelle 4.2). Um das Ende der Evaluationsdurchläufe zu bestimmen, wurde erneut auf die subjektiven und objektiven Endbedingungen aus Abbildung 4.3 zurückgegriffen. Nach zehn Evaluationsdurchläufen waren alle Endbedingungen erfüllt, sodass die Taxonomie als final anzusehen ist.

Tabelle 4.2: Übersicht der Evaluation der Taxonomie
FG = Fokusgruppe, EI = Experteninterview, ILS = Illustratives Szenario

Teilnehmer			**Objekt**	**Methode**		
ID	*Branche*	*Rolle*	*Datenquelle*	*FG*	*EI*	*ILS*
1	Universität Oldenburg	Forscher	OpenCorporates[30], Crunchbase 2013 Snapshot[31]			x
2	Automobilindustrie	Fachexperte, Data Scientist	Interne Datenquellen	x		x
3	Universität Göttingen	Forscher	Wissen über Datenquellen		x	
4	Softwareentwicklung	Data Scientist	Covid, Covid Geo, Wiki		x	x
5	Fotodienstleister	Data Scientist	Wetterdaten	x		x
6	Energieversorger	Fachexperte	Wissen über Datenquellen		x	
7	Energievertrieb	Fachexperte	Wissen über Datenquellen		x	
8	Universität Oldenburg	Forscher	UTKFace, IMBD-WIKI	x		x
9	Energieversorger	Data Scientist	Wissen über Datenquellen		x	
10	Finanzdienstleistungen	Data Scientist	Interne Datenquellen	x		x

4.4 Finale Datenquellen-Taxonomie

In diesem Abschnitt wird die finale Datenquellen-Taxonomie (siehe Abbildung 4.4) beschrieben, die aus 16 Dimensionen besteht:

D1: Zugänglichkeit Die Dimension Zugänglichkeit wurde aus der Taxonomie von Zrenner et al. (2017) übernommen und adressiert gleichzeitig die gleichnamige siebte IQ-Dimension von R. Y. Wang und Strong (1996). Die Dimension besitzt die Charakteristika $C_{1j} = \{$*Intern*, *Extern(offen)*, *Extern(geschlossen)*$\}$. Hier wird zwischen internen und externen Datenquellen aus Sicht des Anwenders der Taxonomie unterschieden. Bei den externen Datenquellen gibt es noch die Unterscheidung, ob man für den Zugriff bspw. Logindaten benötigt (*Extern(geschlossen)*) oder ob diese ohne Barrieren zugreifbar (*Extern(offen)*) sind .

D2: Lizenz Die Dimension Lizenz ist während der drei Entwicklungsiterationen entstanden. Sie besitzt die Charakteristika $C_{2j} = \{$*Angabe Open-Source Lizenz, Provider eigene Lizenz, Keine Angabe*$\}$. Unter dem Charakteristikum *Angabe Open-Source Lizenz* soll die vorliegende Open-Source Lizenz wie bspw. MIT oder BSD-3-Clause angegeben werden. Oftmals werden

Dimension	Charakteristik						
D1: Zugänglichkeit	Intern			Extern (offen)		Extern (geschlossen)	
D2: Lizenz	Angabe Open Source Lizenz			Provider eigene Lizenz		Keine Angabe	
D3: Nutzung nach Lizenzablauf	Uneingeschränkt nutzbar			Keine Nutzung und Löschung		Keine Angabe	
D4: Preismodell	Mengengesteuert	Zeitgesteuert		Einmalige Kosten		Kostenfrei	
D5: Interface	API (Datenformat(e))			GUI		Manueller Download (Dump, Datenformate)	
D6: Umfang	Vollständig	Selbst selektierter Auszug der Daten (Selektionskriterien angeben)		Zur Verfügung gestellter Auszug der Daten (Selektionskriterien angeben)			
D7: Sprache	Originäre Sprache(n)			Übersetzt in Sprache(n)		Keine Sprache	
D8: Update	Echtzeit			Zeitintervall		Keine Angabe	
D9: Vorverarbeitung	Schema erstellt oder Metadaten generiert			Metadaten vom Provider		Datenformat beibehalten	
D10: Datenstruktur	Schema (strukturiert)			Schemalos (semistrukturiert oder unstrukturiert)			
D11: Datenstand	Datum oder Versionsangabe						
D12: Abgebildeter Zeitraum	Zeitpunkt oder Zeitraum					Keine Angabe	
D13: Realwelt-Entität	Unternehmen	Person	Produkt	Organisation	Patent	Geografischer Standort	[...]
D13a: Anzahl Datensätze	Angabe Anzahl					Keine Angabe	
D13b: Datenvolumen	Angabe des Volumens					Keine Angabe	
D13c: Anzahl beschreibender Attribute	Anzahl der beschreibenden Attribute					Keine Angabe	
D14: Gesamtes Datenvolumen	Angabe des Datenvolumens					Keine Angabe	
D15: Anzahl Tabellen/Dateien	Angabe der Anzahl Tabellen oder Dateien					Keine Angabe	
D16: Informationsmehrwert	Patente von Unternehmen	Produktinformationen	M&A Daten	Bilanzdaten	Konzernstrukturen	Geographische Verteilung	[...]

Abbildung 4.4: Datenquellen-Taxonomie zur Datenquellenauswahl im Datenintegrationsprozess

bei kommerziellen Datenquellenanbietern individuelle Lizenzvereinbarungen abgeschlossen. Dann sollte das Charakteristikum *Provider eigene Lizenz* ausgewählt werden. Wenn nichts über die Lizenz bekannt ist wird *Keine Angabe* ausgewählt.

D3: Nutzung nach Lizenzablauf Die Dimension Nutzung nach Lizenzablauf ist durch die Evaluation mit den Praxispartnern entstanden. Sie besitzt die Charakteristika $C_{3j} = \{$*Uneingeschränkt nutzbar*, *Keine Nutzung und Löschung*, *Keine Angabe*$\}$. Die Dimension soll die Datenquellen dahingehend beschreiben, wie mit den Daten nach dem Lizenzende umgegangen werden muss.

D4: Preismodell Die Dimension Preismodell ist aus der Taxonomie von Zrenner et al. (2017) übernommen worden. Sie besitzt die Charakteristika $C_{4j} = \{$*Mengenbasiert, Zeitorientiert, Einmalige Kosten, Kostenfrei, Eigene Datenquelle*$\}$. Diese Dimension soll das Preismodell der Datenquelle beschreiben. In dieser Dimension ist eine Mehrfachauswahl möglich, da bspw. eine Kombinaton aus Mengenbasiertem und Zeitorientiertem Preismodell möglich sind, wie bspw. bei OpenCorporates mit 20000 Anfragen pro Monat im Basistarif. Wenn eine interne Datenquelle klassifiziert wird, sollte hier das Charakteristikum *Eigene Datenquellen* ausgewählt werden.

D5: Interface Die Dimension Interface ist während der drei Entwicklungsiterationen entstanden und soll die Zugriffsmöglichkeiten des Anwenders auf die Datenquelle beschreiben. Hierzu dienen die Charakteristika $C_{5j} = \{$*API, GUI, Manueller Download, Datenträger*$\}$. Eine Mehrfachauswahl ist möglich. Bei der Auswahl der Charakteristika sollte zusätzlich

angegeben werden, welche Datenformate wie XML, JSON oder CSV bereitgestellt werden. Das Charakteristikum *Datenträger* wird ausgewählt, wenn die Datenquelle bspw. über eine Festplatte oder einen USB-Stick zur Verfügung gestellt wird.

D6: Umfang Diese Dimension soll beschreiben, welcher Umfang der Datenquelle zur Einordnung in die Taxonomie vorliegt. Hierzu sollen die Charakteristika $C_{6j} = \{$*Vollständig, Selbst selektierter Auszug, Auszug vom Provider*$\}$ genutzt werden. Wenn die Datenquelle nicht vollständig vorliegt, soll angegeben werden anhand welcher Kriterien durch den Anwender oder den Provider die Selektion eines Auszugs der Daten erfolgte. Die Dimension ist während der drei Entwicklungsiterationen definiert worden. Beispielsweise wird von Crunchbase ein Auszug der Daten aus dem Jahr 2013 zur Verfügung gestellt.

D7: Sprache Die Dimension Sprache beschreibt die vorliegende Sprache in der Datenquelle. Die Dimension ist während des Entwicklungsprozesses entstanden und wurde in der Evaluation erweitert. Mit den Charakteristika $C_{7j} = \{$*Sprache(n), Übersetzt in Sprache(n), Keine Sprache*$\}$ soll die Datenquelle beschrieben werden. Es sollen die Sprachen angegeben werden, die in der Datenquelle auftreten wie bspw. deutsch oder englisch. Während der Evaluation wurde eine Datenquelle klassifiziert, die durch den Datenprovider in eine einheitliche Sprache übersetzt worden ist, wofür das Charakteristikum *Übersetzt in Sprache(n)* aufgenommen worden ist. Wenn die Datenquelle keine Sprache enthält, sondern bspw. nur aus numerischen Werten besteht, wird *Keine Sprache* ausgewählt. Auch dieses Charakteristikum ist während der Evaluation eines Praxispartners entstanden, der eine Sensordatenquelle klassifiziert hat.

D8: Update Die Dimension Update beschreibt mit den Charakteristika $C_{8j} = \{$*Echtzeit, Zeitintervall, Keine Angabe*$\}$ die Aktualisierung der vorliegenden Datenquelle. Dabei kann die Datenquelle in Echtzeit stetig aktualisiert werden oder in einem bestimmten Zeitintervall, das angegeben werden sollte. Ist nichts über eine Aktualisierung der Datenquelle bekannt, wird *Keine Angabe* ausgewählt. Diese Dimension ist im Entwicklungsprozess entstanden und aus der neunten IQ-Dimension von R. Y. Wang und Strong (1996) abgeleitet worden.

D9: Datenvorverarbeitung Mit der Dimension Datenvorverarbeitung soll beschrieben werden, ob die Datenquelle bereits vorverarbeitet worden ist und auf dieser Basis die Klassifizierung mit der Taxonomie vorgenommen wird. Hierzu sollen die Charakteristika $C_{9j} = \{$*Schema erstellt, Metadaten vom Provider, Datenformat beibehalten*$\}$ genutzt werden. Das Charakteristikum *Schema erstellt* soll ausgewählt werden, wenn bspw. JSON-Dateien in ein strukturiertes Format überführt werden, um einen ersten Überblick über die Datenquelle zu erhalten. Wenn ein Datenanbieter zu unstrukturierten Daten wie bspw. Newsdaten strukturierte Metadaten liefert, sollte das Charakteristikum *Metadaten vom Provider* ausgewählt werden. Liegt bereits eine strukturierte Datenstruktur vor, wird oftmals keine Änderung an der Datenstruktur vorgenommen und das Charakteristikum *Datenformat beibehalten* sollte ausgewählt werden. Die Dimension Datenvorverarbeitung ist während des Evaluationspro-

zesses mit den Praxispartnern entstanden.

D10: Datenstruktur Die Dimension Datenstruktur beschreibt mit den Charakteristika $C_{10j} = \{$*Schema(strukturiert), Schemalos(semi-strukturiert oder unstrukturiert)*$\}$ ob die Datenquelle strukturiert, semi-strukturiert oder unstrukturiert ist. Die Dimension ist im Entwicklungsprozess entstanden.

D11: Datenstand Die Dimension Datenstand besitzt ein Charakteristikum $C_{11j} = \{$*Datum oder Versionsangabe*$\}$ mit dem der Datenstand in Form eines Zeitstempels oder einer Versionsangabe angegeben werden soll. Die Dimension ist im Entwicklungsprozess entstanden.

D12: Abgebildeter Zeitraum Die Dimension Abgebildeter Zeitraum ist während der Evaluation entstanden. Mit den Charakteristika $C_{12j} = \{$*Zeitraum, Zeitpunkt, Keine Angabe*$\}$ soll beschrieben werden, welchen Zeitraum oder Zeitpunkt die Daten in der Datenquelle abdecken. Beispielsweise existieren in der Datenquelle USPTO Patendaten seit 1976, wohingegen die Datenquelle Appanion Daten zum Zeitpunkt Juli 2019 bereitstellt.

D13: Realwelt-Entität Mit dieser Dimension soll beschrieben werden, welche Realwelt-Entitäten in der Datenquelle repräsentiert werden. In der Taxonomie in Abbildung 4.4 ist in der letzten Zelle ([...]) angedeutet, dass die Charakteristika bei Bedarf um weitere Entitäten ergänzt werden können. Aus dem Entwicklungs- und Evaluationsprozess sind die Charakteristika $C_{13j} = \{$*Unternehmen, Person, Produkt, Patent, Geografischer Standort*$\}$ entstanden. Diese Dimension ist relevant für den Datenintegrationsprozess, da identifiziert werden kann ob und über welche Realwelt-Entität die Datenquelle mit anderen Datenquellen verbunden werden könnte.

D13a: Anzahl Datensätze; D13b: Datenvolumen; 13c Anzahl beschreibender Attribute Die Dimensionen 13a, 13b und 13c sollen, für jede in der Datenquelle vorhandene Realwelt-Entität ausgefüllt werden. Die Angabe wie viele eindeutige Datensätze, wie groß das Datenvolumen und wie viele beschreibende Attribute pro Realwelt-Entität existieren, sollen bei der Bewertung des Wertes der Datenquelle helfen. Durch die objektiv quantifizierbaren Kriterien kann abhängig vom Anwendungsfall eingeschätzt werden, ob die Datenquelle potenziell nützlich ist oder nicht. Zusätzlich dient die Anzahl der beschreibenden Attribute als erstes Indiz für die Durchführung des Datenintegrationsprozesses, da abgeschätzt werden kann anhand wie vieler Attribute ein Vergleich der Datensätze durchgeführt werden könnte. Weiterhin kann über die Korrelation der eindeutigen Anzahl an Datensätzen, der Anzahl der Attribute und dem Datenvolumen abgeschätzt werden, wie vollständig die Datenquelle ist und ob die Datenquelle einen angemessenen Umfang (IQ-Dimension 19) und somit auch Relevanz (IQ-Dimension 2) für den jeweiligen Anwendungsfall bietet. Andere Taxonomien wie die von Zrenner et al. (2017) verwenden hier Charakteristika wie hoch, mittel, niedrig, die sehr subjektiv sind. Dies führt dazu, dass diese Dimension nicht nützlich ist, da sub-

jektive Kriterien schwer verglichen werden können. Mit numerischen Kriterien, wie in dieser Taxonomie, können Datenquellen objektiv verglichen werden.

D14: Gesamtes Datenvolumen Diese Dimension soll das gesamte Datenvolumen der Datenquelle erfassen. Wenn dies nicht vorhanden ist, wird das Charakteristikum *Keine Angabe* genutzt. Auch dieses objektive Kriterium dient dem Bewerten der Datenquelle im Vergleich zu anderen Datenquellen.

D15: Anzahl Tabellen/Dateien Diese Dimension soll die Anzahl der vorhandenen Tabellen oder Dateien der Datenquelle beschreiben. Mit diesem objektiven Kriterium soll eine erste Einschätzung ermöglicht werden, ob der Umfang angemessen (IQ-Dimension 19) und die Informationen relevant (IQ-Dimension 2) sein könnten R. Y. Wang und Strong (1996).

D16: Informationsmehrwert Diese Dimension Informationsmehrwert ist während des Entwicklungsprozesses entstanden und während des Evaluationsprozesses weiterentwickelt worden. Mit dieser Dimension und den Charakteristika hatten die Praxisexperten und Forscher in der Evaluation die größten Verständnis- und Anwendungsschwierigkeiten. Diese Dimension soll dazu dienen das Big Data Merkmale Value und die IQ-Dimension 2 Mehrwert möglichst objektiv für die Datenquelle zu erfassen. In der finalen Taxonomie (siehe Abb. 4.4) sind einige Charakteristika abgebildet, wie bspw. Produktinformationen, Bilanzdaten oder Konzernstrukturen. Eine Mehrfachauswahl ist möglich und die Charakteristika sollen erweiterbar sein, wenn weitere Datenquellen mit neuen Informationsmehrwerten erfasst werden. Wichtig ist, dass die Charakteristika nicht zu detailliert und nicht zu grob erfasst werden, da zum einen der Erfassungsaufwand aber zum anderen auch der Informationsgehalt des Charakteristikums wichtig sind. Eine wichtige Anforderung für die Operationalisierung der Taxonomie ist es, dass die Charakteristika dieser Dimension zentral gepflegt und erweitert werden, damit die Charakteristika der Dimension disjunkt bleiben.

4.5 Anwendungsbeispiel der Datenquellen-Taxonomie

Nachdem die finale Taxonomie mit ihren 16 Dimensionen vorgestellt worden ist, soll in diesem Kapitel ein Anwendungsbeispiel der Taxonomie erfolgen. Dazu werden die zwei Datenquellen Crunchbase und USPTO Patent mit der Taxonomie klassifiziert. Das Ergebnis der klassifizierten Datenquellen ist in Abbildung 4.5 zu sehen. Um den Nutzen der Taxonomie für Data Scientisten und Fachexperten aufzuzeigen wird in Abschnitt 4.5.1 die Datenintegrations-Perspektive und in Abschnitt 4.5.2 die Fachliche-Perspektive auf die Taxonomie beschrieben.

Dimension	Crunchbase				USPTO Patent			
D1: Zugänglichkeit	Extern (geschlossen)				Extern (offen)			
D2: Lizenz	Provider eigene Lizenz				Provider eigene Lizenz			
D3: Nutzung nach Lizenzablauf	Keine Angabe				Uneingeschränkte Nutzung			
D4: Preismodell	Zeitorientiert				Kostenfrei			
D5: Interface	API, GUI, Manueller Download				API, GUI, Manueller Download (XML)			
D6: Umfang	Snapshot aus dem Jahr 2013				Selbst selektierter Auszug (20.08 – 27.08.2019)			
D7: Sprache	Englisch				Englisch			
D8: Update	Echtzeit				Wöchentlich			
D9: Vorverarbeitung	Datenformat beibehalten				Schema erstellt			
D10: Datenstruktur	Strukturiert (MySQL Dump)				Semistrukturiert (XML)			
D11: Datenstand	2013				20.08 – 27.08.2019			
D12: Abgebildeter Zeitraum	Keine Angabe				1976 - heute			
D13: Realwelt-Entität	Unternehmen	Person	Produkt	Geografischer Standort	Unternehmen	Person	Patent	Geografischer Standort
D13a: Anzahl Datensätze	118.342	117.318	25.059	7.976	110.984	209.123	114.138	24.682
D13b: Datenvolumen	166 MB				12 MB	25 MB	14 MB	12 MB
D13c: Anzahl beschreibender Attribute	40				10	8	7	3
D14: Gesamtes Datenvolumen	300 MB				200 MB			
D15: Anzahl Tabellen/Dateien	10 Tabellen				8 Tabellen			
D16: Informationsmehrwert	Weltweite Unternehmen, Finanz- und Börseninformationen, Unternehmensstrukturen, Beziehungen von Unternehmen, Personen und Produkten, Akademische Ausbildung von Mitarbeitern, M&A Informationen				Patentanmeldungen in den USA, Patentzusprüche in den USA, Beziehungen zwischen Personen, Unternehmen und Patenten, Technologieklassifikationen			

Abbildung 4.5: Anwendung der Taxonomie auf die Datenquellen Crunchbase und USPTO

4.5.1 Datenintegrations-Perspektive

Aus der Datenintegrations-Perspektive sollten die technischen Anforderungen für die Integration der Datenquellen betrachtet werden. Hier könnten bspw. folgende Fragen gestellt werden:

1. **Über welche Realwelt-Entität(en) können die Datenquellen integriert werden?** Die Dimension D13 liefert die Informationen, dass die zwei Datenquellen über die Entitäten Unternehmen, Person oder Geografischer Standort miteinander integriert werden könnten. Eine Integration über die Entität Patent der Datenquelle USPTO Patent ist nicht möglich, da Crunchbase diese Entität nicht enthält.

2. **Existieren beschreibende Attribute für den Vergleich der Realwelt-Entitäten?** Die Anzahl der Attribute, die für den Vergleich herangezogen werden können, ist in Dimension D13a enthalten. Diese Information dient für die erste Einschätzung wie erfolgreich und aufwändig die Integration werden könnte, denn je mehr zu vergleichende Attribute im Datenintegrationsprozessschritt Schema Matching identifiziert werden, desto höher ist die Chance auf gute Ergebnisse Christen (2019). Die Crunchbase Realwelt-Entitäten sind in einer gemeinsamen Tabelle modelliert, die aus 40 Attributen besteht. Die USPTO Patent enthält 10 Attribute für die Realwelt-Entität Unternehmen. Dies bedeutet für die Integration, dass der kleinste gemeinsame Nenner von zehn Attributen mit 40 Attributen abgeglichen (Schema Matching) werden muss. Es wird angenommen, dass mehr Attribu-

te das Ergebnis des Datenintegrationsprozesses verbessern, diesen aber gleichzeitig auch aufwändiger machen.

3. **Wie hoch ist die Datenmenge?** Liegt lediglich ein Ausschnitt der Datenquelle vor (D6) und soll mit diesem zunächst die Datenintegration durchgeführt werden, ist die Dimension 11 relevant. In dieser Dimension ist dokumentiert, ob die ursprüngliche Datenstruktur beibehalten oder vorverarbeitet worden ist. Beispielsweise wurde die ursprüngliche XML Struktur (D10) der USPTO Patent Datenquelle in eine strukturierte Form (D9) mit acht Tabellen (D15) überführt. Über die Anzahl der Datensätze (D13a) und das Datenvolumen (D13b) kann die benötigte Rechenkapazität für die Datenintegration abgeschätzt werden. Sollen die Datenquellen Crunchbase und USPTO Patent über die Realwelt-Entität Unternehmen integriert werden, müssten 118.342 x 110.984 Datensätze miteinander verglichen werden. An dieser Stelle kann der Data Scientist eine erste Einschätzung erhalten, ob ein Blockingverfahren eingesetzt werden sollte, um die Anzahl der Vergleiche im RL-Prozess zu reduzieren.

4. **Wie ist die Datenquelle strukturiert?** Die Struktur der Datenquelle kann aus D10 abgelesen werden. Der Aufwand des Datenintegrationsprozesses erhöht sich, wenn semi- oder unstrukturierte Datenquelle vorliegen, da für den Datenintegrationsprozess strukturierte Daten benötigt werden.

5. **Wie kann auf die Datenquelle zugegriffen werden?** Die Dimension D5 bietet die Information, wie auf die Datenquelle zugegriffen werden kann. Sowohl Crunchbase als auch USPTO Patent bieten eine API, GUI und den manuellen Download als Interface an.

4.5.2 Fachliche-Perspektive

Die Datenquellenauswahl aus der fachlichen-Perspektive hängt oftmals vom Einsatzzweck der Datenquellen ab. Deshalb stellt die Taxonomie objektive und dadurch vergleichbare Dimensionen und Charakteristika zur Verfügung, die abhängig vom Anwendungsfall genutzt, bewertet und priorisiert werden können. Mit der vorliegenden Taxonomie können aus der fachlichen-Perspektive bspw. die folgenden Fragen beantwortet werden:

1. **Beinhalten die Datenquellen den benötigten Informationsmehrwert?** Die Informationsmehrwerte der Datenquelle sind in Dimension 16 abgebildet. In Abbildung 4.5 ist für die Datenquelle Crunchbase zu sehen, dass diese bspw. die Informationsmehrwerte Finanzinformationen, Börseninformationen und M&A Informationen enthält. Die Datenquelle USPTO Patent enthält Informationen über Patentanmeldungen in Amerika und deren Technologieklassifikation. Benötigt ein Fachexperte Informationen über die M&A Transaktionen von Unternehmen, die Patente in Amerika besitzen, dann können diese

beiden Datenquellen hilfreich sein. Benötigt ein Fachexperte Informationen über Patente, die in Europa angemeldet worden sind, ist keine der beiden Datenquellen hilfreich. Zusätzlich kann der Fachexperte über die Dimension 6 beurteilen, auf welcher Grundlage die Dimension 11 - 16 für eine Datenquelle erfasst worden sind. Mit den Dimensionen D13, D13a, D13b, D13c kann der Fachexperte zudem abschätzen, ob die Datenquellen einen ausreichenden Umfang für den jeweiligen Anwendungsfall bereitstellen. Während der Zusammenarbeit mit den Praxispartnern wurde häufig darüber diskutiert, in welchen Regionen, Ländern oder Kontinenten bestimmte Unternehmensdatenquellen eine vollständige und hohe Datenqualität aufweisen und in welchen Regionen, Ländern und Kontinenten nicht. Eine deskriptive Statistik über die Verteilung der in der Datenquelle vorhandenen Unternehmen je Land, wie in Abbildung 4.6 zu sehen, liefert die benötigten Informationen für den Vergleich der Datenquellen. Diese deskriptive Statistik kann genutzt werden, um zu entscheiden, ob der Umfang der Datenquelle für den jeweiligen Anwendungsfall ausreichend ist. In der Taxonomie wurde das Charakteristikum *Geografische Verteilung* als Verweis auf die deskriptive Statistik modelliert. Die Statistik sollte nicht Bestandteil der Taxonomie werden, um diese einfach zu halten. Die Statistik sollte eher als Ergänzung geführt werden, da sie in der Praxis einen hohen Mehrwert erzeugt und in Richtung des Data Profiling geht.

Crunchbase	
Land	**Anzahl**
Null	75089
USA	25032
GBR	3713
IND	2198
CAN	1817
DEU	909
...	

USPTO	
Land	**Anzahl**
US	54208
JP	17183
KR	7611
CN	6009
DE	5354
TW	3081
...	

Abbildung 4.6: Geografische Verteilung der Unternehmen in den Datenquellen Crunchbase und USPTO Patent

2. Sind die Datenquellen vertrauenswürdig genug? Die Vertrauenswürdigkeit einer Datenquelle ist ein schwer messbares Kriterium. Es wird angenommen, dass die Vertrauenswürdigkeit vor allem durch den Umgang mit den Daten selbst eingeschätzt werden kann. Die IQ-Dimensionen Glaubwürdigkeit (1), Vollständigkeit (10), Genauigkeit (4) und Interpretierbarkeit (5) von R. Y. Wang und Strong (1996) werden durch die Big Data Eigenschaft Veracity abgedeckt. Die Taxonomie soll durch die Angabe des Namens des Datenanbieters, des Lizenzmodells in Dimension 2, des Preismodells in Dimension 4 und der Dimensionen 13a, 13b und 13c bei einer ersten Einschätzung der Vertrauenswürdigkeit der Datenquelle unterstützen.

3. **Sind die Daten aktuell und reichen weit genug in die Vergangenheit zurück?** Die Aktualität der Datenquellen kann aus Dimension 8 und 12 abgelesen werden. In Dimension 12 wird angegeben, ob die Datenquelle nur einen Zeitpunkt oder ob sie einen Zeitraum repräsentiert. Beispielsweise liefert die Datenquelle USPTO Patent Daten von 1976 bis heute. In Dimension 8 wird angegeben, ob und wie die Datenquelle aktualisiert wird. Die Datenquelle USPTO Patent wird z.B. wöchentlich aktualisiert. Die Aktualisierungshäufigkeit der Datenquelle ist relevant für die Operationalisierung einer Entscheidungsunterstützung. Denn eine Entscheidung, die täglich getroffen werden muss, erfordert oft eine Datenquelle mit täglich aktualisierten Informationen.

4. **Wie ist das Lizenzmodell der Datenquellen?** Mit den Dimensionen Lizenz (D2) und Nutzung nach Ablauf der Lizenz (D3) erhält der Fachexperte die Information über das Lizenzmodell der Datenquelle. Zudem relevant für den Fachexperten ist die Nutzung der Daten nach Ablauf der Lizenz. Der Grund dafür ist, dass alle Anwendungen, die auf Basis der Datenquelle entwickelt wurden, nach Ablauf der Lizenz ggf. nicht mehr verwendet werden dürfen.

5. **Was kostet die Datenquelle?** Mit der Dimension 4 Preismodell kann der Fachexperte die Kosten der Datenquelle abschätzen und bei der wirtschaftlichen Bewertung des jeweiligen Anwendungsfalls mit einbeziehen.

Neben den zuvor beschriebenen Fragen kann der Fachexperte die Taxonomie nutzen, um bspw. die Auswahl von Datenprovidern durchzuführen. Durch das Klassifizieren der Datenquellen der Datenprovider kann der Fachexperte diese untereinander vergleichen und auf Basis der ausgefüllten Taxonomien eine Auswahl treffen. Um den Aufwand zu reduzieren, könnte die Taxonomie von den Datenprovidern ausgefüllt werden und der Fachexperte erhält eine Übersicht der potenziell zu erwerbenden Datenquellen.

5 Konzept Datenquellen-unabhängiger Record Linkage-Prozesse

In diesem Kapitel wird das Vorgehen zur Entwicklung Datenquellen-unabhängiger RL-Prozesse zur Integration der Datenquellen beschrieben. Für die Entwicklung Datenquellen-unabhängiger RL-Prozesse wird, im Gegensatz zum aktuellen Stand der Forschung, in dieser Arbeit der Prozessschritt Data Preparation miteinbezogen. Die Auswahl der Data Preparation Verfahren hängt wiederum von den zu Grunde liegenden Datenproblemen der Datenquellen ab. Die Datenprobleme können beliebig komplex werden, wenn die Datenquellen nicht eingeschränkt werden. Daher ist das Vorgehen dieser Arbeit, die zu integrierenden Datenquellen auf eine spezifische Realwelt-Entität wie Person oder Unternehmen zu reduzieren, um für diese Datenquellen-unabhängige RL-Prozesse zu entwickeln. Durch dieses Vorgehen können die Datenquellen und die darin enthaltenen Datenprobleme reduziert werden, um die Komplexität der RL-Prozess Entwicklung zu reduzieren. In dieser Arbeit wird angenommen, dass eine Realwelt-Entität wie ein Unternehmen in Datenquellen häufig durch wiederkehrende gemeinsame Informationen repräsentiert wird, wie bspw. Name, Adresse und Rechtsform. Da für den RL-Prozess ausschließlich die Informationen einer Realwelt-Entität relevant sind, die in beiden zu integrierenden Datenquellen vorliegen, sollten diese fokussiert werden. Für die häufig vorkommenden Informationen einer Realwelt-Entität wird angenommen, dass sich über verschiedene Datenquellen häufig vorkommende Datenintegrationsprobleme identifizieren und generalisieren lassen, sodass ein Datenquellen-unabhängiger RL-Prozess entwickelt werden kann. Um diese Annahmen des Vorgehens zur Entwicklung Datenquellen-unabhängiger RL-Prozesse zu prüfen, wird in diesem Kapitel die Teilforschungsfrage 3 beantwortet (siehe Abb. 5.1):

Welche Datenintegrationsprobleme existieren in Datenquellen mit der Realwelt-Entität Unternehmen?

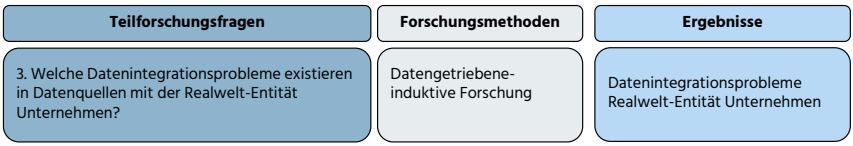


Abbildung 5.1: Einordnung der Konzeption in das gesamte Forschungsvorgehen

Für die Beantwortung der Teilforschungsfrage wird die Realwelt-Entität Unternehmen fokussiert. Zum einen konnten für die Realwelt-Entität Unternehmen die meisten realen und praxisrelevanten Datenquellen identifiziert und zugänglich gemacht werden und zum anderen

wurde durch das Forschungsprojekt TRACE, welches als Evaluation für die Arbeit dient, der Fokus auf die Realwelt-Entität Unternehmen bestimmt. Zur Beantwortung der Teilforschungsfrage wird die datengetriebene-induktive Forschung angewandt, die im Folgenden Abschnitt beschrieben wird.

5.1 Datengetriebene-induktive Forschung

Die datengetrieben-induktive Forschung ist eine Forschungsmethode der Wirtschaftsinformatik. Mit dieser Forschungsmethode können interessante und neue Erkenntnisse aus Daten gewonnen werden, die mehr auf Praxisevidenz als auf wissenschaftlicher Literatur basieren und Nutzung und Nutzen für die Praxis versprechen. Die datengetriebene-induktive Forschung findet häufig Einsatz, wenn keine theoretischen Erkenntnisse für die jeweilige Forschungsfrage existieren (vgl. Robra-Bissantz & Strahringer, 2020, S. 162; Grover & Lyytinen, 2015, S. 285). In der Publikation von Maass, Parsons, Purao, Storey und Woo (2018) wird ein Framework entwickelt, dass die datengetriebene-induktive und die theorie-getriebene Forschung verbindet (siehe Abb. 5.2), da diese laut der Autoren nicht gänzlich voneinander getrennt werden können. Wenn ausschließlich Korrelationen, Trends und Muster aus Daten gewonnen werden, ohne die Domänen-Theorie einzubeziehen, könnten diese möglicherweise nicht zu dauerhaften wissenschaftlichen Erkenntnissen führen. Umgekehrt könnte der Fokus auf die Domänen-Theorie und der Verwendung von wenigen Daten dazu führen, dass weitere Erkenntnisse nicht entdeckt werden, die in größeren Datenmengen enthalten sein könnten (vgl. Maass et al., 2018).

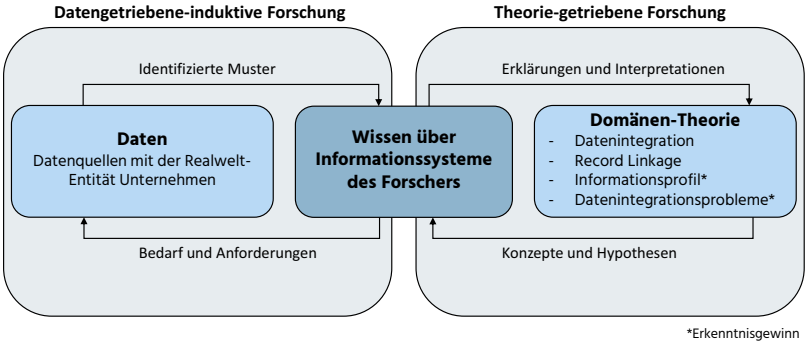


Abbildung 5.2: Datengetriebene-induktive Forschung - angelehnt an Maass et al. (2018)

In Abbildung 5.2 ist das Framework von Maass et al. (2018) auf diese Arbeit übertragen abgebildet. Die zentrale Verbindung zwischen der datengetriebenen-induktiven und der theorie-

getriebenen Forschung ist das Wissen über Informationssysteme des Forschers. Auf Basis der Domänen-Theorie zu Datenintegration und im speziellen RL können die Anforderungen und der Bedarf an die Datenquellen ermittelt werden und somit können die relevanten Daten ausgewählt werden. Die ausgewählten Datenquellen mit der Realwelt-Entität Unternehmen werden dann analysiert, um ein Informationsprofil mit den am häufigst vorkommenden Informationen zur Realwelt-Entität Unternehmen zu erstellen. Das resultierende Informationsprofil stellt einen Erkenntnisgewinn der Domänen-Theorie dar, zu dem bisher keine Theorie existiert, sodass schwerpunktmäßig datengetrieben-induktiv vorgegangen werden muss. Das Informationsprofil wird anschließend verwendet, um erneut datengetrieben-induktiv die Datenintegrationsprobleme der am häufigst vorkommenden Informationen in den Datenquellen zu analysieren. Die identifizierten Datenintegrationsprobleme sind ein weiterer Erkenntnisgewinn für die Domänen-Theorie.

Die identifizierten Datenintegrationsprobleme dienen dann als Grundlage, um einen Datenquellen-unabhängigen RL-Prozess für die Realwelt-Entität Unternehmen zu entwickeln, der diese Probleme meistert.

5.2 Ausgewählte Datenquellen

Die identifizierten und ausgewählten Datenquellen für die datengetriebene-induktive Forschung werden in diesem Abschnitt vorgestellt. Die wesentliche Anforderung besteht darin, dass die Datenquelle die Realwelt-Entität Unternehmen repräsentiert. Weiterhin musste die Datenquelle zugänglich sein, um die modellierten Informationen zur Realwelt-Entität Unternehmen erfassen zu können.

In Tabelle 5.1 sind alle identifizierten und genutzten Datenquellen für die Erstellung des Informationsprofils für die Realwelt-Entität Unternehmen aufgeführt. Für die Klassifikation und den Vergleich der Datenquellen wurden die Dimensionen Zugänglichkeit (D1), Preismodell (D4), Umfang (D6), Anzahl Datensätze (D13a) und Anzahl Attribute (D13c) aus der Datenquellen-Taxonomie (siehe Abb. 4.4) herangezogen. Insgesamt konnten 18 Datenquellen identifiziert und zugänglich gemacht werden, die alle die Realwelt-Entität Unternehmen beinhalten. Alle Datenquellen sind externe Datenquellen, weshalb in der Spalte Zugänglichkeit lediglich zwischen den Charakteristika offen und geschlossen unterschieden wird. Elf der 18 Datenquellen sind kostenfrei und sieben der 18 Datenquellen sind kommerziell erhältlich. Elf der 18 Datenquellen stehen auszugsweise zur Verfügung und sieben der 18 Datenquellen sind vollständig vorhanden. Die Anzahl der Datensätze reicht von 2 bis zu 24.618.332. Für die Erstellung des Informationsprofils ist die Anzahl der Attribute relevant, die in einer Bandbreite von 1 bis 67 liegt. Die kommerziellen Datenquellen Bureau van Dijk, Crunchbase Open Data Map, Crunchbase Snapshot 2013, Databyte, OpenCorporates, Owler, Uscompa-

nylist - Business und Uscompanylist - Company stellen einen kostenfreien Auszug der Daten für diese Forschungsarbeit bereit. Die kommerzielle Datenquelle Capital IQ hat einen einmaligen Export der vollständigen Datenquelle für dieses Forschungsvorhaben zur Verfügung gestellt. Die Datenquelle DeepMatcher Company ist der einzige RL-Benchmark Datensatz, der die Realwelt-Entität Unternehmen enthält. Die Datenquelle Handelsregister wird auf der Website `https://offeneregister.de/` mit dem Datenstand Januar 2019 zum Download angeboten. Die übrigen kostenfreien Datenquellen wie AlphaVantage, GLEIF, USPTO oder Wikidata konnten heruntergeladen oder über einen API Zugriff zugänglich und nutzbar gemacht werden.

Tabelle 5.1: Übersicht der verwendeten Datenquellen

Name	Zugänglichkeit - extern	Preismodell	Umfang	Unternehmen	
				Datensätze	Attribute
AlphaVantage (o. J.)	geschlossen	kostenfrei	Auszug	2	7
Appanion (o. J.)	offen	kostenfrei	Vollständig	279	3
Bureau van Dijk (o. J.)	geschlossen	kommerziell	Auszug	68.162	25
Capital IQ (o. J.)	geschlossen	kommerziell	Vollständig	24.618.442	19
Crunchbase Open Data Map (o. J.)	geschlossen	kommerziell	Auszug	589.343	17
Crunchbase Snapshot 2013 (o. J.)	geschlossen	kommerziell	Auszug	118.342	40
Databyte (o. J.)	geschlossen	kommerziell	Auszug	14.651	33
DeepMatcher Company (o. J.)	offen	kostenfrei	Vollständig	28.200	1
Enigma NASDAQ (o. J.)	offen	kostenfrei	Vollständig	3411	3
Enigma Nike (o. J.)	offen	kostenfrei	Vollständig	565	6
GLEIF (o. J.)	offen	kostenfrei	Vollständig	1.377.003	67
Handelsregister (o. J.)	offen	kostenfrei	Vollständig	5.305.727	25
OpenCorporates (o. J.)	geschlossen	kommerziell	Auszug	2.525.601	36
Owler (o. J.)	geschlossen	kommerziell	Auszug	10	30
UScompanylist Business (o. J.)	geschlossen	kommerziell	Auszug	147	19
UScompanylist Company (o. J.)	geschlossen	kommerziell	Auszug	107	22
USPTO (o. J.)	offen	kostenfrei	Auszug	110.984	10
Wikidata (o. J.)	offen	kostenfrei	Auszug	534.609	12

5.3 Informationsprofil Realwelt-Entität Unternehmen

In diesem Abschnitt werden die identifizierten Datenquellen aus Tabelle 5.1 genutzt, um das Informationsprofil für die Realwelt-Entität Unternehmen zu bestimmen. Simonini et al. (2018) definieren das Informationsprofil als eine „representation of a real-world entity in data

sources“. Mit den zur Verfügung stehenden Datenquellen wurde erfasst, welche Informationen in den Datenquellen vorhanden sind, um die Realwelt-Entität Unternehmen zu repräsentieren. Dabei wurden die vorhandenen Informationen in die drei Bereiche (1) Unternehmensname, (2) Adresse und (3) weitere Informationen unterteilt (siehe Abb. 5.3). Danach wurden alle 18 Datenquellen analysiert und die vorkommenden Informationen extrahiert und den Bereichen zugeordnet. Dieses Vorgehen soll anhand der Datenquelle Capital IQ beispielhaft aufgezeigt werden, da diese eines der umfangreichsten Informationsprofile enthält. In Tabelle 5.2 ist ein Beispieldatensatz aufgeführt, der das Informationsprofil für das Unternehmen Volkswagen AG zeigt.

Tabelle 5.2: Beispiel Unternehmensdatensatz der Datenquelle Capital IQ

Attributname	**Datensatz**
companyid	377732
companyname	Volkswagen AG
alternatecompanynames	Volkwagen, Volkswagen St. (VW), Volkswagen Aktiengesellschaft, VW
city	Wolfsburg
streetaddress	Berliner Ring 2
streetaddress2	
streetaddress3	
streetaddress4	
zipcode	38440
webpage	www.volkswagenag.com
state	Lower Saxony
country	Germany
isocountry2	DE
isocountry3	DEU
simpleindustrydescription	Automobiles
businessdescription	Volkswagen AG manufactures and sells automobiles primarily in Europe, North America, South America, and the Asia-Pacific. The company operates in four segments: Passenger Cars and Light Commercial Vehicles, Commercial Vehicles, Power Engineering, and Financial Services. [...]

Nach der Analyse aller Datenquellen wurden drei Informationen für den Bereich Unternehmensname erfasst. Zunächst die Information Name, die im Capital IQ Beispiel im Attribut COMPANYNAME vorhanden ist. Die Information Name ist in 17 von 18 Datenquellen vorhanden (siehe Abb. 5.3). Dabei spielt die Datenqualität hinsichtlich Vollständigkeit der Daten an dieser Stelle noch keine Rolle. Wenn eine Information in einer Datenquelle mindestens bei einem Datensatz vorhanden ist, wird sie in das Informationsprofil für die Datenquelle aufgenommen. Die Information Rechtsform befindet sich im Capital IQ Beispieldatensatz ebenfalls im Attribut COMPANYNAME. Die Rechtsform Information ist in 17 der 18 Daten-

quellen vorhanden (siehe Abb. 5.3). Die Information zu alternativen Namen ist im Capital IQ Beispieldatensatz im Attribut ALTERNATECOMPANYNAMES vorhanden und ist insgesamt in drei der 18 Datenquellen vorhanden. Die vollständige Übersicht über den Informationsprofilbereich Unternehmensname ist in Anhang B.1 zu finden.

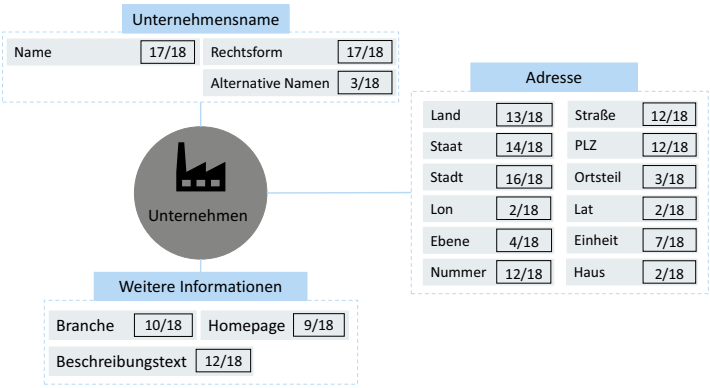


Abbildung 5.3: Informationsprofil Realwelt-Entität Unternehmen

Der Bereich Adresse wurde zunächst nach Comber und Arribas-Bel (2019) in die Adressbestandteile Haus, Einheit, Ebene, Nummer, Ortsteil, Straße, Stadt, Postleitzahl (PLZ), Staat und Land unterteilt (siehe Abb. 5.3). Durch die Analyse der Datenquellen wurden die Bestandteile Longitude und Latitude ergänzt. Die Informationen Einheit, Haus, Ebene, Straße und Nummer kommen in der Capital IQ Datenbank in den Attributen STREETADDRESS, STREETADDRESS2, STREETADDRESS3 und STREETADDRESS4 vor. Im Beispieldatensatz in Tabelle 5.2 sind die Informationen Straße und Nummer vorhanden. Über alle Datenquellen sind die Informationen Haus in zwei von 18, Ebene in vier von 18, Einheit in sieben von 18, Nummer in zwölf von 18 und Straße in zwölf von 18 Datenquellen vorhanden. Die ergänzten Informationen Longitude und Latitude sind jeweils in zwei von 18 Datenquellen vorhanden. Der Ortsteil ist in drei von 18 Datenquellen vorhanden. Die Information Stadt ist in 16 von 18 Datenquellen vorhanden. Im Capital IQ Beispieldatensatz befindet sich die Information Stadt im Attribut CITY. Die PLZ ist in 12 von 18 Datenquellen vorhanden wie bspw. im Attribut ZIPCODE im Capital IQ Beispieldatensatz. Die Information zum Staat ist in 14 von 18 Datenquellen vorhanden. Im Beispieldatensatz ist die Information im Attribut STATE abgebildet. Die Information Land ist in 13 von 18 Datenquellen vorhanden. In der Capital IQ Datenbank werden in den Attributen COUNTRY, ISOCOUNTRY2 und ISOCOUNTRY3 Informationen zum Land modelliert. Die vollständige Übersicht über den Informationsprofilbereich Adresse ist in Anhang B.2 zu finden.

Der Bereich weitere Informationen umfasst Informationen wie die Branche, einen Beschreibungstext und eine Homepage (siehe Abb. 5.3). Ein Beschreibungstext ist in zwölf von 18 Datenquellen vorhanden. Im Capital IQ Beispieldatensatz ist der Beschreibungstext im Attribut BUSINESSDESCRIPTION vorhanden. Die Branchen Information ist in zehn von 18 Datenquellen vorhanden und im Beispieldatensatz in Attribut SIMPLEINDUTRYDESCRIPTION vorhanden. Informationen zur Homepage sind in neun von 18 Datenquellen vorhanden wie im Beispieldatensatz im Attribut WEBPAGE.

5.4 Datenintegrationsprobleme der Realwelt-Entität Unternehmen

Nachdem das Informationsprofil mit den Bereichen Unternehmensname, Adresse und weitere Informationen aufgestellt worden ist, wird nun der Fokus auf die häufigsten Informationen gelegt, um Datenintegrationsprobleme zu identifizieren. In einem ersten Schritt soll der Name, die Rechtsform und der gesamte Adressbereich fokussiert werden, da diese Informationen am häufigsten auftreten (siehe Abb. 5.3). Den Unternehmensnamen und die Rechtsform zu fokussieren, bestätigt zudem die Publikation von Schild und Schultz (2017), die den Unternehmensnamen als wichtigstes Kriterium zur Unterscheidung von Unternehmen ansehen. Gleichzeitig sagen Schild und Schultz (2017), dass die Rechtsform die wichtigste Information ist, die im Unternehmensnamen enthalten ist und extrahiert werden sollte. In dieser Arbeit wird die von Schild und Schultz (2017) getroffene qualitative Aussagen mit quantitativen Zahlen unterstützt. Denn in allen 18 Datenquellen ist die Rechtsform Bestandteil des Unternehmensnamens und liegt in keinem separaten Attribut vor.

Der RL-Prozess hat die Besonderheit, dass die Auswahl der Verfahren und Algorithmen in vorgelagerten Prozessschritten Auswirkungen auf die nachgelagerten Prozessschritte hat. Am Beispiel der Unternehmensrechtsform kann die Bedeutung der Data Preparation für den nachfolgenden Prozessschritt der Comparison aufgezeigt werden. In Abbildung 5.4 ist die Ausgangssituation des Beispiels abgebildet. Es liegen die Datenquellen A und B vor. Datenquelle A enthält die beiden Unternehmen MTU Aero Engines AG und MTU Aero Engines Gesellschaft mbh. Datenquelle B enthält die beiden Unternehmen MTU Aero Engines Gmbh und MTU Aero Engines Aktiengesellschaft. Für den Menschen ist es schnell zu erkennen, dass das Unternehmen mit der ID 1 aus Datenquelle A zum Unternehmen mit der ID 2 aus Datenquelle B und ID 2 aus Datenquelle A zu ID 1 aus Datenquelle B gehört, siehe Spalte Match in Abb. 5.4. In der unteren Tabelle in Abbildung 5.4 ist aufgeführt welche Ähnlichkeit in Prozent die String Similarity Measures Levenshtein, JaroWinkler, Jaccard und Soft TF/IDF bei jeder Datensatzkombination liefern. Aus der Tabelle ist zu entnehmen, dass keine der String Similarity Measures den beiden Tupeln, die ein Match sind, die höchst Ähnlichkeit zuweist. Daraus folgt, dass die Rechtsform im Unternehmensnamen zu Problemen im RL-Prozessschritt Comparison führt.

Datenquelle A

ID	Name
1	MTU Aero Engines AG
2	MTU Aero Engines Gesellschaft mbh

Datenquelle B

ID	Name
1	MTU Aero Engines Gmbh
2	MTU Aero Engines Aktiengesellschaft

Name	Name	Levenshtein	JaroWinkler	Jaccard	Soft TF/IDF	Match
MTU Aero Engines AG	MTU Aero Engines Aktiengesellschaft	51	89	**60**	**43**	**Yes**
MTU Aero Engines AG	MTU Aero Engines Gmbh	**81**	**96**	**60**	**43**	No
MTU Aero Engines Gesellschaft mbh	MTU Aero Engines Aktiengesellschaft	**68**	90	50	71	No
MTU Aero Engines Gesellschaft mbh	MTU Aero Engines Gmbh	63	**92**	**50**	**76**	**Yes**

Abbildung 5.4: Auswirkung der Rechtsform auf die Berechnung der Ähnlichkeitsmaße

Im Prozessschritt Data Preparation soll zunächst die Rechtsform im Unternehmensnamen identifiziert und in ein separates Attribut extrahiert werden, um den Namen ohne Rechtsform vergleichen zu können. Weiterhin sollte die Rechtsform standardisiert werden, damit diese ebenfalls über Ähnlichkeitsmaße verglichen werden kann. Denn die Rechtsform AG mit der Rechtsform Aktiengesellschaft zu vergleichen, stellt die String Similarity Measures ebenfalls vor Probleme. In Abbildung 5.5 ist für das Beispiel aus Abbildung 5.4 dargestellt, wie das Ergebnis der Rechtsform-Extraktion und Standardisierung die Probleme für den RL-Prozessschritt Comparison beseitigt. Durch die Extraktion und Standardisierung der Rechtsform kann der Unternehmensname über die Levenshtein-Distanz und die Rechtsform auf Gleichheit verglichen werden. Dadurch können die beiden korrekten Tupel als Match identifiziert werden, da nur bei diesen der Name und die Rechtsform übereinstimmt.

Name	Rechtsform	Name	Rechtsform	Levenshtein	Gleiche Rechtsform	Match
MTU Aero Engines	AG	MTU Aero Engines	AG	100	1	**Yes**
MTU Aero Engines	AG	MTU Aero Engines	GmbH	100	0	No
MTU Aero Engines	GmbH	MTU Aero Engines	AG	100	0	No
MTU Aero Engines	Gmbh	MTU Aero Engines	GmbH	100	1	**Yes**

Abbildung 5.5: Ziel zur Lösung des Rechtsform Problems

Aus diesen gewonnenen Erkenntnissen ergibt sich das erste generalisierte Problem für den RL-Prozess für die Realwelt-Entität Unternehmen, da diese immer die Rechtsform-Problematik mit sich bringen. Durch die Analyse der vorliegenden Datenquellen kann die Problematik weiter beschrieben werden, da es eine Vielzahl von Rechtsformen auf der gesamten Welt gibt, die in den Datenquellen enthalten sind. Allein für die deutsche Rechtsform GmbH wurden neun verschiedene Schreibweisen im Unternehmensnamen identifiziert (siehe Abb. 5.6), die es zu erkennen, zu extrahieren und zu standardisieren gilt.

ID	Unternehmensname
1	Selbstfahrer Union **G.m.b.H.**
2	GIANT Weilerswist g21 **GmbH**
3	FABIUS Vermietungs**gesellschaft mbH**
4	Infrastrukturentwicklungs**gesellschaft** Hilden **mbH**
5	ITM & C **GmbH** International Trade Marketing & Consulting
6	FHS Gabelstapler **Gesellschaft mit beschränkter Haftung**
7	bunse aufzuege **gesellschaft mit beschraenkter haftung**
8	alint 458 grundstueckverwaltung **gesellschaft m.b.h.**
9	**gesellschaft** zur verwertung von leistungsschutzrechten **mit beschraenkter haftung** gvl

Abbildung 5.6: Überblick verschiedener Repräsentationen einer Unternehmensrechtsform am Beispiel der GmbH

Durch das Analysieren der vorhandenen Datenquellen konnten Datenintegrationsprobleme für die Informationsprofilbereich Adresse identifiziert werden, die ebenfalls im RL-Prozessschritt Data Preparation gelöst werden sollten. In Abbildung 5.7 sind fünf klassische Adressdatenprobleme aufgeführt.

1. **Schreibweise und Format:** In Abbildung 5.7 ist in Beispiel 1 zu sehen, dass die Stadt München in Datenquelle A mit „ü“ und in Datenquelle B mit „ue“ unterschiedlich geschrieben wird. Zudem ist das Land in unterschiedlichen Formaten repräsentiert. In Datenquelle A ist das Land durch das ISO-3166-1 ALPHA-2 Kürzel mit DE angegeben, das für Deutschland steht. In Datenquellen B ist das Land ausgeschrieben mit Deutschland. Es existieren weitere Kürzel-Formate u.a. ISO-3166-1 ALPHA-3, das aus drei Buchstaben besteht, wie bspw. DEU für Deutschland. Weitere Kürzel existieren ebenso für subnationale Verwaltungseinheiten oder Staaten wie das ISO 3166-2 Kürzel. Für Deutschland existiert das Kürzel ISO 3166-2:DE, das die Bundesländer umfasst wie NI für Niedersachsen. Diese Adressinformationen sollten standardisiert werden, da bspw. der Vergleich von DE und Deutschland mit String Similarity Measures keine ausreichend hohe Ähnlichkeit liefert.

2. **Sprache:** Ein weiteres Problem ist es, wenn die Datenquellen die Adresse in unterschiedlicher Sprache enthalten. Das Beispiel 2 in Abbildung 5.7 soll dies verdeutlichen. Datenquelle A enthält die Adressinformationen in deutscher Sprache Koeln und Deutschland während Datenquelle B diese in in englischer Sprache Cologne und Germany enthält. Auch dies stellt die String Similarity Measures vor die Herausforderung einen hohen Ähnlichkeitswert zu liefern.

3. **Detailgrad in den Attributen:** Das Beispiel 3 in Abbildung 5.7 soll zeigen, dass Adressinformationen oftmals in verschiedenen Detailgraden vorhanden sind. Das Unternehmen

1. Schreibweise und Format

Name	Ort	Land	Name	Location_City	Location_region
Tawny Al	München	DE	Tawny Al	Muenchen	Deutschland

2. Sprache

Name	Ort	Land	Name	Location_City	Location_region
Deepl	Koeln	Deutschland	Deepl	Cologne	Germany

3. Detailgrad in den Attributen

Name	Ort	Land	Name	Location_City	Location_region
Job pal	Berlin	DE	Job pal	Kreuzberg	Berlin
Retorio	München	Deutschland	Retorio	Garching bei München	Bayern

4. Vollständigkeit und Normalisierungsgrad

Name	Address	Postal Code	Locality	Region	Country
BioNTech SE	An der Goldgrube 12, 55131 Mainz.				Germany
BIONTECH AG	AN DER GOLDGRUBE 12, 55131 MAINZ		55131 MAINZ		ALLEMAGNE

5. Keine Normalisierung und verschiedene Informationen in den selben Attributen

Name	Data_1	Data_2	Data_3
Facebook Inc.	Menlo Park	CA	US
Siemens AG	DE		

Abbildung 5.7: Überblick Adressdatenprobleme der Realwelt-Entität Unternehmen

Job pal wird in Datenquelle A mit der Stadt Berlin und dem Land DE repräsentiert. In Datenquelle B wird es mit der Stadt Kreuzberg und dem Bundesland Berlin repräsentiert. Diesen Adressinformationen einen hohe Ähnlichkeit über String Similarity Measures zuzuweisen ist nicht möglich. Diese Herausforderung des unterschiedlichen Detailgrads der Adressattribute sollte im RL-Prozessschritt Data Preparation gelöst werden.

4. **Vollständigkeit und Normalisierungsgrad:** Eine weitere Herausforderung stellt die Vollständigkeit und der Normalisierungsgrad der Adressinformationen dar. Das Beispiel 4 in Abbildung 5.7 zeigt ein Beispiel für zwei Datensätze aus der Datenquelle OpenCorporates für das Unternehmen Biontech SE. Die Adressinformationen Straße, Hausnummer, Postleitzahl und Stadt sind in einem Attribut Address enthalten, obwohl eigene Attribute Postal Code und Locality existieren. Im zweiten Beispiel kommt hinzu, dass im Attribut Locality erneut mit der Postleitzahl und der Stadt mehrere Informationen in einem Attribut enthalten sind. Dies erschwert den Vergleich der Adressinformationen. Daher sollte die Herausforderung der Vollständigkeit und des Normalisierungsgrads der Adressinformationen im RL-Prozessschritt Data Preparation gelöst werden.

5. **Keine Normalisierung und verschiedene Informationen in den selben Attributen:** Eine weitere Herausforderung bei Adressinformationen, die in den Datenquellen entdeckt worden sind, zeigt Beispiel 5 in Abbildung 5.7. Die Adressattribute sind lediglich durch-

nummeriert und es ist nicht zu erkennen, welche Information in welchem Attribut enthalten ist. Die inhaltliche Analyse der Attribute zeigt zudem auf, dass in den durchnummerierten Attributen unterschiedliche Adressinformationen enthalten sein können. Im ersten Datensatz ist in Attribut DATA_1 mit MENLO PARK eine Stadtinformation und im zweiten Datensatz mit DE eine Landinformation vorhanden. Die fehlende Normalisierung und die unterschiedlichen Adressinformationen in den selben Attributen erschweren den Adressvergleich mit anderen Datenquellen, sodass diese Herausforderung im RL-Prozessschritt Data Preparation gelöst werden sollte.

Neben der Auswahl der geeigneten Algorithmen und Verfahren für die jeweiligen Datenprobleme ist die Bewertung der Ergebnisqualität der Datenintegration eine große Herausforderung. Wie auch die Publikationen von Barlaug (2020) und Doan et al. (2020) erkannt haben, ist die Bewertung der Ergebnisqualität in der Praxis mit großem manuellen Aufwand verbunden. Denn während die Publikationen der RL-Forschung hauptsächlich synthetische Datensets verwenden (vgl. Mudgal et al., 2018), für die Ground-Truth Daten existieren und somit kein manueller Aufwand für die Bewertung der Ergebnisqualität betrieben werden muss, existieren in der Praxis selten Ground-Truth Daten. Diese Herausforderung soll anhand von Abbildung 5.8 erläutert werden.

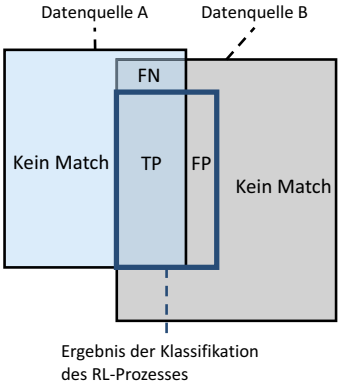


Abbildung 5.8: Herausforderung der Bewertung der Ergebnisqualität beim Record Linkage

In der Praxis soll die Datenquelle A mit der Datenquelle B integriert werden. In Abbildung 5.8 sind alle Mengen, wie TP, FP, FN und Kein Match, die bei der Integration der beiden Datenquellen entstehen beschriftet. Diese vollständigen Informationen erhält man nur, wenn man die Ground-Truth Daten erstellt hat. Diese Ground-Truth erhält man durch das bilden des Kreuzproduktes der beiden Datenquellen und der manuellen Prüfung aller entstandenen

Datensatzpaare. Bereits bei 100 Datensätzen pro Datenquellen entstehen 10.000 Datensatzpaare. Rechnet man mit einer Prüfdauer pro Datensatzpaar von 20 Sekunden, entspricht dies einem Aufwand von 7 PT. Diesen Aufwand, im Vorfeld Ground-Truth Daten zu erstellen, wollen Unternehmen in der Praxis nicht leisten. Beim Einsatz des RL-Prozesses in der Praxis ist dennoch das Ziel, alle korrekten Matches aus beiden Datenquellen zu identifizieren. Die Gesamtmenge der korrekten Treffer besteht aus der TP und FN Menge. Im Prozessschritt Classification des RL-Prozesses werden die Datensatzpaare selektiert, die einen Match darstellen. Die Match Menge ist in Abbildung 5.8 die Summe der TP und FP. Die Match Menge muss ebenfalls manuell überprüft werden, um diese in TP und FP unterteilen zu können, was erneut ungewollten manuellen Prüfaufwand erfordert. Ein fortwährendes Problem bei der Integration von Datenquellen ohne Ground-Truth Daten ist die Identifikation der FN. Die FN sind Datensatzpaare, die einen Match abbilden, allerdings nicht durch die RL Classification als Match identifiziert werden. Zudem wird die Identifikation der FN durch die Vielzahl der vorhandenen Datensätze die keinen Match besitzen erschwert. Dies sind Datensätze die lediglich in einer der beiden Datenquellen existieren und für die somit kein Match existiert. Im Rahmen dieser Arbeit soll ein Ansatz zur Klassifikation der Matches entwickelt werden, der den manuellen Prüfaufwand reduziert.

Zusammenfassend lässt sich festhalten, dass die Informationen zum Unternehmensnamen und den Adressen im RL-Prozessschritt zunächst angereichert und standardisiert werden sollten, um die Herausforderungen für die nachfolgenden RL-Prozessschritte zu reduzieren. Für die identifizierten Herausforderungen sollen nun Algorithmen und Verfahren implementiert werden, die die Data Preparation Komponenten des Datenquellen-unabhängigen RL-Prozesses für die Realwelt-Entität Unternehmen bilden.

6 Prototypische Implementierung des Unternehmen-Matcher

Im vorherigen Kapitel wurde das Konzept vorgestellt, wie die Herausforderungen und Probleme der Datenintegration über die Realwelt-Entität Unternehmen generalisiert werden können. Für die identifizierten Herausforderungen und Probleme wurden Verfahren und Algorithmen implementiert, die den Unternehmen-Matcher bilden. Der Unternehmen-Matcher ist das Ergebnis zur Beantwortung von Teilforschungsfrage 4, siehe Abb. 6.1, und wird in diesem Kapitel vorgestellt.

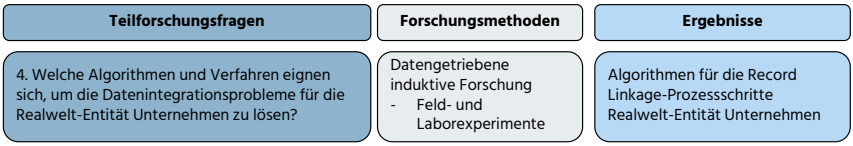


Abbildung 6.1: Einordnung der prototypischen Implementierung in das gesamte Forschungsvorgehen

6.1 Überblick Record Linkage-Prozess Unternehmen-Matcher

In Abschnitt 5.3 wurde herausgearbeitet, dass der Unternehmensname und die Adressdaten die am häufigsten vorkommenden Informationen über Unternehmen in Datenquellen mit der Realwelt-Entität Unternehmen sind. Die Attribute mit Informationen über den Unternehmensnamen und Adressdaten werden in dieser Arbeit fokussiert und es wurden bereits erste Probleme und Herausforderungen der Datenintegration der Realwelt-Entität Unternehmen in Abschnitt 5.4 identifiziert und beschrieben.

Für die einzelnen Probleme und Herausforderungen wurden Softwarekomponenten konzeptioniert und entwickelt, die den Unternehmen-Matcher darstellen, welcher in Abbildung 6.2 abgebildet ist. Der Unternehmen-Matcher bietet einen RL-Prozess, um beliebige Datenquellen mit der Realwelt-Entität Unternehmen ohne die Auswahl und Konfiguration von Algorithmen zu integrieren und reduziert den manuelle Prüfaufwand der Ergebnisse.

Der Unternehmen-Matcher enthält Softwarekomponenten für die klassischen RL-Prozessschritte Data Preparation, Blocking, Comparison, Classification und Evaluation. Zuerst werden die Daten der zu integrierenden Datenquellen im Prozessschritt Data Preparation aufbereitet und in ein einheitliches Datenschema überführt. Hierzu wurde der RechtsformService entwickelt, um die Unternehmensrechtsform im Unternehmensnamen zu klassifizieren,

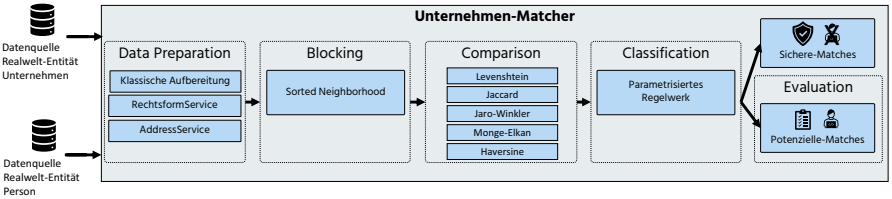


Abbildung 6.2: Gesamter Record Linkage-Prozess des Unternehmen-Matcher

zu extrahieren und zu vereinheitlichen. Die Aufbereitung des Unternehmensnamens wird in Abschnitt 6.3 näher beschrieben. Für die Aufbereitung der Adressdaten und das Behandeln der häufig vorkommenden Adressdatenprobleme wurde der AdressService entwickelt, der in Abschnitt 6.4 detailliert beschrieben wird. Die Auswahl des Sorted Neighbourhood Algorithmus für das Blocking und die Auswahl der Ähnlichkeitsmaße für den Prozessschritt Comparison werden in Abschnitt 6.5 und 6.6 näher beschrieben.

Eine weitere konzeptionierte und entwickelte Komponente des Unternehmen-Matcher ist das parametrisierte Regelwerk zur Einteilung der Datensatzpaare in die Kategorien Match und Kein-Match. Im Gegensatz zu vielen RL-Forschungsarbeiten, die überwachte Lernverfahren für die Klassifikation der Datensatzpaare trainieren und einsetzen, ist für den Unternehmen-Matcher ein parametrisiertes Regelwerk entwickelt worden, um den manuellen Aufwand der Trainingsdatenerstellung zu vermeiden. Das Regelwerk wurde iterativ durch neun Datenintegrationsexperimente entwickelt und besteht aus Regeln, die die Datensatzpaare in sichere-Matches und potenzielle-Matches unterteilen. Der große Vorteil des parametrisierten Regelwerkes wirkt sich auf den letzten Prozessschritt, die Evaluation, aus. Die sicheren-Matches müssen keiner manuellen Prüfung unterzogen werden. Lediglich die potenziellen-Matches sollten manuell überprüft werden, wodurch insgesamt eine Reduktion des manuellen Prüfaufwandes herbeigeführt wird. Das entwickelte parametrisierte Regelwerk wird in Abschnitt 6.7 detailliert beschrieben.

6.2 Forschungsmethode Feldexperiment

Für die Entwicklung der Komponenten des Unternehmen-Matcher wurde das Feldexperiment als methodisches Vorgehen gewählt. Durch Experimente werden Kausalzusammenhänge in kontrollierter Umgebung untersucht, indem die Experimentalvariable wiederholbar manipuliert und die Auswirkung dieser Manipulation ermittelt wird. Der zu untersuchende Kontext wird entweder im Feld, in natürlicher Umgebung, oder im Labor, in künstlicher Umgebung, erforscht. Daher entscheidet die Umgebung des Experiments, ob es sich um ein Feld- oder Laborexperiment handelt (vgl. Wilde & Hess, 2007; Müllerleile, 2019, S. 115).

Aus Sicht des Autors eignen sich Feldexperimente sehr gut für die Lösung von Data Science Problemen. In dieser Arbeit wird RL als eine spezielle Ausprägung von Data Science Problemen definiert, wie in der Publikation von (vgl. Govind et al., 2019), sodass der Einsatz von Feldexperimenten auf RL-Probleme übertragen werden kann. Data Science Probleme besitzen immer einen zu untersuchenden Kontext, wie bspw. das Problem der Rechtsformerkennung im Unternehmensnamen. Um dieses Problem zu lösen, soll die Rechtsform im Unternehmensnamen klassifiziert und extrahiert werden. Dieses Problem sollte idealerweise in seiner natürlichen Umgebung mit realen und praxisrelevanten Datenquellen untersucht werden. Die Experimentalvariablen in Data Science Experimenten sind die verwendeten und kombinierten Algorithmen und Verfahren sowie deren Parametereinstellungen. Die Auswirkungen des explorativen und iterativen Einsatzes verschiedener Algorithmen und Verfahren sowie deren Parametereinstellungen, die Manipulation der Experimentalvariablen, wird meist über die Evaluationsmetriken Precision, Recall, Accuracy und F1-Score gemessen und bewertet.

Für die Durchführung der Feldexperimente wurde zunächst die nötige IT-Infrastruktur aufgebaut. Die IT-Infrastruktur ist in Abbildung 6.3 dargestellt.

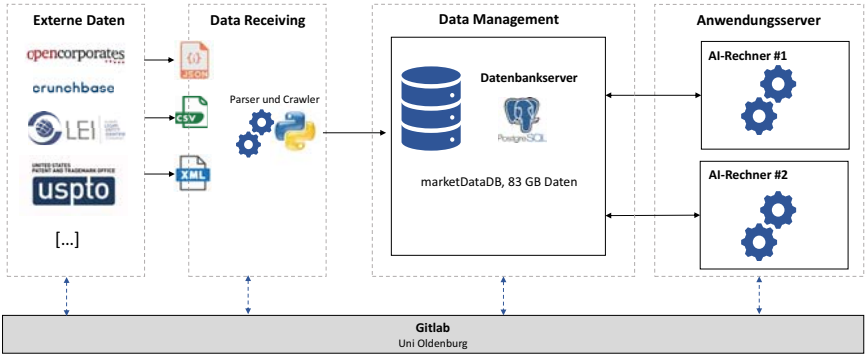


Abbildung 6.3: IT-Infrastruktur für die Durchführung der Feldexperimente

Zunächst wurde ein Datenbankserver mit einer PostgreSQL Datenbank für das Datenmanagement eingerichtet. Für die Forschungsarbeit wurden 83 GB Daten in die Datenbank importiert und verwendet. Zusätzlich wurden die in der Abteilung VLBA zur Verfügung stehenden Server für die Entwicklung der Software eingerichtet. Hierzu wurde eine Python-Entwicklungsumgebung eingerichtet, die u. a. Jupyter[32] enthält. Daraufhin wurden alle zur Verfügung stehenden externen Datenquellen, die in Tabelle 5.1 aufgeführt sind, in die Datenbank importiert. Da jede Datenquelle unterschiedliche technische Schnittstellen mit unter-

[32] `https://jupyter.org/`

schiedlichen Datenaustauschformaten, wie JSON, CSV oder XML, anbietet, wurden für jede Datenquelle python-basierte Importskripte entwickelt, mit denen die Daten in die Datenbank importiert worden sind. Diese reale und praxisrelevante Datenbasis bildet die Grundlage für die Feldexperimente zur Entwicklung der Komponenten des Unternehmen-Matcher.

6.3 Data Preparation - Unternehmensname

Der Unternehmensname ist ein entscheidendes Attribut für das RL der Realwelt-Entität Unternehmen. Eine zentrale Herausforderung für die Data Preparation des Unternehmensnamens stellt die Unternehmensrechtsform dar. Diese Herausforderung wurde in Abschnitt 5.4 beschrieben. In diesem Abschnitt werden die Experimente vorgestellt, die zur Lösung, dem RechtsformService, geführt haben.

Neben dem RechtsformService werden die Unternehmensnamen in der Data Preparation des Unternehmen-Matcher mit klassischen Verfahren aufbereitet. Die Groß- und Kleinschreibung wird behandelt, indem alle Unternehmensnamen klein geschrieben werden. Weiterhin werden Umlaute, wie ä, ö und ü, ersetzt durch ae, oe und ue. Zudem werden Sonderzeichen in den Unternehmensnamen entfernt.

Das Feldexperiment zur Entwicklung der RechtsformService Komponente, mit der die Unternehmensrechtsform klassifiziert und extrahiert wird, wurde durch den Autor bereits in der Publikation Kruse, Awick, Marx Gómez und Loos (2021) veröffentlicht. In der Publikation wird die für die Entwicklung verwendete Datengrundlage beschrieben, die einen Auszug der in Tabelle 5.1 aufgeführten Datenquellen darstellt. Für die Feldexperimente wurde die Komplexität reduziert, indem zunächst ausschließlich deutsche Unternehmensrechtsformen betrachtet worden sind. Die berücksichtigten deutschen Rechtsformen sind in Anhang C.1 aufgeführt. Bei der Auswahl des Ansatzes wurde die Erweiterbarkeit auf internationale Rechtsformen berücksichtigt. Insgesamt wurden die vier Ansätze (1) Bundesbank, (2) Cleanco, (3) Deep Learning und (4) Hybrid entwickelt, verglichen und evaluiert (vgl. Kruse, Awick et al., 2021).

Der (1) Bundesbank-Ansatz basiert auf der Publikation von (vgl. Schild & Schultz, 2017). Der Bundesbank-Ansatz verwendet reguläre Ausdrücke, die mit der Programmiersprache Perl implementiert wurden, um die Rechtsform zu klassifizieren. Für das Feldexperiment wurden diese regulären Ausdrücke mit Python umgesetzt und um die ausgewählten deutschen Rechtsformen (siehe Anhang C.1) erweitert. Der Bundesbank-Ansatz kann die Unternehmensrechtsform ausschließlich klassifizieren und nicht extrahieren. Diese Limitation und der benötigte Aufwand für die Entwicklung und Wartung von regulären Ausdrücken führte dazu, einen weiteren Ansatz zu untersuchen (vgl. Kruse, Awick et al., 2021).

Der nächste Ansatz der im Feldexperiment untersucht worden ist, ist der (2) Cleanco-Ansatz. Der Cleanco-Ansatz basiert auf dem gleichnamigen Github Projekt[33]. Cleanco ist ein regelbasierter Ansatz, der die Rechtsform klassifiziert und extrahiert. Der Ansatz hat im Standard nicht alle deutschen Rechtsformen implementiert. Die fehlenden Rechtsformen wurden nachträglich in das Cleanco Regelwerk integriert. Bei der Evaluation des Ansatzes stellte sich heraus, dass Rechtsformen die aus mehreren Token bestehen wie bspw. „GmbH &. Co. KG“, nicht klassifiziert und extrahiert werden. Die Änderung des Cleanco-Ansatzes um diese technische Limitation zu lösen, bedeutet großen Implementierungsaufwand weshalb ein weiterer Ansatz untersucht worden ist (vgl. Kruse, Awick et al., 2021).

Als Drittes wurde ein Deep Learning-Ansatz konzeptioniert, entwickelt und evaluiert. Der (3) Deep Learning-Ansatz hat das Problem der Rechtsform Klassifikation und Extraktion als Sequence Labeling Problem definiert. Sequence Labeling Probleme werden häufig durch Neuronale Netze mit der Architektur bi-directional LSTM (BI-LSTM) oder BI-LSTM mit Conditional Random Fields (CRF) gelöst. Für das Training der neuronalen Netze sind durch den Autoren dieser Arbeit 10.000 Unternehmensnamen gelabelt worden, um einen Trainings- und Testdatensatz zu erstellen. Für das Labeln der Trainingsdaten wurde das BIO-Tagging Schema ausgewählt. Bei diesem Schema wird jedes Token des Unternehmensnamens mit einem beginnenden Tag (B), einem inneren Tag (I) und einem außerhalb Tag (O) versehen, wie in Abbildung 6.4 für zwei Unternehmensnamen beispielhaft dargestellt.

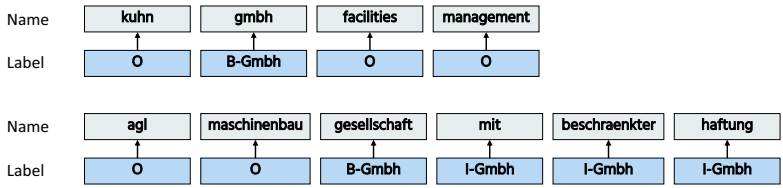


Abbildung 6.4: Aufbau der Trainingsdaten nach BIO-Tagging Schema für Sequence Labeling Problem

Für die Evaluation des Deep Learning-Ansatzes wurden weitere 3.733 Unternehmensnamen manuell gelabelt, um einen Evaluationsdatensatz mit Unternehmensnamen zu erstellen, die das trainierte Modell noch nie gesehen hat. Auf diesem Evaluationsdatensatz lieferte der Deep Learning-Ansatz mit 91,9% korrekt klassifizierter und 91,3% korrekt extrahierter Rechtsformen vielversprechende Ergebnisse. Allerdings hat der Ansatz Rechtsformen wie die „gGmbH“ und die „PartG“ nicht korrekt klassifiziert und klassifizierte diese oftmals falsch als „GmbH“. Zudem erfordert das Labeln der Unternehmensnamen für das Training der neuronalen Netze einen hohen manuellen Aufwand, da jedes Token im Unternehmensnamen

[33] `https://github.com/psolin/cleanco`

mit einem Label versehen werden muss. Die Schwierigkeiten bei der Klassifikation einiger Rechtsformen und der hohe manuelle Aufwand für die Erweiterung des Ansatzes um weitere Rechtsformen haben dazu geführt, dass ein weiterer Ansatz untersucht worden ist (vgl. Kruse, Awick et al., 2021).

Der vierte Ansatz ist der (4) Hybrid-Ansatz, der ebenfalls durch den Autor dieser Arbeit konzeptioniert, entwickelt und evaluiert wurde. Dieser Ansatz stellt den RechtsformService des Unternehmen-Matcher dar. Im Hybrid-Ansatz werden regelbasierte Ansätze mit Machine Learning kombiniert. Daher wurde die Klassifikation und Extraktion der Unternehmensrechtsform in die Teilaufgaben (1) Identifizierung der Rechtsform-relevanten Token, (2) Klassifikation der Rechtsform und (3) Extraktion der Rechtsform Token unterteilt (vgl. Kruse, Awick et al., 2021). Die Teilaufgaben sind in Abbildung 6.5 dargestellt und werden im Folgenden näher erläutert:

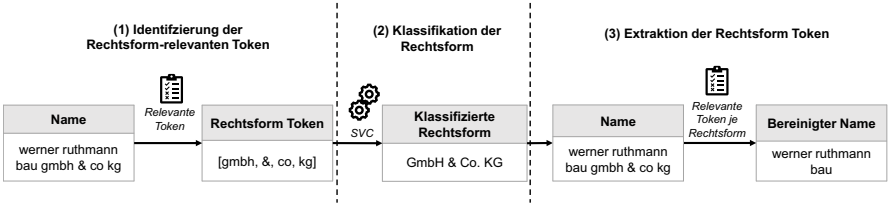


Abbildung 6.5: RechtsformService Hybrider-Ansatz

1. **Identifikation der Rechtsform-relevanten Token:** Zur Klassifikation der Rechtsform sind lediglich die Rechtsform Token im Unternehmensnamen entscheidend. Für den Unternehmensnamen „werner ruthmann bau gmbh & co kg“ , der in Abbildung 6.5 als Beispiel dient, sind dies die Token „gmbh“, „&“, „co“ und „kg“. Im RechtformService werden zunächst alle Rechtsform-relevanten Token, die in einer Liste gepflegt sind (siehe Listing 6.1), identifiziert. Anhand der ermittelten Rechtsform-relevanten Token wird dann im nächsten Schritt die Rechtsform bestimmt (vgl. Kruse, Awick et al., 2021).

```
list_legal_form_indicators = ['ag', 'aktiengesellschaft', 'aktien', 'mbh',
 'gmbh', 'beschraenkter', 'haftung', 'gesellschaftmbh', 'cogmbh', 'cokg',
 'ucokg', '&cokg', 'gmbhco', 'gmbh&co', 'gesmbh', 'gesellschaft',
 'gemeinnuetzige', 'ggmbh', 'mbh.', 'gmbh&cokg', 'gmbh&', 'gmbhcokg', 'gmbh
 &', 'ohg', ... ]
```

Listing 6.1: Auszug der Rechtsform-relevanten Token

2. **Klassifikation der Rechtsform:** Die Klassifikation der Rechtsform erfolgt nun auf Ba-

sis der extrahierten Token. Für die Erstellung eines Regelwerkes zur Rechtsform Klassifikation wird der Support Vector Classifier (SVC), ein Machine Learning-Verfahren, verwendet, da das manuelle Erstellen eines Regelwerkes zu kompliziert, zeitaufwändig und wartungsintensiv ist. Als Trainingsdaten dienten erneut die 10.000 gelabelten Datensätze, die lediglich in ihrem Aufbau verändert wurden. In Abbildung 6.6 ist der veränderte und benötigte Aufbau der Trainingsdaten dargestellt. Es wird lediglich ein Label für jeden Unternehmensnamen benötigt, während beim BIO-Tagging Schema ein Label für jedes Token des Unternehmensnamens benötigt wird (siehe Abb. 6.4). Die extrahierten Token wurden durch eine Multilabel-Binarization in einen numerischen Vektor transformiert, sodass ein Trainings- und Testdatensatz für den SVC generiert werden konnte. Der SVC erzielte eine Precision von 99,69%, einen Recall von 99,71% und damit einen F1-Score von 99,70%. Mit diesen Evaluationsmetriken war der SVC minimal besser als der zum Vergleich genutzte Random Forrest, sodass der SVC für den RechtsformService genutzt wird (vgl. Kruse, Awick et al., 2021).

Name	Label
kuhn gmbh facilities management	GmbH
agl maschinenbau mit beschraenkter haftung	GmbH

Abbildung 6.6: Aufbau der Trainingsdaten für Klassifikationsalgorithmus

3. **Extraktion der Rechtsform Token:** Im RechtsformService wird die klassifizierte Rechtsform, im Beispiel in Abbildung 6.5 die „Gmbh & Co. KG“, dazu verwendet, die Rechtsform-relevanten Token aus dem Unternehmensnamen zu extrahieren, um das gewünschte Ergebnis, den Namen ohne Rechtsform, zu erhalten. Im Beispiel in Abbildung 6.5 ist der bereinigte Name „werner ruthmann bau“ dargestellt. Die Extraktion der Rechtsform-relevanten Token erfolgt erneut über einen regelbasierten Ansatz. Für jede Rechtsform sind die möglichen Rechtsform Token in einer Liste gepflegt. Wenn die Rechtsform für den Unternehmensnamen klassifiziert wurde, werden alle in der Liste und zur jeweiligen Rechtsform gepflegten Token, die im zu bereinigenden Unternehmensnamen enthalten sind, entfernt (vgl. Kruse, Awick et al., 2021).

```
1  dict_legal_form_extraction = {'replace_ag': ['a.g.', 'ag',
   'aktiengesellschaft', 'aktien gesellschaft', 'a g', 'a. g.', 'a. g',
   'aktien', 'gesellschaft', 'ag.'], 'replace_gmbh': ['m.b.h.', 'g.m.b.h.',
   'gesellschaft', 'mit', 'beschraenkter', 'haftung', 'mbh', 'gmbh',
   'gesellschaftmbh'],
2    ... }
3
```

Listing 6.2: Auszug der Rechtsform-relevanten Token zur Rechtsformextraktion

Der Hybrid-Ansatz wurde ebenfalls mit dem Evaluationsdatensatz, der 3.733 Unternehmensnamen mit der Rechtsform als Label enthält, evaluiert. Auf Basis des Evaluationsdatensatzes klassifizierte der Hybrid-Ansatz 96,2% der Rechtsformen korrekt und extrahierte 91,6% der Rechtsform Token korrekt aus dem Unternehmensnamen. Die Erweiterung des Hybrid-Ansatzes ist mit weniger Aufwand verbunden als beim Deep Learning-Ansatz. Daher wurde der Hybrid-Ansatz für den RechtsformService gewählt und implementiert. Der RechtsformService wurde erweitert, sodass dieser in der aktuellen Version 57 verschiedene internationale Rechtsformen klassifizieren und extrahieren kann. In Tabelle 6.1 wird der Output des RechtsformService dargestellt. Dem RechtsformService wird der Unternehmensname (*company_name*) übergeben und als Rückgabewerte erhält man die Rechtsform (*legal_form*), den bereinigten Unternehmensname (*cleaned_name*) und die Rechtsform-relevanten Token (*indicators*) (siehe Tab. 6.1).

Tabelle 6.1: Beispielhafter Output des RechtsformService

company_name	**legal_form**	**cleaned_name**	**indicators**
Volkswagen Aktiengesellschaft	AG	volkswagen	[aktien, gesellschaft]
Volkswagen AG	AG	volkswagen	[ag]
CEWE Stiftung & Co KGaA	stiftung_co_kgaa	cewe	[kgaa, stiftung, co]
Unviersitaet Oldenburg	NaN	unviersitaet oldenburg	[]
Tesla Limited	Limited	tesla	[limited]

In Tabelle 6.1 ist zu sehen, dass der RechtsformService die verschiedenen Schreibweisen der Rechtsform Aktiengesellschaft klassifiziert, extrahiert und standardisiert. Mit dem Beispieldatensatz „Tesla Limited“ wird eine internationale Rechtsform dargestellt, um die der RechtsformService in dieser Forschungsarbeit u.a. erweitert worden ist.

6.4 Data Preparation - Adressdaten

Die Herausforderungen der Adressdaten für die Datenintegration sind in Abschnitt 5.4 bereits beschrieben worden. In diesem Abschnitt wird die entwickelte Lösung, der AdressService, vorgestellt.

Die uneinheitliche Strukturierung der Adressdaten erschwert das Schema Matching. Daher ist das Ziel des AdressService, die unterschiedlich strukturierten Adressdaten in eine einheitliche Struktur zu überführen. In Abbildung 6.7 ist die für den Unternehmen-Matcher vorgesehene Ziel-Datenstruktur der Adressdaten abgebildet.

Die Adresse kann als Text- oder Geokoordinatendarstellung repräsentiert werden. Die Textdarstellung enthält fünf Detailebenen wobei Detailebene eins eine Land Angabe enthält und die ungenaueste Adressangabe darstellt. Mit jeder Detailebene wird die Adressangabe

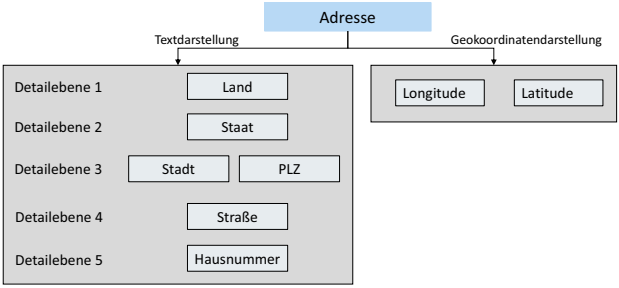


Abbildung 6.7: Ziel-Datenstruktur der Adresse

präziser. Detailebene zwei enthält zusätzlich den Staat. Detailebene drei enthält zusätzlich die Stadt oder die Postleitzahl (PLZ). Detailebene vier enthält die Straße und Detailebene fünf die Hausnummer. Die Textdarstellung der Adresse kann für jede Detailebene auch über Geokoordinaten in Form von Longitude und Latitude dargestellt werden.

In allen Datenquellen, die in der Experimentumgebung vorliegen (siehe Tab. 5.1), liegen die Adressdaten in Textdarstellung vor. Für eine hohe RL-Ergebnisqualität benötigt der Unternehmen-Matcher eine normalisierte, möglichst vollständige und einheitlich dargestellte Adresse für den Vergleich der Adressdaten in den RL-Prozessschritten Comparison und Classification. Unter normalisiert wird verstanden, dass die Adressdaten atomar in den Attributen Land, Staat, Stadt, PLZ, Straße und Hausnummer vorliegen. Die Analyse der Datenquellen für die Feldexperimente hat gezeigt, dass die Adressdaten normalisiert, teilweise normalisiert sowie nicht normalisiert vorliegen. Die Adressdaten liegen teilweise normalisiert vor, wenn Attribute mit und ohne atomare Werbebereiche vorliegen. Die Adressdaten werden als nicht normalisiert bezeichnet, wenn kein Attribut einen atomaren Wertebereich besitzt. Unter der einheitlichen Darstellung der Adressdaten wird verstanden, dass Adressdaten wie das Land bspw. einheitlich ausgeschrieben (Deutschland), im ISO-3166-1 ALPHA-2 Format (DE) oder im ISO-3166-1 ALPHA-3 Format (DEU) vorliegen. Herausforderungen der nicht einheitlichen Darstellung existieren für die Adressdaten-Attribute jeder Detailebene. Für das RL ist eine vollständige Adresse, ohne fehlenden Angaben, der optimale Zustand. Die Analyse der vorliegenden Datenquellen zeigte, dass die Adressdaten sich in ihrer Vollständigkeit unterscheiden. Daher soll der AdressService die Adressdaten, sofern möglich, um fehlende Angaben anreichern.

Um die genannten Herausforderungen und Probleme mit den Adressdaten zu lösen, wurden die beiden Frameworks Libpostal[34] und Nominatim[35] identifiziert, die im AdressService

[34] `https://github.com/openvenues/libpostal`
[35] `https://github.com/osm-search/Nominatim`

genutzt werden. Beide Frameworks bieten einen großen Funktionsumfang und stehen als Open-source Software zur Verfügung, weshalb diese in dieser Arbeit verwendet werden.

Libpostal: Libpostal ist ein in der Programmiersprache C entwickeltes Framework zur Normalisierung von weltweiten Adressdaten. Das Framework nutzt Natural Language Processing (NLP) Methoden und verwendet OpenStreetMap Daten. Das Framework ist unter der MIT Lizenz als Open-source veröffentlicht. In dieser Arbeit wird das auf Libpostal aufbauende PyPostal[36] Framework, welches in Python entwickelt ist, verwendet. In Listing 6.3 sind Beispiele für die Verwendung des PyPostal Frameworks dargestellt.

```
from postal.parser import parse_address

parse_address('Ammerlaender Heerstrasse 114-118 26129 Oldenburg DEU')
# Out:
[('ammerlaender heerstrasse', 'road'),
('114-118', 'house_number'),
('26129', 'postcode'),
('oldenburg', 'city'),
('deu', 'country')]

parse_address('Kreuzberg, Friedrichshain-Kreuzberg, Berlin, Germany')
# Out:
[('kreuzberg', 'suburb'),
('friedrichshain-kreuzberg', 'city_district'),
('berlin', 'city'),
('germany', 'country')]

parse_address('Oldenburg DE')
# Out:
[('oldenburg', 'city'), ('de', 'country')]

parse_address('198 CHAMPION COURT SAN JOSE CA 95134 USA')
# Out:
[('198', 'house_number'),
('champion court', 'road'),
('san jose', 'city'),
('ca', 'state'),
('95134', 'postcode'),
('usa', 'country')]
```

Listing 6.3: Beispiele für die Anwendung des PyPostal Framework

[36] https://github.com/openvenues/pypostal

In Listing 6.3 ist in Zeile 3 die Funktion zum Aufruf für die Normalisierung der Adresse dargestellt. Mit PyPostal werden die einzelnen Adressangaben mit Labeln versehen, wie in Zeile 5-9 für das erste Beispiel dargestellt. Das Beispiel in Zeile 22-29 zeigt, dass PyPostal internationale Adressen, wie im dargestellten Fall eine amerikanische Adresse, normalisieren kann. Für die Normalisierung der Adresse verwendet Libpostal das Conditional Random Fields (CRF) Modell, das auf einer Trainingsdatenmenge von einer Milliarde weltweiter Adressdatensätze trainiert worden ist. Neben dem lateinischen Alphabet kann PyPostal auch Adressen, die über das arabische oder chinesische Alphabet repräsentiert sind, verarbeiten.

Nominatim: Nominatim ist ein unter der GPL-2.0 Lizenz veröffentlichtes Open-source Framework zur Geokodierung von Adressdaten, die in Text- oder Geokoordinatendarstellung vorliegen. Nominatim basiert auf OpenStreetMap Daten. Über die Nominatim Website ist die Benutzeroberfläche zum Testen des Frameworks zu finden[37]. Die Benutzeroberfläche zeigt die beiden Anfrage-Möglichkeiten für Adressdaten, (1) die Simple-Anfrage und (2) die Strukturierte-Anfrage. Beide Anfrage-Möglichkeiten können auch technisch über eine API genutzt werden. Bei der API Funktion Simple-Anfrage werden die Adressdaten als einzelner String übergeben, wie bspw. „Ammerländer Heerstraße 118 26129 Oldenburg DEU“. Bei der API Funktion Strukturierte-Anfrage wird die Adresse strukturiert mit den Variablen *housenumber_street*, *city*, *county*, *state*, *country* und *postal_code* übergeben. Als Rückgabe liefert die Nominatim API die Adressdaten im JSON-Format. Mit Hilfe von Nominatim können Adressdaten vereinheitlicht und angereichert werden. Durch Nominatim kann bspw. die Land Angabe einheitlich ausgeschrieben und im ISO-3166-1 ALPHA-2 Format dargestellt werden. Zusätzlich können dem Nominatim Service Adressdaten in einer Vielzahl von Sprachen übergeben werden und auf eine festgelegte Sprache, wie Englisch oder Deutsch, vereinheitlicht zurückgegeben werden. Bspw. enthält Nominatim für die Stadt Berlin 193 verschiedene Übersetzungen [38].

Damit das Nominatim Framework für den Unternehmen-Matcher genutzt werden kann, wurde ein Nominatim-Server in der vorhandenen IT-Infrastruktur für die Experimente eingerichtet. Das Einrichten eines eigenen Nominatim-Server wird für Power-User empfohlen und erfolgte nach der Anleitung, die auf der Nominatim Website hinterlegt ist[39]. Zudem wurde ein Python-basiertes Programm entwickelt, das die API Funktionalitäten des Nominatim-Server kapselt. Die Funktionalitäten mit beispielhaftem Input und Output sind in Listing 6.4 dargestellt.

[37] `https://nominatim.openstreetmap.org/ui/search.html`
[38] `https://nominatim.openstreetmap.org/ui/details.html?osmtype=N&osmid=240109189&class=place`
[39] `https://nominatim.org/release-docs/latest/admin/Installation/`

```
##### Simple-Anfrage ######
enrich_address_data_nominatim("Ammerlaender Heerstrasse 118 26129 Oldenburg DE")

# Out:
[{'place_id': 157162574,
'lat': '53.148925399999996',
'lon': '8.182053307335966',
'address': {'building': 'V01 - Verwaltung',
'house_number': '114-118',
'road': 'Ammerlaender Heerstrasse',
'suburb': 'Haarentor',
'city': 'Oldenburg',
'state': 'Lower Saxony',
'postcode': '26129',
'country': 'Germany',
'country_code': 'de'}}]

##### Strukturierte-Anfrage ######
enrich_address_data_nominatim_detail(house_street="118 Ammerlaender
    Heerstrasse", city="Oldenburg", country="Deutschland")

# Out:
[{'place_id': 157162574,
'lat': '53.148925399999996',
'lon': '8.182053307335966',
'address': {'building': 'V01 - Verwaltung',
'house_number': '114-118',
'road': 'Ammerlaender Heerstrasse',
'suburb': 'Haarentor',
'city': 'Oldenburg',
'state': 'Lower Saxony',
'postcode': '26129',
'country': 'Germany',
'country_code': 'de'}}]

```

Listing 6.4: Beispiel für die Anwendung des Nominatim Framework

Die Funktion ENRICH_ADDRESS_DATA_NOMINATIM kapselt die Simple-API-Anfrage. Der Funktion wird die Adresse als ein String Attribut übergeben und es werden sofern vorhanden die Attribute *lat*, *lon*, *house_number*, *road*, *suburb*, *city*, *state*, *postcode*, *country* und *country_code* zurückgeliefert (siehe Zeile 5-16 Listing 6.4). Das Beispiel zeigt, wie

die übergebene Adresse um Informationen wie bspw. die PLZ, die Latitude und die Longitude angereichert wird. Weiterhin wird der zweistellige Ländercode zusätzlich in die ausgeschriebene Darstellung der Land Angabe, siehe Zeile 15-16, überführt. Die API-Funktionalitäten wurden mit dem Parameter versehen, dass die Rückgabewerte in englischer Sprache zurückgeliefert werden.

Nachdem die beiden Teilkomponenten PyPostal und Nominatim des entwickelten AdressService vorgestellt worden sind, wird nun die gesamte Funktionalität des AdressService vorgestellt. In Abbildung 6.8 ist die Funktionalität des AdressService abgebildet.

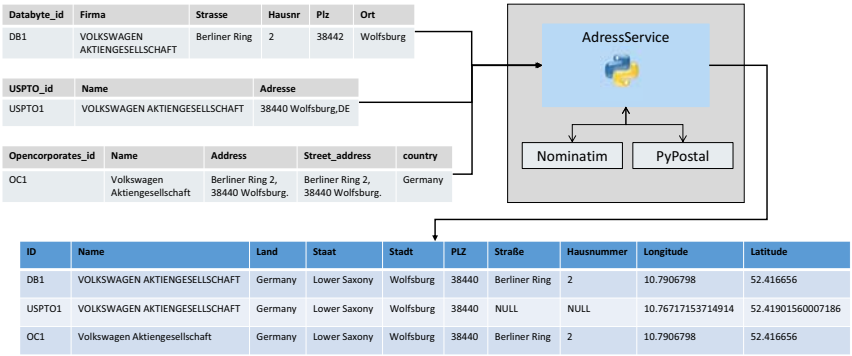


Abbildung 6.8: Funktionsweise des AdressService

Dem AdressService können beliebig strukturierte Adressdaten übergeben werden. Wie in Abbildung 6.8 dargestellt wird ein Adressdatensatz aus der Databyte Datenquelle, ein Datensatz aus der USPTO Datenquelle und ein Datensatz aus der OpenCorporates Datenquelle übergeben. Alle drei Datensätze enthalten unterschiedliche Attributsbezeichnungen, sind unterschiedlich strukturiert, normalisiert und uneinheitlich dargestellt. Die Logik des AdressService normalisiert die Adressdatensätze mit Hilfe von PyPostal. Anschließend erfolgt ein Mapping der Attribute auf die benötigten Parameter für das Nominatim Framework. Durch das Nominatim Framework werden die normalisierten Adressdatensätze vereinheitlicht und angereichert. Aufgrund von mangelnder Datenqualität der Adressdatensätze wie fehlenden oder falsche Werte, kann es vorkommen, dass das Nominatim Framework keine Rückgabewerte liefert. Für diese Fälle wurde eine Logik im AdressService entwickelt, die schrittweise unterschiedliche Adressattribute aus der API-Anfrage herauslässt, um die bestmöglichen Rückgabewerte zu erhalten. In Abbildung 6.8 sind die drei Datensätze nach der Data Preparation durch den AdressService abgebildet. Die drei Datensätze wurden ver-

einheitlicht und in die Ziel-Datenstruktur überführt. Zusätzlich wurden Attribute wie die Longitude und Latitude angereichert. Durch die Data Preparation der Adressdaten mit dem entwickelten AdressService können die Algorithmen der darauffolgenden RL-Prozessschritte mit der Ziel-Datenstruktur weiterarbeiten.

6.5 Blocking Verfahren

In diesem Abschnitt werden die für den Unternehmen-Matcher verwendeten Algorithmen für den RL-Prozessschritt Blocking beschrieben. Der RL-Prozessschritt Blocking bildet in der RL-Forschung einen eigenen Forschungsbereich. In der Publikation von Papadakis, Skoutas et al. (2020) wird ein umfangreicher Überblick über den aktuellen Stand der Forschung im Bereich Blocking gegeben. Da in dieser Arbeit ein ganzheitlicher RL-Prozess für die Realwelt-Entität Unternehmen entwickelt wird, basiert die Auswahl der Blocking-Verfahren auf den Ergebnissen von Papadakis, Skoutas et al. (2020). In ihrer Publikationen beschreiben sie, dass das Standard-Blocking und das Sorted Neighbourhood Blocking die am meisten verwendeten und erfolgreichsten Verfahren sind (vgl. Papadakis, Skoutas et al., 2020). Bestätigt wird dies ebenfalls durch die Publikation von l. Koumarelas et al. (2020). Daher werden für den Unternehmen-Matcher zunächst diese beiden Verfahren implementiert und untersucht.

Um die beiden Verfahren zu untersuchen, wurden zunächst vier Feldexperimente erstellt. Die vier Feldexperiment mit den ID's 1, 2, 3 und 4 sind in Tabelle 6.2 dargestellt. Die ersten vier Experimente wurden mit einer Datenquellengröße durchgeführt, sodass die Anzahl aller vorhandenen korrekten Matches manuell bestimmt werden konnte und somit Ground-Truth Daten vorliegen. Für das Experiment mit der ID 1 werden die AIStartups und die Crunchbase Datenquelle genutzt, zwischen denen 212 korrekte Matches manuell identifiziert wurden. Für das Experiment mit der ID 2 wird ein Auszug der Crunchbase Datenquelle und der Wikidata Datenquelle verwendet, zwischen denen 3.895 korrekte Matches manuell identifiziert wurden. Für das Experiment mit der der ID 3 wird die CompanyList Datenquelle und die vollständige Wikidata Datenquelle verwendet, zwischen denen 82 korrekte Matches manuell identifiziert wurden. Das Experiment mit der ID 4 wird mit einem Auszug der Wikidata Datenquelle und der OpenCorporates Datenquelle durchgeführt, zwischen denen 2.787 korrekte Matches manuell identifiziert wurden.

Tabelle 6.2: Feldexperiment Blocking Verfahren
PC = Pair Quality; RR = Reduction Ratio

ID	Datenquellen	Anz. Ds.	korrekte Matches	Kreuzprodukt	Standard Blocking			Sorted Neighbourhood		
					Tupel	PC	RR	Tupel	PC	RR
1	AIStartups Crunchbase	279 589.343	212	164.426.697	8.871.906	**99,06%**	94,60%	3.683	98,58%	**99,998%**
2	Crunchbase Wikidata	4.434 4.177	3895	18.520.818	813.885	**99,77%**	95,61%	64.701	99,67%	**99,65%**
3	CompanyList Wikidata	84 534.609	82	44.907.156	2.347.413	**96,34%**	94,77%	3.734	**96,34%**	**99,992%**
4	Wikidata OpenCorporates	2.758 2.793	2787	7.703.094	392.484	**90,38%**	94,90%	28.152	89,09%	**99,63%**

Der Unternehmensname ist das häufigst vorhandene Attribut, sodass der Unternehmensname, nach der Data Preparation durch den RechtsformService, als Blocking Key herangezogen wird. Für das implementierte Standard Blocking Verfahren wird der erste Buchstabe des aufbereiteten Unternehmensnamens verwendet. Das Standard Blocking über den ersten Buchstaben des Unternehmensnamens wurde über einen SQL-Join implementiert, wie in Listing 6.5 exemplarisch für Experiment 4 dargestellt.

```
1 -- Standard Blocking
2 SELECT *
3 FROM exp_4_opencorporates_wikidata.oc_legal_form t
4 join exp_4_opencorporates_wikidata.wikidata_legal_form t2
5 on (substring(t.cleaned_name,1,1) = substring(t2.cleaned_name,1,1))
```

Listing 6.5: SQL Implementierung des Standard Blocking erster Buchstabe des Unternehmensnamens

Für das Sorted Neighbourhood Blocking Verfahren wurde auf die Implementierung des Algorithmus aus dem Python Record Linkage-Toolkit[40] zurückgegriffen. Das Python Record Linkage-Toolkit ist ein unter der BSD-3-Clause lizensiertes Open-source Framework. In Listing 6.6 ist die Implementierung des Sorted Neighbourhood Blocking mit dem Algorithmus aus dem Python Record Linkage-Toolkit dargestellt. Als Blocking Key wird der aufbereitete Unternehmensname verwendet, siehe Zeile vier. Der Sorted Neighbourhood Algorithmus benötigt als Parameter die Window-Size. Diese wurde auf dreizehn gesetzt, wie in Zeile vier in Listing 6.6 zu sehen. Anschließend wird der Algorithmus in Zeile fünf ausgeführt und in Zeile sechs werden die generierten Tupel in einen DataFrame überführt.

```
1 #### Sorted Neighbourhood ####
2 import recordlinkage as rl
3
4 indexer = rl.SortedNeighbourhoodIndex(left_on='df_a_company_name_prep',
     right_on='df_b_company_name_prep', window=13)
5 record_pairs = indexer.index(df_a, df_b)
6 df_record_pairs = record_pairs.to_frame()
```

Listing 6.6: Python Implementierung des Sorted Neighbourhood Blocking

Durch die Anzahl der korrekten Matches können die Blocking-Verfahren Standard Blocking und Sorted Neighbourhood über die Kennzahl Pair Completeness (PC) und Reduction Ratio (RR) verglichen und evaluiert werden (siehe Tab. 6.2).

In Experiment eins erzielte das SB eine PC von 99,06% und ist damit 0,48% besser als das SN.

[40] https://github.com/J535D165/recordlinkage

Das SB erzielt eine RR von 94,60% gegenüber dem Kreuzprodukt der beiden Datensätze. Mit einer RR von 99,998% ist das SN rund 5,4% besser als das SB. In Experiment zwei erzielte das SB eine PC von 99,77% und ist damit 0,1% besser als das SN. Das SB erzielte eine RR von 95,61% gegenüber dem Kreuzprodukt der beiden Datensätze. Mit einer RR von 99,65% ist das SN rund 4% besser als das SB. In Experiment drei erzielte das SB eine PC von 96,34%. Das SN erzielte ebenfalls eine PC von 96,34%. Mit einer RR von 99,992% ist das SN rund 5,2% besser als die RR des SB. In Experiment vier erzielte das SB eine PC von 90,38% und ist damit 1,3% besser als die PC des SN. Die RR des SN ist rund 4,7% besser als die RR des SB.

Insgesamt ist die PC des SN minimal niedriger über alle vier Experimente als die PC des SB, was bedeutet, dass einige korrekte Matches durch das SN eliminiert werden. Allerdings ist die RR des SN über alle vier Experiment deutlich höher als die RR des SB, weshalb viele FP Tupel eliminiert werden. Daher wird für den weiteren Verlauf der Arbeit das Sorted Neighbourhood Verfahren für den Unternehmen-Matcher eingesetzt.

6.6 Comparison

In diesem Abschnitt werden die für den Unternehmen-Matcher verwendeten Algorithmen und Verfahren für den RL-Prozessschritt Comparison beschrieben. In Abbildung 6.9 ist exemplarisch dargestellt, wie die Tupel nach der Verarbeitung des Unternehmen-Matcher durch die RL-Prozessschritte Data Preparation und Blocking an den Prozessschritt Comparison übergeben werden. Als Attribute für den Vergleich der Tupel werden die Attribute bereinigter *Name*, *Rechtsform*, *Stadt*, *PLZ*, *Straße*, *Hausnummer*, *Land*, *Ländercode*, *Latitude* und *Longitude* für beide Datensätze übergeben. Das Tupel bestehend aus Datensatz A und Datensatz B in Abbildung 6.9 stellt ein Match dar. Zusätzlich sind in Abbildung 6.9 die für die vorhandenen Attribute verwendeten Ähnlichkeitsmaße dargestellt.

Die Berechnung der Ähnlichkeit zwischen den bereinigten Unternehmensnamen erfolgt mit den Ähnlichkeitsmaßen Levenshtein, Jaro-Winkler und Monge-Elkan. Diese Ähnlichkeitsmaße wurden ausgewählt, da mit der Levenshtein-Distanz minimale Zeichenunterschiede wie bspw. vertauschte Buchstaben, erkannt werden können. Durch die Jaro-Winkler-Distanz wird der Anfang des Unternehmensnamens höher gewichtet und mögliche fehlende Ergänzungen am Ende des Unternehmensnamens fallen nicht so hoch ins Gewicht, wie es bei der Levenshtein-Distanz passieren würde. Mit der Monge-Elkan-Distanz sollen neben Zeichenunterschieden auch die unterschiedliche Reihenfolge der Token im Unternehmensnamen berücksichtigt werden.

Die Ähnlichkeitsmaße wurden für den Unternehmen-Matcher auf Basis des py-stringmat-

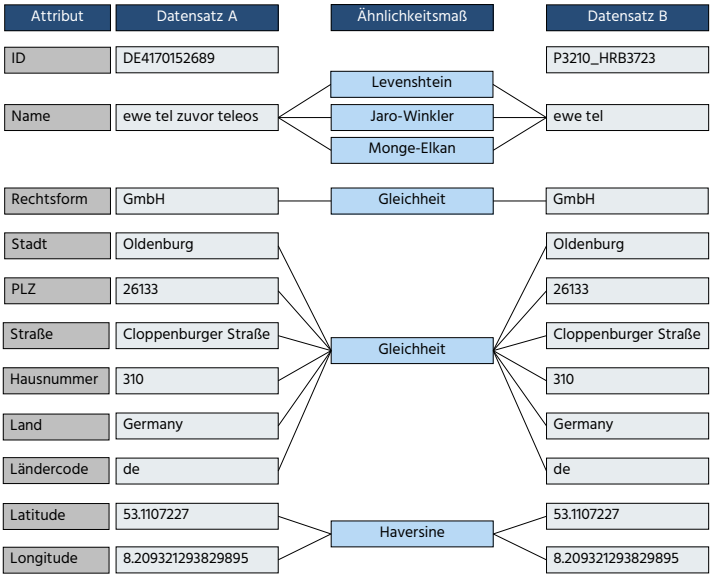


Abbildung 6.9: Verwendete Ähnlichkeitsmaße im Prozessschritt Comparison

ching[41] Frameworks aus dem Magellan Projekt implementiert. Die Python Implementierung der Ähnlichkeitsmaße ist in Listing 6.7 dargestellt. Von Zeile vier bis Zeile neun ist die Funktion für die Berechnung der Levenshtein-Distanz dargestellt. Von Zeile elf bis Zeile 17 ist die Funktion zur Berechnung der Jaro-Winkler-Distanz dargestellt. Von Zeile 19 bis Zeile 28 ist die Funktion zur Berechnung der Monge-Elkan Distanz dargestellt. Als Tokenizer wird in der für den Unternehmen-Matcher implementierten Monge-Elkan-Distanz der *WhitespaceTokenizer* verwendet, siehe Zeile 27. Die generierten Token werden in der hier implementierten Monge-Elkan-Distanz mit der Levenshtein-Distanz verglichen, siehe Zeile 22 und 28.

```
1 import py_stringmatching as sm
2
3 ########## Levenshtein ############
4 def levenshtein_norm(token1,token2):
5    lev = sm.Levenshtein()
6    if (token1 == None or token2 == None):
7       return None
8    else:
```

[41] https://anhaidgroup.github.io/py_stringmatching/v0.4.2/index.html

```
9       return lev.get_sim_score(str(token1),str(token2))
10
11  ########## JaroWinkler ############
12  def jarowinkler_raw(token1,token2):
13     jarowinkler = sm.JaroWinkler()
14     if (token1 == None or token2 == None):
15        return None
16     else:
17        return jarowinkler.get_raw_score(str(token1),str(token2))
18
19  ########## Monge Elkan ############
20  # Inner Similarity Measure is Levenshtein
21  def monge_elkan_ws_ld(token1,token2):
22     me = sm.MongeElkan(sim_func=sm.Levenshtein().get_sim_score)
23     if (token1 == None or token2 == None):
24        return None
25     else:
26     # create a whitespace tokenizer that returns a set of tokens
27        ws_tok_set = sm.WhitespaceTokenizer(return_set=True)
28        return me.get_raw_score(ws_tok_set.tokenize(str(token1)),
       ws_tok_set.tokenize(str(token2)))
29
30  ############ Token Equal Check #########
31  def check_token_equal(token1,token2):
32     if (token1 == None or token2 == None):
33        return None
34     else:
35        if (token1 == token2):
36           return 1
37        else:
38           return 0
```

Listing 6.7: Python Implementierung der Ähnlichkeitsmaße

Die Rechtsform wird auf Gleichheit geprüft, siehe Abbildung 6.9. Da die Rechtsform durch den RechtsformService in der Data Preparation bereits in eine einheitlich Darstellung transformiert worden ist, wird für die Rechtsform kein anderes Ähnlichkeitsmaß benötigt. Die Python Implementierung der Prüfung auf Gleichheit ist in Listing 6.7 in Zeile 30-38 dargestellt.

Die Attribute *Stadt*, *PLZ*, *Straße*, *Hausnummer*, *Land* und *Ländercode* sind die durch den AdressService aufbereitete Textdarstellung der Adresse, die daher über die Prüfung auf Gleichheit verglichen werden. Die Geokoordinatendarstellung wird über die Attribute La-

titude und Longitude über die Haversine-Distanz verglichen (siehe Abb. 6.9). Die Python Implementierung der Haversine-Distanz für den UNTERNEHMEN-MATCHER basiert auf dem gleichnamigen Python Framework Haversine[42]. Die Implementierung der Funktion ist in Listing 6.8 dargestellt. Der Funktion werden die Latitude und Longitude der beiden Geokoordinaten übergeben und die Distanz wird in Kilometern zurückgegeben, wie in Zeile 16 zu sehen.

```
1  from haversine import haversine, Unit
2  import numpy as np
3
4  def compute_distance_lat_lon(lat1,lng1,lat2,lng2):
5     try:
6        ## city1
7        f_lat1= float(lat1)
8        f_lng1= float(lng1)
9        ## city2
10       f_lat2= float(lat2)
11       f_lng2= float(lng2)
12
13       city1 = (f_lat1,f_lng1)
14       city2 = (f_lat2,f_lng2)
15
16       distance = haversine(city1, city2, unit=Unit.KILOMETERS)
17       return distance
18    except:
19       return np.nan
```

Listing 6.8: Python Implementierung der Haversine-Distanz

Durch die Ähnlichkeitsberechnung der Adresse über die Geokoordinatendarstellung kann u.a. das Problem des Detailgrad in den Adressdaten, das in Abschnitt 5.4 beschrieben ist, gelöst werden. Die Städte *Garching bei München* und *München* über String-Ähnlichkeitsmaße oder die Prüfung auf Gleichheit als sehr ähnlich zu deklarieren ist schwierig. Daher wird für den besseren Vergleich die Distanz in Kilometern genutzt, die in diesem Beispiel ca. 13km beträgt. Ein weiteres Beispiel ist der Umzug eines Unternehmens in eine andere Straße. Während die Straßennamen komplett verschieden sein können, kann die Kilometer-Distanz, wie im Fall des Umzugs eines Unternehmens von der Kurwickstraße in Oldenburg in die Mottenstraße in Oldenburg wenige 100 Meter betragen. Im UNTERNEHMEN-MATCHER wird die Haversine-Distanz zum Vergleich genutzt, wenn die Adressdaten der zu vergleichenden Datensätze die gleiche Detailebene besitzen. Vor allem bei nicht vollständig vorhandenen Adressdaten kann

[42] https://pypi.org/project/haversine/

die Haversine-Distanz zu einer fehlerhaften Ähnlichkeit führen. Existiert für ein in München ansässiges Unternehmen in Datenquelle A bspw. nur die Land Angabe Deutschland und in Datenquelle B die vollständige Adresse, liegt die Haversine-Distanz bei ca. 340 Kilometer. Für die Land Angabe Deutschland liefert der AdressService die Geokoordinaten des Mittelpunktes von Deutschland. Dadurch würde ein Unternehmen aus Erfurt fälschlicherweise näher am Unternehmen aus Datenquelle A liegen, als der korrekte Match aus Datenquelle B. Daher muss die Detailebene bei der Ähnlichkeitsberechnung der Adresse über die Haversine-Distanz berücksichtigt werden.

Das Ergebnis des Unternehmen-Matcher nach dem RL-Prozessschritt Comparison ist in Tabelle 6.3 exemplarisch dargestellt. In Tabelle Tabelle 6.3 sind zehn Tupel dargestellt. Dem Datensatz mit der ID_A sind zehn verschiedene Datensätze zugeordnet. Die Ähnlichkeit zwischen den Datensätzen wird über die berechneten Ähnlichkeitsmaße dargestellt. Der Ländercode steht exemplarisch für die gesamte Adresse in Textdarstellung, die auf Gleichheit geprüft wird. Auf Basis dieser Ähnlichkeitswerte, die für jedes Tupel einen Vektor darstellen, muss im folgenden RL-Prozssschritt Classification entschieden werden, ob es sich um ein Match oder um Kein-Match handelt.

Tabelle 6.3: Ergebnis des RL-Prozessschrittes Comparison

ID_A	ID_B	Levenshtein	Jaro-Winkler	Monge Elkan	Rechtsform	Haversine	Ländercode
1	37744	0.29	0.66	0.34	0.0	113.03	1.0
1	10773	0.28	0.73	0.50	1.0	46.72	1.0
1	26216	0.26	0.75	0.35	1.0	216.80	1.0
1	38339	0.25	0.60	0.26	1.0	131.09	1.0
1	61527	0.19	0.61	0.27	Null	57.51	1.0
1	20969	0.32	0.68	0.35	1.0	45.47	1.0
1	3834	0.25	0.66	0.33	0.0	201.94	1.0
1	37728	0.19	0.71	0.35	1.0	113.03	1.0
1	54238	0.25	0.61	0.23	1.0	178.76	1.0
1	19152	0.19	0.62	0.31	1.0	74.52	1.0

6.7 Classification

In diesem Abschnitt werden die für den Unternehmen-Matcher verwendeten Algorithmen und Verfahren für den RL-Prozessschritt Classification beschrieben. Durch die implementierten Algorithmen für die RL-Prozessschritte Data Preparation und Comparison des Unternehmen-Matcher stehen für die Classification immer die Attribute aus Tabelle 6.3 als Input bereit. Wie bereits in Abschnitt 5.4 beschrieben, ist die Bewertung der Ergebnisqualität eine Herausforderung bei der Datenintegration, da diese oftmals einen großen manuellen Aufwand erfordert. Für die finale Bewertung der klassifizierten Matches in TP und FP

müssen alle Matches manuell geprüft werden. Dies bedeutet in der Praxis einen hohen manuelle Aufwand, der selten ökonomisch ist, da oftmals mehr als eine Datenquelle integriert wird und mit jeder zu integrierenden Datenquelle der Prüfaufwand steigt. Daher ist das Ziel dieser Arbeit für den Unternehmen-Matcher einen Algorithmus für den RL-Prozessschritt Classification zu entwickeln, bei dem nicht alle Matches manuell geprüft werden müssen. Weiterhin ist das Ziel dieser Arbeit, einen Algorithmus für die Classification zu entwickeln, der Datenquellen-unabhängig ist und damit auf neue zu integrierende Datenquellen übertragen werden kann.

Damit dieses Ziel erreicht wird, werden die Matches aus der Klassifikation nicht mehr ausschließlich in TP und FP unterteilt sondern das Ziel ist es, diese in sichere-Matches und potenzielle-Matches zu unterteilen (siehe Abb. 6.10).

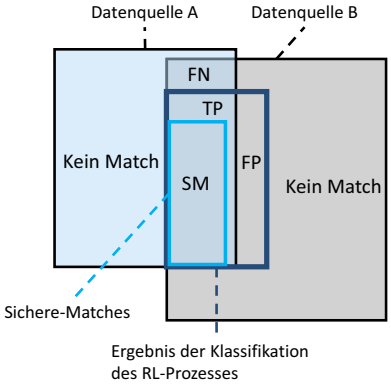


Abbildung 6.10: Lösung der Bewertung der Ergebnisqualität beim Record Linkage

Sichere-Matches: Ein sicherer-Match soll ein TP Datensatz sein, sodass dieser nicht manuell überprüft werden muss. Die sicheren-Matches bilden einen Teil der TP aus der gesamten Menge der Matches, die aus den TP und FP besteht (siehe Abb. 6.10). Der Klassifikationsalgorithmus soll möglichst viele TP als sichere-Matches generieren, um den manuelle Aufwand zu reduzieren.

Potenzielle-Matches Die potenziellen-Matches sind die Matches abzüglich der sicheren-Matches. Daher enthalten die potenziellen-Matches TP und FP Datensatzpaare, weshalb eine manuelle Prüfung erforderlich ist. Dennoch sollten möglichst wenige FP als potenzielle-Matches generiert werde, um den manuellen Prüfaufwand zu reduzieren.

Wie das Literaturreview zum aktuellen Stand der Technik in Abschnitt 3.2.8 zeigt, exis-

tieren mit regelbasierten Ansätzen, supervised und unsupervised Learning Verfahren eine Vielzahl an unterschiedlichen Algorithmen und Verfahren für die Klassifikation der Tupel in Match und Kein-Match bzw. für den Unternehmen-Matcher in sicheres-Match und potenzielles-Match. Supervised Verfahren haben den Nachteil, dass für das Training gelabelte Daten benötigt werden. Auch das Erstellen von gelabelten Trainingsdaten erfordert manuellen Aufwand, der vermieden werden soll. Zudem fällt dieser manuelle Aufwand für jede neue zu integrierende Datenquelle an. Auch die Autoren Gschwind et al. (2019), Köpcke (2014) und Christen (2012a) sehen den hohen manuellen Aufand und die damit verbundenen Kosten für den Einsatz von supervised Learning Verfahren im RL-Prozessschritt Classification als Herausforderung an. Daher soll in dieser Arbeit für den Unternehmen-Matcher ein parametrisiertes Regelwerk entwickelt werden, was auch im Literaturreview, siehe Abschnitt 3.2.8, als am häufigsten verwendetes Verfahren identifiziert wurde.

Das Regelwerk des Unternehmen-Matcher soll (1) die Datensatzpaare in sichere-Matches und potenzielle-Matches unterteilen sowie (2) Datenquellen-unabhängig für die Realwelt-Entität Unternehmen anwendbar sein. Um ein Datenquellen-unabhängiges Regelwerk zu entwickeln, wurden mehrere Feldexperimente iterativ durchgeführt, bis keine bestehende Regel mehr verändert wurde und keine Regeln hinzugefügt oder entfernt wurde.

Für die Feldexperimente zur Entwicklung des Regelwerkes wurden zunächst die vier Experimente, die bereits für die Entwicklung des Blocking-Verfahrens genutzt wurden, verwendet (siehe Tab. 6.2). Mit diesen vier Experimenten wurden die ersten Iterationen zur Entwicklung des Regelwerkes durchgeführt. Im Anschluss wurden fünf weitere Experimente durchgeführt, bis das Regelwerk nicht mehr verändert werden musste. Die insgesamt neun durchgeführten Experimente sind Tabelle 6.4 aufgeführt.

Die Experimente eins und drei enthalten jeweils eine Datenquelle mit wenigen Datensätzen, jeweils 279 und 84. Die Experimente zwei und vier bestehen aus einem Auszug der Datenquellen Crunchbase und Wikidata sowie Wikidata und OpenCorporates. Daher wurden für diese vier Experimente Ground-Truth Daten manuell erstellt, sodass die gesamte Anzahl korrekter Matches bekannt ist. Das Feldexperiment 5 wurde mit den gesamten Datensätzen der Datenquellen Wikidata und Crunchbase durchgeführt. In den Experimenten 6 bis 9 wurden neue Datenquellenkombinationen gebildet, die ebenfalls mit den gesamten Datensätzen integriert werden sollten. Die Experimente 5 bis 9 entsprechen somit den praxis- und unternehmensrelevanten Problemstellungen, für die keine Ground-Truth Daten aufgrund des hohen manuellen Aufwands erzeugt werden, sodass keine Angabe über die gesamten korrekten Matches möglich ist. In den Feldexperimenten fünf bis neun werden die in den Entwicklungsiterationen durch das Regelwerk klassifizierten Matches manuell geprüft, um die jeweilige Regel zu prüfen, ob diese sichere-Matches oder potenzielle-Matches generiert.

Das entwickelte Regelwerk für den Unternehmen-Matcher ist in Abbildung 6.11 auszugs-

Tabelle 6.4: Beschreibung der Feldexperimente

ID	Datenquellen	verwendete Attribute	Anz. Ds.	Korrekte Matches
1	AIStartups	name, ort	279	212
	Crunchbase	name, location_city, location_region, location_country_code	589.343	
2	Crunchbase	name, location_city, location_region, location_country_code	4.434	3895
	Wikidata	company_name, housenr, street, postalcode, city, country	4.177	
3	CompanyList	competitor	84	82
	Wikidata	company_name, housenr, street, postalcode, city, country	534.609	
4	Wikidata	company_name, housenr, street, postalcode, city, country	2.758	2787
	OpenCorporates	name, address, street_address, postal_code, locality, region, country	2.793	
5	Wikidata	company_name, housenr, street, postalcode, city, country	534.609	n.a.
	Crunchbase	name, location_city, location_region, location_country_code	589.343	
6	CapitalIQ	companyname, streetaddress, city, zipcode, state, country	2.910	n.a.
	Wikidata	company_name, housenr, street, postalcode, city, country	534.609	
7	USPTO	assignee	26.801	n.a.
	Crunchbase	name, location_city, location_region, location_country_code	589.343	
8	Databyte	Firma, Strasse, Hausnr, Plz, Ort	14.651	n.a.
	BureauVanDijk	companyname, address, addresspostalcode, addresscountry	68.162	
9	Handelsregister	company_name, registered_address, registered_office, federal_state	5.305.727	n.a.
	BureauVanDijk	companyname, address, adresscity, addresspostalcode, addresscountry	68.162	

weise dargestellt. Das Regelwerk umfasst insgesamt 45 verschiedene Regeln. Die einzelnen Regeln werden nacheinander auf die Menge der verbleibenden Tupel angewendet. Die Funktionsweise des Regelwerks soll am Beispiel des neunten Experiments, siehe Abbildung 6.11, erläutert werden.

Durch das Blocking und die Comparison des Unternehmen-Matcher wurden für das neunte Experiment 1.025.870 Tupel mit ihren Ähnlichkeitsmaßen berechnet und erzeugt. Auf diese Tupel wird nun die erste Regel des Regelwerkes ausgeführt. Die erste Regel klassifiziert 3.620 Tupel als Matches. In Experiment neun wird die Datenquelle Bureau van Dijk als führende Datenquelle angesehen. Dass heißt, dass jedem Buerau van Dijk Datensatz durch das Blocking bis zu 13 verschiedene Handelsregister Datensätze zugeordnet worden sind, die einen möglichen Match darstellen. Für 3.620 Bureau van Dijk Datensätze wurde durch Regel eins ein Match generiert, sodass die übrigen Tupel, die diese Bureau van Dijk Datensätze enthalten, als Kein-Match aussortiert werden können. Aus diesem Grund reduziert sich die Anzahl der Tupel um 60.053 und nicht um 3.620. Die verbleibenden 965.817 Tupel werden dann mit der darauffolgenden Regel, in diesem Fall Regel zwei, auf weitere Matches überprüft. In Experiment neun werden der letzten Regel noch 146.817 Tupel übergeben.

Die 45 Regeln sind während der Durchführung der neun Experimente iterativ weiterentwickelt worden. Durch Regel eins werden die Tupel mit vollständig vorhandenen Attributen selektiert. Das heißt, dass der Unternehmensname exakt übereinstimmen muss, die Adresse muss bei beiden Datensätzen auf Detailebene fünf vorliegen und exakt übereinstimmen sowie die

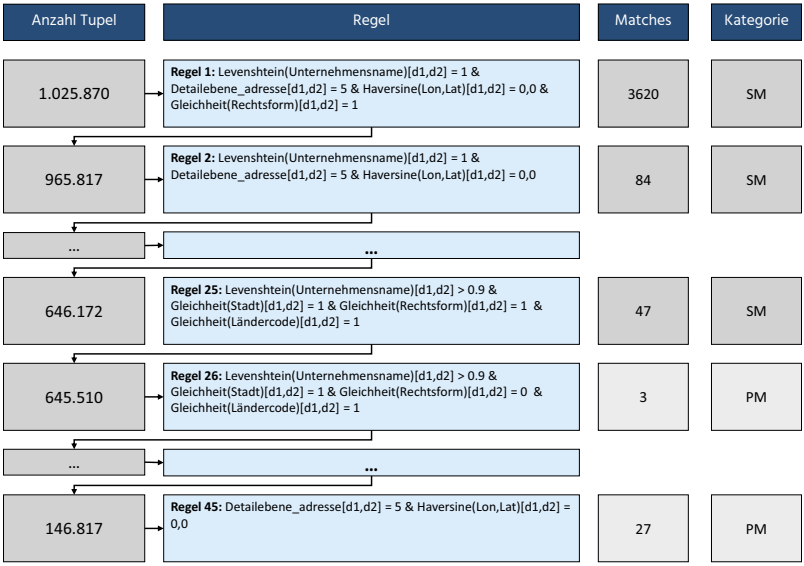


Abbildung 6.11: Auszug des Regelwerks des Unternehmen-Matcher

Rechtsform sollte übereinstimmen. Da diese Regel über alle Experimente hinweg ausschließlich TP Tupel selektiert hat, wird die Regel zur Kategorie sichere-Matches zugeordnet. Regel zwei selektiert Tupel, die einen übereinstimmenden Namen besitzen und die eine vollständige Adresse auf Detailebene fünf besitzen und exakt übereinstimmen. Der Unterschied zu Regel eins ist, dass die Rechtsform nicht geprüft wird, sodass Tupel, die keine Rechtsform besitzen oder die Rechtsform fehlt selektiert werden. Auch diese Regel selektierte über alle Experimente ausschließlich TP Tupel, sodass die Regel der Kategorie sichere-Matches zugeordnet wurde. Die folgenden Regeln wurden über die Experimente kontinuierlich angepasst und erweitert, dass bis einschließlich Regel 25 alle Regeln des Regelwerkes mit der Kategorie sichere-Matches versehen worden sind. Dabei prüft Regel 25 auf die verbliebenen Tupel die Levenshtein-Ähnlichkeit des Unternehmensnamens auf größer als 90%, die Gleichheit der Stadt, die Gleichheit des Ländercodes sowie die Gleichheit der Rechtsform. Die Regeln 26 bis 45 sind der Kategorie potenzielle-Matches zugeordnet, da über die neun Experimente nicht nur TP Tupel sondern auch FP Tupel selektiert worden sind. Daher sollten die Ergebnisse aus diesen Regeln für eine eindeutige Bestimmung der TP Tupel manuell geprüft werden. Durch Regel 26 werden Tupel selektiert, deren Levenshtein-Ähnlichkeit des Unternehmensnamens größer als 90% ist, deren Stadt gleich ist, deren Ländercode gleich ist und deren Rechtsform ungleich ist. Die Regel 45 selektiert abschließend Tupel, deren Adresse auf De-

tailebene fünf vorliegt und exakt übereinstimmt. Der Unternehmensname wird in Regel 45 nicht berücksichtigt.

In Tabelle 6.5 sind die Ergebnisse des finalen Regelwerkes für die neun Experimente dargestellt.

Tabelle 6.5: Übersicht der Ergebnisse des Unternehmen-Matcher für die Feldexperimente

ID	M	TP	FN	FP	SM	SM/TP	PM	TP-PM	TP-PM/PM	Manueller Aufwand
1	212	209	3	3	169	80,9%	43	40	93,0%	0,05 PT
2	4.392	3.866	29	526	2.293	59,3%	2.099	1.573	74,9%	2,0 PT
3	124	79	3	45	0	0,0%	124	79	63,7%	0,1 PT
4	2.513	2.444	343	69	728	29,8%	1.785	1.716	96,1%	1,8 PT
5	111.906	23.626	n.a.	88.280	16.122	68,2%	95.784	7.504	7,8%	95,0 PT
6	620	284	n.a.	336	107	37,7%	513	177	34,5%	0,5 PT
7	5.303	3.372	n.a.	1.931	755	22,4%	4.548	2.617	57,5%	4,5 PT
8	2.544	1.366	n.a.	1.178	1.024	75,0%	1.520	342	22,5%	1,5 PT
9	69.116	30.174	n.a.	38.942	24.553	81,4%	44.563	5.621	12,6%	44,5 PT

Tabelle 6.5 enthält die ID des Experiments. Zudem werden die durch das Regelwerk selektierte Anzahl an Matches (M) und die darin enthaltenen TP sowie die FP dargestellt. Für die ersten vier Experimente liegt zudem die Anzahl der FN vor, da eine gesamte Ground-Truth Datenmenge manuell erstellt worden ist. Die weitere Bewertung des Regelwerks für die Experimente besteht aus der Menge an sicheren-Matches (SM) und dem Anteil der sicheren-Matches an der Gesamtmenge der TP (SM/TP). Die Anzahl der potenziellen-Matches (PM) ist ebenfalls in Tabelle 6.5 dargestellt. Mit der Spalte TP-PM ist die Anzahl der TP Tupel aus den PM dargestellt. Die Quote der TP von der selektierten Menge an PM ist der Spalte TP-PM/PM zu entnehmen. In der letzten Spalte ist der manuelle Aufwand aufgeführt, der benötigt wurde, um die PM manuell zu prüfen. Lediglich für die Experimente fünf und neun wurden nicht alle Matches manuell geprüft. Für diese beiden großen Mengen an generierten Matches mit jeweils 111.906 und 69.116 wurden über alle Regeln hinweg zufällig 20% der Matches manuell geprüft. Die Ergebnisse der Stichprobenprüfung wurden für die Auswertung in Tabelle 6.5 auf die Gesamtheit hochgerechnet.

Experiment eins hat mit einer SM/TP Quote von 80,9% den zweithöchsten Wert. Der benötigte manuelle Aufwand von 0,05 PT zum Prüfen der PM entspricht dem zweitniedrigsten Wert. Da beide Datenquellen einen Schwerpunkt für Startup-Unternehmen besitzen, existieren viele

Matches. Trotz der Anzahl von 4.434 und 4.177 Unternehmen in Experiment zwei, beträgt der manuelle Aufwand für das Prüfen der PM lediglich 2 PT. Soll kein manueller Aufwand betrieben werden, werden 59,3% der TP bereits über SM generiert. Experiment drei generierte als einziges keine SM. Es wurden lediglich 124 PM generiert, die mit 0,1 PT manuellem Aufwand überprüft werden mussten. Die Begründung liegt darin, dass SM durch den Unternehmen-Matcher nicht generiert werden können, wenn lediglich der Unternehmensname ohne Rechtsform für die Datenintegration vorliegt. In Experiment vier wurden 29,8% der TP über SM generiert. Die TP-PM/PM Quote von 96,1% ist die höchste aller Experimente. Die Quote bedeutet, dass fast jedes selektierte Tupel aus den PM einem TP Tupel entspricht. In Experiment fünf klassifizierte der Unternehmen-Matcher mit 111.906 die meisten Matches. Insgesamt konnten 68,2% der TP über SM generiert werden. Die Quote der TP-PM/PM von 7,5% ist die geringste von allen Experimenten und gleichzeitig wird mit 95,0 PT der höchste manuelle Aufwand benötigt, um die 7.504 TP in den 95.784 PM zu identifizieren. In Experiment sechs wurden 284 TP generiert. Von diesen konnte der Unternehmen-Matcher 37,7% als SM selektieren. Für die übrigen 177 TP aus PM wurde ein manueller Prüfaufwand von 0,5 PT benötigt. In Experiment sieben wurden 3.372 TP generiert. Mit einer Quote von 22,4% SM erzielte das Regelwerk in diesem Experiment das zweitschlechteste Ergebnis. Die TP-PM/PM Quote von 57,5% ist das drittschlechteste Ergebnis aller Experimente. Insgesamt wurde ein manueller Aufwand von 4,5 PT benötigt, um die 4.548 PM zu überprüfen. In Experiment sieben hat der Unternehmen-Matcher mit 75,0% die drittbeste SM/TP Quote erzielt. Für die manuelle Prüfung der 1.520 PM wurden 1,5 PT benötigt, um weitere 342 TP zu identifizieren. Experiment neun stellt mit 69.116 selektierten Matches das zweitgrößte Experiment. Mit einer SM/TP Quote von 81,4% wurde in Experiment neun das beste Ergebnis erzielt. Die Quote von 12,6% der TP-PM/PM ist die zweitschlechteste, was zugleich einen manuellen Prüfaufwand von 44,5 PT bedeutet, um die PM zu prüfen und weitere 5.621 TP zu identifizieren.

7 Evaluation des Unternehmen-Matcher

In diesem Kapitel wird die Evaluation des in dieser Arbeit entwickelten Unternehmen-Matcher für das Datenquellen-unabhängige Integrieren von Unternehmensdaten beschrieben (siehe Abb. 7.1). Der entwickelte Unternehmen-Matcher wird innerhalb der Fallstudien bei den Evaluationspartnern Volkswagen AG (VW), EWE TEL GmbH (EWE), CEWE Stiftung & Co. KGaA (CEWE) sowie dem Oldenburgisch-Ostfriesischen Wasserverband (OOWV) eingesetzt und evaluiert.

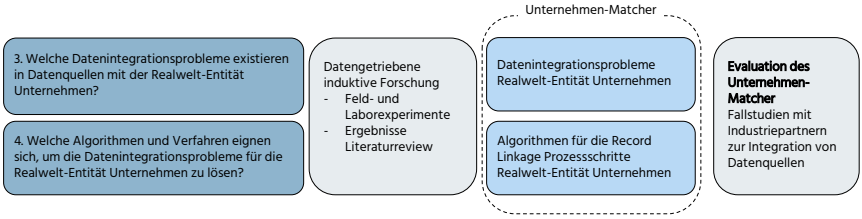


Abbildung 7.1: Forschungsvorgehen zur Evaluation des Unternehmen-Matcher

7.1 Forschungsmethode Fallstudie

Die Forschungsmethode Fallstudie eignet sich, um einen Erklärungsgrund (Warum?) oder einen Erklärungsansatz (Wie?) einer Problemstellung zu finden. Mit der Fallstudie wird ein aktuelles Phänomen, der Fall, eingehend und in seinem realen Kontext untersucht. Auf Basis von zuvor entwickelten theoretischen Erkenntnissen wird das Design, die Datenerhebung und die Analyse einer Fallstudie abgeleitet (vgl. R. K. Yin, 2018; Ridder, 2017).

Die Forschungsmethode Fallstudie steht in der Wissenschaft aufgrund der fehlenden Generalisierbarkeit der gewonnenen Erkenntnisse zum Teil in der Kritik. Dennoch besteht in der Wissenschaft ein gemeinsamer Konsens, dass die Fallstudienforschung die Generalisierbarkeit durch das Durchführen von mehreren Fällen sicherstellen kann (vgl. C.-Y. Yin, 2018; Ridder, 2017; Tsang, 2014). Im Gegensatz zur Fallstudie wird bei der Forschungsmethode Experiment das Phänomen oder der Fall bewusst aus seinem Kontext herausgelöst und nur das relevante Phänomen analysiert. In dieser Arbeit werden mehrere Fälle im Kontext Unternehmen durchgeführt, sodass in dieser Evaluation die Forschungsmethode Fallstudie verwendet wird. Die Durchführung einer Fallstudie kann in die Phasen (1) Planung, (2) Durchführung und (3) Analyse unterteilt werden (vgl. C.-Y. Yin, 2018). Der Prozess mit dem definierten Inhalt zur Durchführung der Fallstudie in dieser Arbeit ist in Abbildung 7.2 dargestellt und wird im Folgenden beschrieben.

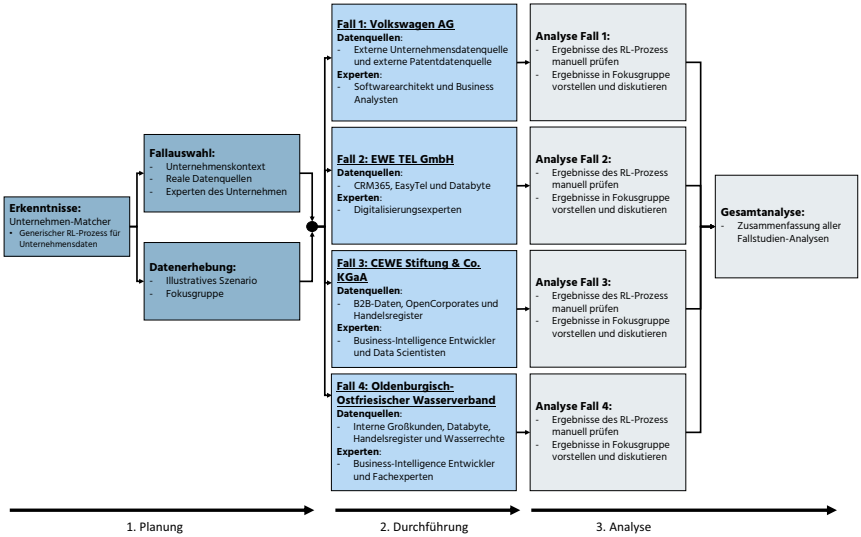


Abbildung 7.2: Fallstudienprozess zur Evaluation des Unternehmen-Matcher - eigene Darstellung in Anlehnung an C.-Y. Yin (2018)

Planung: Die Fallstudie dient der Evaluation des entwickelten Datenquellen-unabhängigen RL-Prozesses für Unternehmensdaten, dem Unternehmen-Matcher. Durch die Fallstudie sollen folgende Hypothesen im Unternehmenskontext geprüft werden:

Hypothese 1: Datenquellen-unabhängiger Ansatz Der Unternehmen-Matcher kann ohne weitere Anpassungen auf neue Datenintegrationsfälle bestehend aus zwei beliebigen Datenquellen, die die Realwelt-Entität Unternehmen beinhalten, übertragen werden.

Hypothese 2: Reduktion manueller Prüfaufwand Die durch den Unternehmen-Matcher generierten sicheren-Matches sind alle TPs, sodass diese nicht mehr manuell überprüft werden müssen.

Während der Durchführung der Experimente innerhalb der prototypischen Implementierung (siehe Abschnitt 6) lag der Fokus auf den Datenquellen und der Kontext des Einsatzes im Unternehmen wurde bewusst ausgeklammert. Durch die Fallstudie wird der Unternehmen-Matcher nun hinsichtlich der technischen und fachlichen Einsatzpotenziale im Unternehmenskontext evaluiert. Für die Evaluation konnten die Unternehmen VW, EWE, CEWE und OOWV gewonnen werden. Diese stellen Datenquellen und Experten für die Durchführung der Fallstudie bereit. Jedes Unternehmen stellt einen Fall

der Fallstudie dar, sodass die Ergebnisse über die vier Fälle generalisiert werden können. Für die Datenerhebung sollen zunächst illustrative Szenarien[43] durchgeführt werden. In diesen wird der Unternehmen-Matcher für die Datenquellen des jeweiligen Unternehmens eingesetzt. Mit der Forschungsmethode Fokusgruppe[44] werden qualitative Daten erhoben, indem die Ergebnisse mit den Experten des jeweiligen Unternehmens bewertet und diskutiert werden.

Durchführung: Die Durchführung der vier Fälle geschieht nach einem standardisierten Vorgehen, das im Folgenden beschrieben wird. Zunächst werden vom jeweiligen Unternehmen die für die Evaluation zur Verfügung gestellten Datenquellen vorgestellt. Anschließend werden die illustrativen Szenarien definiert. Ein illustratives Szenario besteht aus zwei Datenquellen des Unternehmens, die mit dem Unternehmen-Matcher integriert werden sollen. Da einige Unternehmen mehr als zwei Datenquellen für die Evaluation zur Verfügung stellen, werden in einigen Fällen mehrere illustrative Szenarien durchgeführt. Durch den Einsatz des Unternehmen-Matcher in den illustrativen Szenarien der vier Fälle wird geprüft, ob der Unternehmen-Matcher *Hypothese 1 Datenquellen-unabhängiger Ansatz* und *Hypothese 2 Reduktion manueller Prüfaufwand* erfüllt.

Zusätzlich werden die Ergebnisse der einzelnen illustrativen Szenarien innerhalb der Fälle mit den Experten der Unternehmen in Fokusgruppen-Workshops diskutiert, um qualitative Daten zur Bewertung des Einsatzes des Unternehmen-Matcher im Unternehmenskontext zu erheben. Innerhalb der Fokusgruppen-Workshops werden die von den Unternehmen beabsichtigten fachlichen Anwendungsfälle, die durch die Integration der zur Verfügung gestellten Datenquellen ermöglicht werden, wieder aufgegriffen und mit den Ergebnissen des Unternehmen-Matcher präsentiert. Daraufhin wird das Feedback der Experten zu den Ergebnissen eingeholt und die Ergebnisse werden an einzelnen Fällen detailliert analysiert, um zu evaluieren, ob der Einsatz des Unternehmen-Matcher im Unternehmen Mehrwerte schafft.

Analyse: Im letzten Prozessschritt, der Analyse, werden die durchgeführten illustrativen Szenarien und die Fokusgruppen-Workshops der vier Fälle zunächst einzeln analysiert und dargestellt. Bereits während der Durchführung der illustrativen Szenarien wird *Hypothese 1 Datenquellen-unabhängiger Ansatz* überprüft. Sollte der Unternehmen-Matcher in einem illustrativen Szenario nicht angepasst werden müssen, ist Hypothese 1 bestätigt. Die Evaluation des Unternehmen-Matcher in den illustrativen Szenarien erfolgt über die klassischen Evaluationsmetriken von RL-Prozessen (siehe Abschnitt 2.2.1.4). Für die Evaluation der *Hypothese 2 Reduktion manueller Prüfaufwand* werden die sicheren-Matches, manuell überprüft, um zu bestätigen, dass ausschließlich TP generiert wurden. Die poten-

[43] Beschreibung der Forschungsmethode in Abschnitt 4.3
[44] Beschreibung der Forschungsmethode in Abschnitt 4.3

ziellen-Matches werden manuell überprüft, um die darin enthaltenen TP zu identifizieren. Die Bestimmung des Recall für die illustrativen Szenarien der vier Fälle stellt auch für diese Evaluation eine Herausforderung dar, da für diese alle korrekten Matches benötigt werden, die nur über den manuellen Vergleich aller möglichen Datensatzpaare ermittelt werden können. Daher wird innerhalb der illustrativen Szenarien die Anzahl der FN, die für eine Berechnung des Recall notwendig wären, über eine Stichprobenprüfung ermittelt.

Innerhalb der Fokusgruppen-Workshops werden die Ergebnisse der illustrativen Szenarien durch die Experten der Unternehmen bewertet und Feedback zu den Einsatzpotenzialen des Unternehmen-Matcher erhoben und für jeden der vier Fälle zusammengefasst. Vor allem die Bewertung der Anzahl der korrekten Matches im Verhältnis zu den Kein-Matches und die Stichprobenprüfung der Kein-Matches zur Ermittlung möglicher FN soll diskutiert werden.

Zum Abschluss werden die Ergebnisse aus der Analyse der vier Fälle in einer Gesamtanalyse zusammengefasst und dargestellt.

In den folgenden Abschnitten werden die vier Fälle und die Ergebnisse der Durchführung und die Analyse der einzelnen Fälle präsentiert.

7.2 Fall 1: Volkswagen AG

Der erste Fall wird mit VW durchgeführt. VW ist einer der größten Automobilhersteller der Welt und besitzt zehn Marken[45] aus fünf europäischen Ländern: Volkswagen, Volkswagen Nutzfahrzeuge, ŠKODA, SEAT, CUPRA, Audi, Lamborghini, Bentley, Ducati und Porsche. Darüber hinaus bietet der VW Konzern ein weitere Marken und Geschäftsbereichen an, wie bspw. Finanzdienstleistungen, wozu die Händler- und Kundenfinanzierung, das Leasing, das Bank- und Versicherungsgeschäft sowie das Flottenmanagement zählen [46].

Abbildung 7.3: VW Logo - Volkswagen (o. J.)

Der Fall wird im Rahmen der Forschungskooperation TRACE[47] zwischen der VW AG und der Universität Oldenburg durchgeführt. Die VW AG hat mit der Forschungskooperation das Ziel, eine datenbasierte Sicht auf das Unternehmensumfeld und Märkte zu entwickeln. Dabei lizensiert die VW AG eine Vielzahl von externen Datenquellen ein, um relevante Daten- und Informationen analysieren zu können. Unter anderem werden eine externe Unternehmensdatenquelle und eine externe Patentdatenquelle für Analysen verwendet. Die Datenprovider

[45] Stand: 20.02.2022
[46] vgl. `https://www.volkswagenag.com/de/group.html`
[47] `https://projekt-trace.de/`

der beiden Datenquellen werden aufgrund des Geschäftsgeheimnisses von Volkswagen nicht genannt. Bisher werden die beiden Datenquellen einzeln analysiert und mit großem Aufwand wird versucht, die Erkenntnisse zu verknüpfen. Das Ziel dieses Falls soll es sein, die beiden Datenquellen mit dem Unternehmen-Matcher zu integrieren. Durch die Integration der beiden Datenquellen wird es der VW AG ermöglicht, die Informationen aus beiden Datenquellen zu analysieren und relevante Erkenntnisse zu generieren.

In Abbildung 7.4 sind Beispieldatensätze der Patentdatenquelle und der Unternehmensdatenquelle dargestellt. Die Patentdatenquelle erfasst jährlich 2,5 Millionen Patente die in 49 Millionen Patentfamilien eingruppiert werden. Dabei können einem Patent mehrere Assignee zugeordnet sein. Ein Assignee, der ein Patent einreicht, kann wiederum eine Person oder ein Unternehmen sein. Die Unternehmensdatenquelle enthält über 24 Millionen Unternehmensprofile. Ein solches Unternehmensprofil ist exemplarisch in Abbildung 7.4 für die VW AG dargestellt. Auch die Unternehmensdatenquelle wächst und verändert sich kontinuierlich. Aufgrund der großen und stetig wachsenden Datenmenge beider Datenquellen ist das manuelle Verknüpfen der Assignee aus den Patenten zu den Unternehmensprofilen aus der Unternehmensdatenquelle für VW keine Option.

Patentdatenquelle

Publication Number	Assignee
US20170132510A1	Facebook Inc.,Menlo Park,CA,US | FACEBOOK INC
EP3166025A1	Facebook Inc.,Menlo Park, CA 94025,US,101278072
WO2019105974A1	VOLKSWAGEN AKTIENGESELLSCHAFT,DE | AUDI AG,DE

Unternehmensdatenquelle

Attribut	Inhalt
Companyid	377732
Companyname	Volkswagen AG
City	Wolfsburg
Streetaddress	Berliner Ring 2
Streetaddress2	
Streetaddress3	
Streetaddress4	
Zipcode	38440
Country	Germany
Isocountry2	DE
Isocountry3	DEU

Abbildung 7.4: Fall 1 Volkswagen - Beispieldatensätze der Unternehmensdatenquelle und Patentdatenquelle

7.2.1 Illustratives Szenario

Zu Beginn des illustrativen Szenarios erfolgte das Schema Matching. Der Unternehmen-Matcher erwartet ein Attribut mit dem Inhalt des Unternehmensnamens und alle Attribute, die Adressdaten beinhalten können. Für die Unternehmensdatenquelle wurden für den Un-

ternehmensnamen das Attribut *companyname* und für die Adressdaten die Attribute *city*, *streetaddress*, *zipcode* und *country* übergeben.

Die Patentdatenquelle musste zunächst aufbereitet werden, bevor das Schema Matching für den Unternehmen-Matcher durchgeführt werden konnte. In Abbildung 7.4 ist zu sehen, dass jedem Patent mehrere Assignee zugeordnet sein können, die im selben Attribut (Assignee) über Pipe-Zeichen getrennt sind. Zuerst wurden die zugehörigen Assignee zu einem Patent in Zeilen separiert (siehe Abb. 7.5). Anschließend wurden die Komma-separierten Werte in die Attribute *data_0*, *data_1*, *data_2*, *data_3* und *data_4* unterteilt. Nun konnten das Attribut *data_0* als Unternehmensname und die Attribute *data_1*, *data_2*, *data_3* und *data_4* als Adressdaten an den Unternehmen-Matcher übergeben werden. Während die Adressdaten der Unternehmensdatenquelle strukturiert sind, befinden sich die Adressdaten in der Patentdatenquelle in unterschiedlichen Attributen. In Abbildung 7.5 ist zu sehen, dass die Adressdaten zum Land teilweise in Attribut *data_1* oder *data_3* vorliegen.

Patentdatenquelle

Publication Number	Assignee	Data_0	Data_1	Data_2	Data_3	Data_4
US20170132510A1	Facebook Inc.,Menlo Park,CA,US | FACEBOOK INC	Facebook Inc.	Menlo Park	CA	US	
		FACEBOOK INC				
EP3166025A1	Facebook Inc.,Menlo Park, CA 94025,US,101278072	Facebook Inc.	Menlo Park	CA 94025	US	101278072
WO2019105974A1	VOLKSWAGEN AKTIENGESELLSCHAFT,DE | AUDI AG,DE	VOLKSWAGEN AKTIENGESELLSCHAFT	DE			
		AUDI AG	DE			

Abbildung 7.5: Datenquellen-spezifische Aufbereitung Patentdatenquelle

Für die Durchführung des illustrativen Szenarios wurden von VW drei Datensätze aus der Patentdatenquelle zur Verfügung gestellt. Datensatz 1 umfasst 1.953 Patente, Datensatz 2 2.919 Patente und Datensatz 3 764 Patente (siehe Abb. 7.6). Das Ziel war es, die Assignee der Patente aus den drei Datensätzen mit einem der 24.618.442 Unternehmen aus der Unternehmensdatenquelle zu verbinden. In Abbildung 7.6 ist der Ablauf der illustrativen Szenarien für jeden Datensatz abgebildet.

Im ersten illustrativen Szenario wurden dem Unternehmen-Matcher mit Datensatz 1 1.953 Patente, die 1.562 verschiedene Assignee beinhalten, übergeben (siehe Abb. 7.6). Zwischen einem Patent und einem Assignee besteht eine m:n-Beziehung, weshalb die Anzahl der Patente von der Anzahl der Assignee abweichen kann. Der Unternehmen-Matcher hat 1.302 Matches generiert. Die 1.302 Matches werden durch das Regelwerk des Unternehmen-Matcher in 745 sichere-Matches und 557 potenzielle-Matches unterteilt. Die manuelle Prüfung der 557 potenziellen-Matches ergab 210 TP. Für die Evaluation und Prüfung von *Hypothese 2 Reduktion manueller Prüfaufwand* wurden auch die 745 sicheren-Matches überprüft. Alle sicheren-Treffer sind TP, wodurch Hypothese 2 bestätigt wird (siehe Abb. 7.6 und Tab.

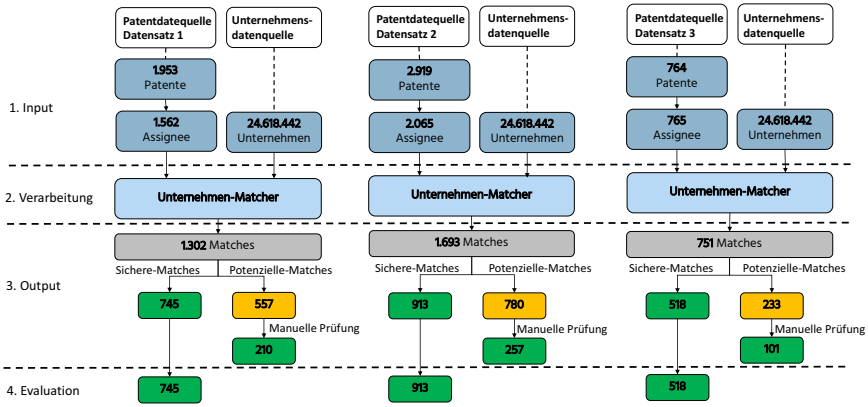


Abbildung 7.6: Fall 1 Volkswagen - Ablauf der illustrativen Szenarien

7.1). Mit dem Unternehmen-Matcher wurden 1.302 Matches generiert von denen 955 TP sind, was einer Precision von 73,3% entspricht. Der Anteil der sicheren-Matches, die nicht mehr manuell überprüft werden müssen, an den gesamten TP beträgt 78%. Für die restlichen 22% der TP wurde ein manueller Prüfaufwand von 0,56 PT[48] erbracht, um die 557 potenziellen-Matches zu überprüfen und weitere 210 TP zu identifizieren (siehe Tab. 7.1).

Tabelle 7.1: Fall 1 Volkswagen - Ergebnisse des Unternehmen-Matcher in den illustrativen Szenarien
PT = Personentage

	Datensatz 1	**Datensatz 2**	**Datensatz 3**
Matches	1302	1693	751
TP	955	1170	619
Precision	73,3%	69,1%	82,4%
Sichere-Matches	745	913	518
TP sichere-Matches	745	913	518
TP sichere-Matches / TP	78,0%	78,0%	83,7%
Potenzielle-Matches	557	780	233
TP potenzielle-Matches	210	257	101
TP potenzielle-Matches / TP	22,0%	22,0%	16,3%
Manueller Aufwand	0,56 PT	0,78 PT	0,24 PT

[48] Im Rahmen der Experimente zur Entwicklung des Unternehmen-Matcher wurden ca. 1000 Datensatzpaare pro Arbeitstag manuell geprüft. Daher wird für die Auswertung angenommen, dass auch ein Mitarbeiter im Unternehmen 1000 Datensatzpaare pro Arbeitstag prüfen kann.

Im zweiten illustrativen Szenario wurden dem Unternehmen-Matcher mit Datensatz 2 2.919 Patente, die 2.065 verschiedene Assignee beinhalten, übergeben. Der Unternehmen-Matcher hat 1.693 Matches generiert. Die 1.693 Matches wurden durch das Regelwerk des Unternehmen-Matcher in 913 sichere-Matches und 780 potenzielle-Matches unterteilt. Die manuelle Prüfung der 780 potenziellen-Matches ergab 257 TP. Für die Evaluation und Prüfung von *Hypothese 2 Reduktion manueller Prüfaufwand* wurden auch die 913 sicheren-Matches überprüft. Alle sicheren-Matches sind TP, wodurch Hypothese 2 bestätigt wird (siehe Abb. 7.6). Mit dem Unternehmen-Matcher wurden 1.693 Matches generiert von denen 1.170 TP sind, was einer Precision von 69,1% entspricht. Der Anteil der sicheren-Matches, die nicht mehr manuell überprüft werden müssen, an den gesamten TP beträgt 78,0%. Für die restlichen 22,0% der TP wurde ein manueller Prüfaufwand von 0,78 PT erbracht, um die 780 potenziellen-Matches zu überprüfen und weitere 257 TP zu identifizieren (siehe Tab. 7.1).

Im dritten illustrativen Szenario wurden dem Unternehmen-Matcher mit Datensatz 3 764 Patente, die 765 verschiedene Assignee beinhalten, übergeben. Der Unternehmen-Matcher hat 751 Matches generiert. Die 751 Matches wurden durch das Regelwerk des Unternehmen-Matcher in 518 sichere-Matches und 233 Treffer-Vorschläge unterteilt. Die manuelle Prüfung der 233 potenziellen-Matches ergab 101 TP. Für die Evaluation und Prüfung von *Hypothese 2 Reduktion manueller Prüfaufwand* wurden auch die 518 sicheren-Matches überprüft. Alle sicheren-Matches sind TP, wodurch Hypothese 2 bestätigt wird (siehe Abb. 7.6). Mit dem Unternehmen-Matcher wurden 751 Matches generiert von denen 619 TP sind, was einer Precision von 82,4% entspricht. Der Anteil der sicheren-Matches, die nicht mehr manuell überprüft werden müssen, an den gesamten TP beträgt 82,4%. Für die restlichen 16,3% der TP wurde ein manueller Prüfaufwand von 0,24 PT erbracht, um die 233 potenziellen-Matches zu überprüfen und weitere 101 TP zu identifizieren (siehe Tab. 7.1).

Um die Güte des Unternehmen-Matcher vollständig bewerten zu können, muss bewertet werden, ob die Anzahl der TP allen korrekten Matches entspricht oder ob noch FN in den illustrativen Szenarien existieren. Daher wurde zunächst ermittelt, wie viele der Assignee keinem Datensatz der Unternehmensdatenquelle zugeordnet worden sind. In Datensatz 1 wurden 30,5%, in Datensatz 2 31,9% und in Datensatz 3 24,2% der Assignee mit keinem Unternehmen der Unternehmensdatenquelle verknüpft (siehe Tab. 7.3 Kein-Match). Die Kein-Match Assignee Datensätze wurden stichprobenartig überprüft, um mögliche FN zu identifizieren. In Tabelle 7.2 sind die Ergebnisse der Stichprobenprüfung aufgeführt.

Der Tabelle 7.2 ist zu entnehmen, dass 19 der 32 Assignee eine Person repräsentieren. Der Unternehmen-Matcher ist auf das Integrieren von Unternehmensdaten fokussiert, weshalb diese Assignee nicht verknüpft werden können und diese TN darstellen. Weitere 8 der 32 Assignee sind nicht in der Unternehmensdatenquelle vorhanden und stellen damit ebenfalls TN dar. Mit 5 von 32 wurden Assignee identifiziert, die einem Unternehmen der Unter-

Tabelle 7.2: Fall 1 Volkswagen - Stichprobe der Assignee Datensätze der Kategorie Kein Match

Patent	Assignee	Kommentar
197	Coglix Co. Ltd.,KR	TN, da nicht vorhanden
198	Korea University Research and Business Foundation,KR	TN, da nicht vorhanden
240	BRAINSIGHT INC.,KR	TN, da nicht vorhanden
2847	MODU COMMUNICATION CO. LTD.,KR	TN, da nicht vorhanden
646	Intellectual Therapeutics Inc.,US	TN, da nicht vorhanden
684	Nutriomix Inc.,Pasadena, CA 91101,US,101887756	TN, da nicht vorhanden
201	Dong-eui University Industry-Academic Cooperation Foundation,KR	TN, da nicht vorhanden
231	DAS Disign Co. Ltd,KR	TN, da nicht vorhanden
198	AN BYEONGDUK,KR	TN, da Person
203	HWANG Kwan-joong,KR	TN, da Person
204	JI Mi Kyung,KR	TN, da Person
221	WON Sangyeon,KR	TN, da Person
224	Lee Do Kyun,KR	TN, da Person
225	SHIN TAE ZI,KR	TN, da Person
226	HAN HEE JU,KR	TN, da Person
234	KIM Jae Ho,KR	TN, da Person
234	JIN Sang Un,KR	TN, da Person
234	Yoo Kwan Woo,KR	TN, da Person
236	Choi Hee Jae,KR	TN, da Person
453	Ferry Keith George,Burbank,CA,US	TN, da Person
573	Pattnaik Manjula,IN	TN, da Person
573	Pattanaik Balachandra,ET	TN, da Person
573	Balachandran Ruthramurthy,ET	TN, da Person
573	Dwivedi Jaiprakash Narain,IN	TN, da Person
573	Saurav Swapnil,IN	TN, da Person
574	Dhiraj Kapila,IN	TN, da Person
453	Ferry Keith George,Burbank,CA,US	TN, da Person
505	SONY CORPORATION,TOKYO,JP	FN, da vorhanden
1873	Mapbox Inc.,San Francisco,CA,US	FN, da vorhanden
320	BLUBERD BIO INK.,US	FN, da vorhanden
526	Suzhou Gracell Biotechnologies Co. Ltd.	FN, da vorhanden
1917	L'Oreal,Paris,FR	FN, da vorhanden

nehmensdatenquelle hätten zugeordnet werden können und damit FN darstellen. Diese FN bedeuten weiteres Optimierungspotenzial für den Unternehmen-Matcher, um weitere TP zu erzeugen. Die erste Analyse zeigte, dass vor allem die Erweiterung um weitere Blocking-Verfahren und die Berücksichtigung der alternativen Unternehmensnamen dazu führen kann, dass die nicht verknüpften Assignee verknüpft werden.

Daher existieren die Folgenden drei Ursachen für die Kein-Match Assignee Datensätze:

1. Der Assignee des Patents existiert in der Unternehmensdatenquelle nicht.
2. Der Assignee ist kein Unternehmen, sondern eine Person.
3. Der Unternehmen-Matcher verknüpft das vorhandene Unternehmen der Unternehmensdatenquelle nicht.

Durch das Domänenwissen von VW über die Unternehmensdatenquelle ist bekannt, dass die Datenquelle bei asiatischen Unternehmen nicht so vollständig ist, wie bei europäischen und amerikanischen Unternehmen. In Datensatz 1 und Datensatz 2 sind 59% bzw. 58,8% der Assignee asiatische Unternehmen. Daher wurde in Datensatz 3 Patente zur Verfügung gestellt, die mit einem Anteil von 16,9% asiatischer Assignee deutlich weniger asiatische Unternehmen enthalten. Die deutlich höhere Precision vom illustrativen Szenario mit Datensatz 3 gegenüber Datensatz 1 und 2 wird auf die fehlenden asiatischen Unternehmen in der Unternehmensdatenquelle zurückgeführt.

Durch die illustrativen Szenarien mit den zur Verfügung gestellten Datenquellen von VW wurde bewiesen, dass der Unternehmen-Matcher nicht angepasst werden musste und die sicheren Regeln TP, die nicht geprüft hätten werden müssen, geliefert haben. Somit wurden Hypothese 1 und 2 durch diese illustrativen Szenarien bestätigt.

7.2.2 Fokusgruppe

Durch den Fokusgruppen-Workshop mit den Softwarearchitekten und Business Analysten der VW AG werden die Ergebnisse qualitativ bewertet werden. Hierzu wurden den Teilnehmern der Fokusgruppe die Ergebnisse aus Tabelle 7.3 vorgestellt.

Für VW war das Ziel durch die illustrativen Szenarien, jedes Patent über die zugehörigen Assignees mit mindestens einem Unternehmen aus der Unternehmensdatenquelle zu verknüpfen. Mit dem Unternehmen-Matcher wurden in Datensatz 1 69,5%, in Datensatz 2 68,1% und in Datensatz 3 75,8% der Patente mit einem Unternehmen der Unternehmensdatenquelle verknüpft (siehe Tabelle 7.3). Hierzu war für Datensatz 1 ein manueller Prüfaufwand von 0,56 PT, für Datensatz 2 0,78 PT und für Datensatz 3 0,23 PT notwendig (siehe Tabelle 7.1).

Tabelle 7.3: Fall 1 Volkswagen - Gesamtergebnis der illustrativen Szenarien

	Datensatz 1	Datensatz 2	Datensatz 3
Anzahl Patente	1953	2919	764
Mind. 1 Treffer	1357 (69,5%)	1989 (68,1%)	579 (75,8%)
Anzahl Assignee	1562	2065	765
Korrekte Matches	955 (61,1%)	1170 (43,3%)	619 (80,9%)
Kein Match	607 (38,9%)	895 (43,3%)	146 (19,1%)

Im Fokusgruppen-Workshop haben die VW Experten geschildert, dass bisher lediglich über eine manuell erstellte Konkordanzliste eine Auswahl von Unternehmen aus der Unternehmensdatenquelle mit der Patentdatenquelle verknüpft wird. Die Anzahl der ausgewählten Unternehmen ist daher aktuell durch den manuellen Aufwand stark limitiert. Durch den Einsatz des Unternehmen-Matcher wird diese Limitation aufgehoben, da alle Unternehmen aus der Capital IQ Datenquelle bei der Integration mit der Clarivate Derwent Datenquelle berücksichtigt werden können. Mit dem Unternehmen-Matcher werden mehr Patente mit weniger Aufwand als mit der aktuell vorhandenen Konkordanzliste mit Unternehmen aus der Capital IQ Datenquelle verknüpft, sodass die Softwarearchitekten und Business-Analysten von VW großes Einsatzpotenzial im Unternehmen-Matcher sehen.

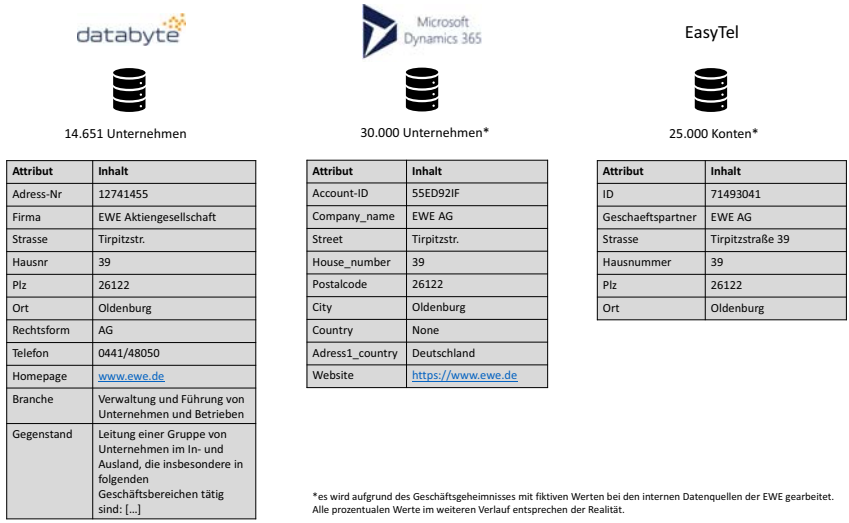
7.3 Fall 2: EWE TEL GmbH

Abbildung 7.7: EWE Logo - EWE (o. J.)

Der zweite Fall wurde mit der EWE TEL GmbH durchgeführt. Die EWE ist als Dienstleister in den Geschäftsfeldern Energie, Telekommunikation und Informationstechnologie tätig. Die EWE hat über 8.800 Mitarbeiter und erzielte rund 5,7 Milliarden Euro Umsatz im Jahr 2020 womit sie zu den großen Energieunternehmen in Deutschland gehört. Das Unternehmen hat seinen Hauptsitz in Oldenburg und befindet sich überwiegend in kommunaler Hand. Die EWE beliefert im Nordwesten Deutschlands, Brandenburg, auf Rügen und in Teilen Polens rund 1,4 Millionen Kunden mit Strom, rund 0,7 Millionen mit Erdgas sowie rund 0,7 Millionen mit Telekommunikationsdienstleistungen (vgl. EWE, o. J.).

Der Fall 2 wurde mit der Digitalisierungsabteilung des Geschäftskundenvertrieb der EWE TEL GmbH durchgeführt. Für die Evaluation des Unternehmen-Matcher wurden die EWE TEL internen Datenquellen EasyTel und Microsoft Dynamics 365 for Sales (folgend CRM365) sowie die externe Datenquelle Databyte verwendet (siehe Abb. 7.8). In Abbildung 7.8 ist für jede Datenquelle ein Beispieldatensatz mit den zur Verfügung stehenden Attributen abgebildet. Die Anzahl der Unternehmen und Konten der internen Datenquellen EasyTel und CRM365 der EWE TEL wurden aufgrund des Geschäftsgeheimnisses verändert und entsprechen fiktiven Werten. Die realen Zahlen wurden so verändert, dass die für die Präsentation der Evaluationsergebnisse relevanten prozentualen Werte der Realität entsprechen. Die 14.651 Unternehmen der Databyte Datenquelle repräsentieren alle Oldenburger Unternehmen aus dem Datenbestand von Databyte.

Laut der Digitalisierungsexperten der EWE TEL ist das CRM365 die zentrale Datenquelle des Geschäftskundenvertriebs. Daher wurden zwei Anwendungsfälle identifiziert und definiert, in denen Datenquellen durch den Unternehmen-Matcher mit dem CRM365 integriert werden. Zuerst wurden die CRM365 Unternehmen mit den Unternehmen der Databyte Datenquelle integriert, um das CRM365 mit externen Informationen anzureichern. Als nächstes wurde die interne Datenquelle EasyTel mit dem CRM365 verknüpft. Die EasyTel Datenquelle verwaltet die Verträge der Unternehmenskunden und ist über keine gemeinsame ID mit dem CRM365 verbunden. Vor ca. drei Jahren wurden alle Unternehmensdatensätze des CRM365 mit der Databyte Datenquelle durch die Databyte GmbH verknüpft, weshalb für dieses illustrative Szenario eine Ground-Truth Datenmenge existiert. Die Integration des CRM365 mit der EasyTel Datenquelle wurde ebenfalls in den letzten Jahren durch ein EWE TEL internes Projekt umgesetzt, weshalb für dieses illustrative Szenario ebenfalls Ground-Truth Daten und zusätzlich manuelle Aufwände in Personentagen für den Vergleich mit dem Unternehmen-Matcher vorliegen.



Attribut	Inhalt
Adress-Nr	12741455
Firma	EWE Aktiengesellschaft
Strasse	Tirpitzstr.
Hausnr	39
Plz	26122
Ort	Oldenburg
Rechtsform	AG
Telefon	0441/48050
Homepage	www.ewe.de
Branche	Verwaltung und Führung von Unternehmen und Betrieben
Gegenstand	Leitung einer Gruppe von Unternehmen im In- und Ausland, die insbesondere in folgenden Geschäftsbereichen tätig sind: [...]

Attribut	Inhalt
Account-ID	55ED92IF
Company_name	EWE AG
Street	Tirpitzstr.
House_number	39
Postalcode	26122
City	Oldenburg
Country	None
Adress1_country	Deutschland
Website	https://www.ewe.de

Attribut	Inhalt
ID	71493041
Geschaeftspartner	EWE AG
Strasse	Tirpitzstraße 39
Hausnummer	39
Plz	26122
Ort	Oldenburg

Abbildung 7.8: Fall 2 EWE - Beispieldatensätze der Datenquellen Databyte, Microsoft Dynamics 365 und EasyTel

7.3.1 Illustratives Szenario

Das erste illustrative Szenario wurde mit den Datenquellen CRM365 und Databyte durchgeführt. Dem Unternehmen-Matcher wurden für die Datenquelle CRM365 das Attribut *company_name* als Unternehmensname und die Attribute *Street*, *House_number*, *Postalcode*, *City*, *Country* und *Adress1_country* als Adressdaten übergeben (siehe Abb. 7.9). Für die Datenquelle Databyte wurde das Attribut *Firma* für den Unternehmensnamen und die Attribute *Strasse*, *Hausnr*, *Plz* und *Ort* als Adressdaten übergeben (siehe Abb. 7.9).

Für die Durchführung des illustrativen Szenarios hat der Datenprovider Databyte[49] 14.651 Datensätze zur Verfügung gestellt. Die 14.651 Datensätze sind ausschließlich Unternehmen mit Firmensitz in Oldenburg. In Abbildung 7.9 ist der Ablauf der illustrativen Szenarien für die Datenquellen CRM365 und Databyte abgebildet. Zuerst wurden die Daten dem Unternehmen-Matcher übergeben. Die Verarbeitung des Unternehmen-Matcher generierte 699 Matches. Die 699 Matches wurden durch das Regelwerk des Unternehmen-Matcher in 651 sichere-Matches und 48 potenzielle-Matches unterteilt. Die manuelle Prüfung der 48 potenziellen-Matches ergab 32 TP. Die 651 sicheren-Matches wurden zur Evaluation

[49] `https://www.databyte.de/`

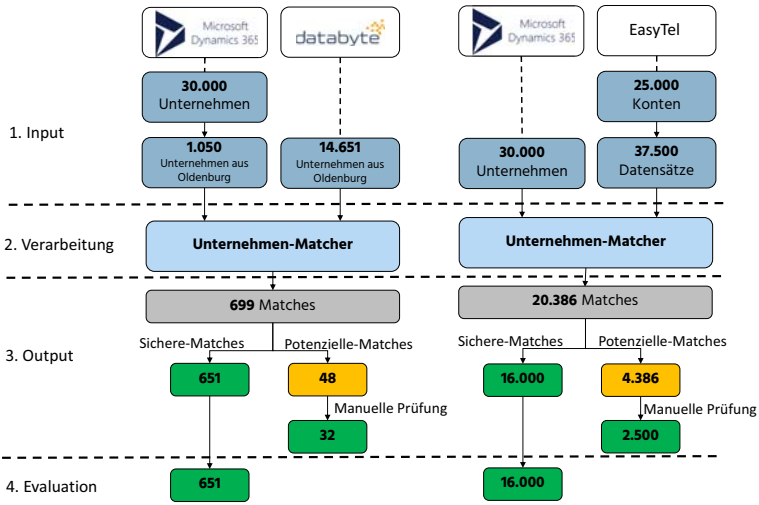


Abbildung 7.9: Fall 2 EWE - Ablauf der illustrativen Szenarien

und Prüfung der *Hypothese 2 Reduktion manueller Aufwand* manuell geprüft. Alle sicheren-Matches sind TP, wodurch Hypothese 2 bestätigt wird. Mit dem Unternehmen-Matcher wurden 699 Matches generiert von denen 683 TP sind, was einer Precision von 97,7% entspricht (siehe Tab. 7.4). Der Anteil der sicheren-Matches, die nicht mehr manuell überprüft werden müssen, an den gesamten TP beträgt 95,3%. Für die restlichen 4,7% der TP wurde ein manueller Prüfaufwand von 0,1 PT erbracht, um die 48 potenziellen-Matches zu überprüfen und weitere 32 TP zu identifizieren (siehe Tab. 7.4).

Das zweite illustrative Szenario wurde mit den EWE TEL internen Datenquellen CRM365 und Easytel durchgeführt. Das Ziel war es, jedes der 25.000 EasyTel Konten mit einem Unternehmensdatensatz aus dem CRM365 zu verknüpfen. Zunächst wurden die Daten dem Unternehmen-Matcher übergeben (siehe Abb. 7.9). Zu einem EasyTel Konto können mehrere Unternehmensnamen und verschiedene Adressangaben existieren. In diesen Fällen wurden alle vorhandenen Kombinationen aus Adressangabe und Unternehmensname selektiert und an den Unternehmen-Matcher übergeben. Daher wurden aus den 25.000 EasyTel Konten 37.500 Datensätze, die an den Unternehmen-Matcher übergeben wurden. Die Verarbeitung des Unternehmen-Matcher hat 20.386 Matches generiert. Die 20.386 Matches wurden durch das Regelwerk des Unternehmen-Matcher in 16.000 sichere-Matches und 4.386 potenzielle-Matches unterteilt. Die manuelle Prüfung der 4.386 potenziellen-Matches ergab 2.500 TP. Für die Evaluation der *Hypothese 2 Reduktion manueller Prüfaufwand* wur-

Tabelle 7.4: Fall 2 EWE TEL - Ergebnisse des Unternehmen-Matcher in den illustrativen Szenarien
PT = Personentage; UM = Unternehmen-Matcher; DP = Databyte Projekt; IP = Internes Projekt

	CRM365 - Databyte		**EasyTel - CRM365**	
	UM	**DP**	**UM**	**IP**
Matches	699	k. A.	20.386	k. A.
TP	683	819	18.500	19.000
Precision	97,7%	k. A.	90,7%	k. A.
Sichere-Matches	651	k. A.	16.000	k. A.
TP sichere-Matches	651	k. A.	16.000	k. A.
TP sichere-Matches/TP	95,3%	k. A.	86,5%	k. A.
Potenzielle-Matches	48	k. A.	4.386	k. A.
TP potenzielle-Matches	32	k. A.	2.500	k. A.
TP potenzielle-Matches/TP	4,7%	k. A.	13,5%	k. A.
Manueller Aufwand	0,1 PT	k. A.	4,4 PT	>100 PT

den die sicheren-Matches überprüft. Alle sicheren-Matches sind TP, wodurch Hypothese 2 bestätigt wird (siehe Tab. 7.4). Mit dem Unternehmen-Matcher wurden 20.386 Matches generiert, von denen 18.500 TP sind, was einer Precision von 90,7% entspricht (siehe Tab. 7.4). Der Anteil der sicheren-Matches, die nicht mehr manuell überprüft werden müssen, an den gesamten TP beträgt 86,5%. Für die restlichen 13,5% der TP wurde ein manueller Prüfaufwand von 4,4 PT erbracht, um die 4.386 potenziellen-Matches zu überprüfen und weitere 2.500 TP zu identifizieren (siehe Tab. 7.4).

Um die Güte der Ergebnisse für die beiden illustrativen Szenarien durch den Unternehmen-Matcher vollständig bewerten zu können, muss bewertet werden, ob die Anzahl der TP allen existierenden korrekten Matches entspricht oder ob FN in den illustrativen Szenarien existieren. Da für beide illustrativen Szenarien Ergebnisse aus früheren Projekten vorliegen, werden diese Ergebnisse zusammen mit den Ergebnissen des Unternehmen-Matcher als Ground-Truth angesehen.

Für das illustrative Szenario CRM365-Databyte existieren für die 1050 Unternehmen 874 korrekte Matches, die sich aus der Summe der TP durch den Unternehmen-Matcher und dem Databyte Projekt zusammensetzen. Für den Unternehmen-Matcher bedeutet dies eine Anzahl von 191 FN und einen Recall von 78,1 %. Für das Databyte Projekt bedeutet dies eine Anzahl von 55 FN und einen Recall von 93,7 %. Da der Unternehmen-Matcher TP generiert hat, die im Databyte Projekt nicht generiert wurden und umgekehrt, erreicht keiner der beiden Ansätze einen Recall von 100%. Allerdings können noch weitere FN für die verbleibenden 176 CRM365 Unternehmensdatensätze, die keinen Match besitzen, existieren, die im Rahmen dieser Arbeit nicht weiter untersucht worden sind. Für das illustrative Szenario

Tabelle 7.5: Fall 2 EWE TEL - Gesamtergebnis der illustrativen Szenarien

	CRM365-Databyte		**EasyTel - CRM365**	
	Unternehmen-Matcher	**Databyte Projekt**	**Unternehmen-Matcher**	**Internes Projekt**
Anzahl Unternehmen	1.050	1.050	25.000	25.000
Korrekte Matches	683 (65,1%)	819 (78,0%)	18.500 (74,0%)	19.000 (76,0%)
Kein Match	367 (44,9%)	231 (22,0%)	6.500 (26,0%)	6.000 (24,0%)

EasyTel-CRM365 existieren für die 25.000 Konten 19.000 korrekte Matches. Diese korrekten Matches wurden in einem internen EWE TEL Projekt ermittelt. Die Experten der EWE TEL gehen davon aus, dass für die 24,0% der Konten, die der Kategorie Kein Match (siehe Tab. 7.5) zugeordnet werden, weitere korrekte Matches existieren. Dennoch werden die 19.000 korrekten Matches für diese Arbeit als Gesamtmenge aller korrekte Matches für die Berechnung des Recall angenommen. Für den Unternehmen-Matcher bedeutet dies eine Anzahl von 500 FN und einen Recall von 97,4%. Da wir die Anzahl der TP aus dem internen EWE TEL Projekt als die Gesamtmenge aller korrekten Matches annehmen, erzielte das interne Projekt einen Recall von 100%.

Durch die illustrativen Szenarien mit den Datenquellen der EWE TEL wurde bewiesen, dass der Unternehmen-Matcher nicht angepasst werden musste und die sicheren-Matches TP geliefert haben. Somit wurden Hypothese 1 und 2 durch diese illustrativen Szenarien bestätigt.

7.3.2 Fokusgruppe

Die Ergebnisse der durchgeführten illustrativen Szenarien wurden den Digitalisierungsexperten der EWE TEL in Fokusgruppen-Workshops vorgestellt und diskutiert. Die vorgestellten und diskutierten Ergebnisse der Integration der CRM365 Datenquelle mit der Databyte Datenquelle sind in Tabelle 7.4 und 7.5 aufgeführt. Außerdem sind die Ergebnisse der Datenintegration aus dem Projekt mit Databyte, welches vor 3 Jahren durchgeführt worden ist, in Tabelle 7.4 und 7.5 dargestellt. Mit dem Unternehmen-Matcher wurden 683, 65,1%, von den 1.050 Oldenburger Unternehmen des CRM365 mit einem Unternehmen in der Databyte Datenquelle verknüpft. Für die Integration der beiden Datenquellen war ein manueller Prüfaufwand von 0,1 PT notwendig.

Im Projekt mit Databyte hat die EWE TEL ihre gesamten Unternehmenskundendatensätze an den Datenprovider als CSV-Datei geliefert und nach einigen Wochen die CSV-Datei inklusive der verknüpften Databyte Datensätze zurück erhalten. Von den 1.050 Oldenburger Unternehmen im CRM365 sind 819, 78,0%, mit einem Databyte Datensatz verknüpft worden. Den Digitalisierungsexperten der EWE TEL ist nicht bekannt, wie viel Aufwand in

PT oder mit welchem Automatisierungsgrad der Datenprovider Databyte die Datenintegration durchgeführt hat, weshalb ein Vergleich mit dem Unternehmen-Matcher für diese Aspekte nicht möglich ist.

Obwohl der Unternehmen-Matcher gegenüber dem Databyte Projekt 12,9% weniger Unternehmensdatensätze des CRM365 mit Databyte Unternehmen integriert hat, sehen die Digitalisierungsexperten großes Potenzial im Einsatz des Unternehmen-Matcher. Dies liegt vor allem darin begründet, dass der Unternehmen-Matcher eine Datenprovider-unabhängige Lösung darstellt und keine Kundendaten der EWE TEL speichert. Die Unternehmenskundendaten der EWE TEL sind ein sensibles und geschäftskritisches Asset, dass bisher an Datenprovider herausgegeben werden musste, um die Unternehmenskundendaten mit externen Daten zu integrieren und anzureichern. Durch den Einsatzes des Unternehmen-Matcher verlassen die Unternehmenskundendatensätze die IT-Infrastruktur der EWE TEL nicht. Weiterhin sehen sie großes Potenzial im Einsatz des Datenprovider-unabhängigen Unternehmen-Matcher darin, dass mit geringem Aufwand weitere externe Datenquellen von Datenprovidern integriert und getestet werden können, um einen Vergleich durchzuführen und mögliche Einsatzszenarien entwickeln zu können. Neben dem geringeren Aufwand wurde auch die verkürzte Projektlaufzeit von Datenintegrationsprojekten von den Digitalisierungsexperten als Vorteil genannt. Während das Databyte Projekt in einen Zeitraum von mehreren Wochen durchgeführt worden ist, hat der Einsatz des Unternehmen-Matcher inklusive des initialen Setup[50] in der EWE TEL Infrastruktur und dem manuelle Prüfen der Datensätze mit 6 PT Aufwand ca. 1,5 Wochen gedauert. Zusätzlich hat der Unternehmen-Matcher ca. 5% der CRM365 Unternehmen mit einem Databyte Datensatz verknüpft, der im Databyte Projekt nicht verknüpft worden ist. Im Databyte Projekt wurden ca. 14% der CRM365 Unternehmen mit einem Databyte Datensatz verknüpft, die vom Unternehmen-Matcher nicht verknüpft worden sind. Die Digitalisierungsexperten fordern für einen kommerziellen Einsatz des prototypischen Unternehmen-Matcher eine annähernd gleiche Trefferquote wie die im Databyte Projekt.

Im zweiten illustrativen Szenario wurden die EasyTel Konten mit den CRM365 Unternehmen verknüpft. Die im Fokusgruppen-Workshop vorgestellten und diskutierten Ergebnisse durch den Einsatz des Unternehmen-Matcher im Vergleich zum intern durchgeführten Projekt sind in Tabelle 7.5 dargestellt.

Der Unternehmen-Matcher konnte 18.500 von 25.000 EasyTel Konten mit einem CRM365 Unternehmen verknüpfen (siehe Tab. 7.5). Zudem wurden 4.386 Treffer-Vorschläge generiert die mit einem manuellen Aufwand von 4,4 PT geprüft worden sind. Mit dem initialen Setup des Unternehmen-Matcher in der EWE TEL Infrastruktur hatte dieses illustrative

[50] Das initiale Setup umfasste das Einrichten der Python-Umgebung sowie der Datenquellenanbindungen zum Ausführen des Unternehmen-Matcher auf dem Notebook der EWE.

Szenario einen Aufwand von 15 Personentagen und erstreckte sich über einen Zeitraum von 3 Wochen. Vor drei Jahren hat die EWE TEL ein internes Projekt durchgeführt, um die EasyTel Konten mit dem CRM365 zu integrieren. Das Projekt dauerte über ein Jahr und hat in Summe über 100 PT Aufwand benötigt. Durch das Projekt sind 19.000 der 25.000 EasyTel Konten mit einem CRM365 Unternehmen verknüpft worden. Der prototypische Unternehmen-Matcher hat lediglich 2% weniger EasyTel Konten mit einem CRM365 Unternehmen verknüpft. Da im internen Projekt die Verknüpfung größtenteils manuell erfolgt ist, passierten hier Bearbeitungsfehler. Durch den Unternehmen-Matcher wurden 3,3% dieser falsch mit CRM365 Unternehmen verknüpften EasyTel Konten identifiziert.

Die Digitalisierungsexperten der EWE TEL bewerteten dieses illustrative Szenario als sehr erfolgreich. Die Ergebnisqualität des Unternehmen-Matcher weicht um 2% vom durchgeführten internen Projekt ab. Allerdings hätte der Unternehmen-Matcher den Aufwand um 90% und die Projektlaufzeit um 11 Monate gegenüber dem durchgeführten internen Projekt reduziert. Sie betonten zudem die positiven Auswirkungen auf den operativen Betrieb im Unternehmen, wenn die Integration von zwei vertriebsrelevanten Systemen nicht über 12 Monate durchgeführt wird, sondern auf 3 Wochen verkürzt werden kann. Großes Potenzial im Einsatz des Unternehmen-Matcher sehen die Digitalisierungsexperten darin, dass dieser aufgrund des deutlich reduzierten Aufwands auch kontinuierlich eingesetzt werden kann, um die Datenquellen EasyTel und CRM365 kontinuierlich miteinander zu verknüpfen und damit die Stammdatenqualität kontinuierlich sicherzustellen und zu erhöhen.

7.4 Fall 3: CEWE Stiftung & Co. KGaA

Der dritte Fall wurde mit CEWE durchgeführt. CEWE ist ein europäischer Fotodienstleister und Anbieter im kommerziellen Online-Druck. CEWE bietet Produkte im Fotofinishing an, wie das CEWE FOTOBUCH, Wandbilder, Kalender, klassische Fotoabzüge sowie eine Vielfalt an Fotogeschenken. Gemeinsam mit ihren Partnern (bspw. dm und Müller) bietet CEWE ihren Kunden Multi-Channel-Konzepte in Form stationärer Läden und Online-Shops in verschiedenen Ländern. An den Standorten in Europa sind derzeit c.a. 4.000 Mitarbeitende beschäftigt. Das Vertriebsgebiet von CEWE erstreckt sich über 21 Länder in Europa. CEWE produziert seine Produkte europaweit in 14 Betriebsstätten (vgl. CEWE, o. J.).

Abbildung 7.10: CEWE Logo - CEWE (o. J.)

Der Fall 3 wird mit der Abteilung Corporate Information Management (CIM) von CEWE durchgeführt. Für die Evaluation des Unternehmen-Matcher wurden interne B2B-Daten zur Verfügung gestellt (siehe Abb. 7.11). Die internen B2B-Daten wurden zunächst mit der

Datenquelle OpenCorporates und anschließend mit der Datenquelle Handelsregister integriert. In Abbildung 7.11 ist für jede der drei Datenquellen ein Beispieldatensatz abgebildet.

Das Ziel der Data Scientisten der Abteilung CIM ist es, durch die Integration der kommerziellen Datenquelle OpenCorporates und der frei verfügbaren Datenquelle Handelsregister das Einsatzpotenzial von externen Daten für ihre B2B-Daten zu überprüfen.

122.184 Unternehmen

Attribut	Inhalt
Company_number	P3210_HRB208214
Name	CEWE Stiftung & Co. KGaA
Jurisdiction_code	de
Current_status	Currently Registered
Registered_adress_in_full	Meerweg 30-32, 26133 Oldenburg.
Street_adress	Meerweg 30-32, 26133 Oldenburg.
Locality	Null
Region	Null
Postal_code	Null
Country	Germany

B2B-Daten

86.635 Unternehmen

Attribut	Inhalt
ID	1
Name	HybridSupply GmbH
Mailingstreet	Stellmacherstraße
Mailingcity	Lübeck
Mailingstate	Null
Mailingpostalcode	23556
Mailingcountry	Deutschland

Handelsregister

5.305.727 Unternehmen

Attribut	Inhalt
Company_number	P3210_HRA1548
Name	CeWe Color AG & Co. OHG
Registered_address	Meerweg 30-32, 26133 Oldenburg.
Current_status	Removed
Federal_state	Lower Saxony
Registered_office	Oldenburg
Registrar	Oldenburg (Oldenburg)

Abbildung 7.11: Fall 3 CEWE - Beispieldatensätze der Datenquellen OpenCorporates, B2B-Daten und Handelsregister

7.4.1 Illustratives Szenario

Das erste illustrative Szenario wurde mit den Datenquellen B2B-Daten und OpenCorporates durchgeführt. Für die Datenquelle B2B-Daten wurden dem Unternehmen-Matcher das Attribut *Name* als Unternehmensname und die Attribute *Mailingstreet*, *Mailingcity*, *Mailingstate*, *Mailingpostalcode* und *Mailingcountry* als Adressdaten übergeben (siehe Abb. 7.11). Für die Datenquelle OpenCorporates wurde das Attribut *Name* für den Unternehmensnamen und die Attribute *Registered_adress_in_full* und *Country* als Adressdaten übergeben. Die Datenquelle OpenCorporates enthält noch weitere Adressangaben wie *Locality*, *Region* oder *Postal_code* (siehe Abb. 7.11). Da diese selten gefüllt sind und alle relevanten Adressangaben in den Attributen *Registered_adress_in_full* und *Country* vorhanden sind, wurden ausschließlich diese verwendet. Für die Datenquelle Handelsregister wurde dem Unternehmen-Matcher das Attribut *Name* für den Unternehmensnamen und die Attribute *Registered_adress*, *Federal_state* und *Registered_office* als Adressdaten übergeben.

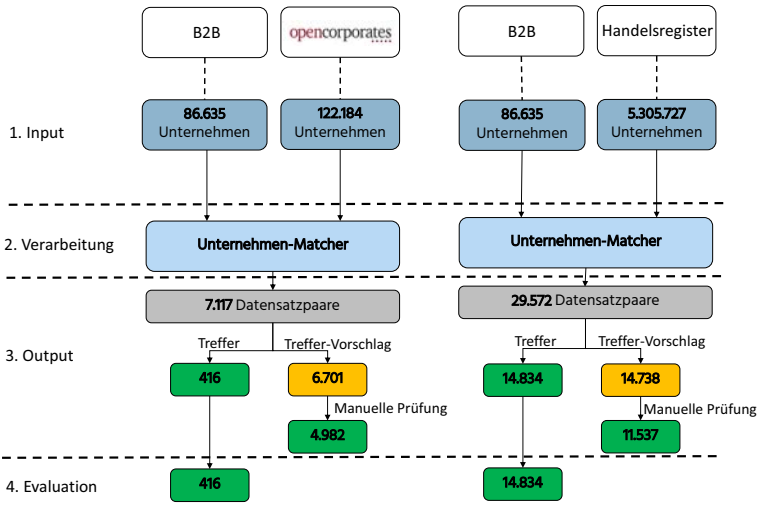


Abbildung 7.12: Fall 3 CEWE - Ablauf der illustrativen Szenarien

In Abbildung 7.12 ist der Ablauf des illustrativen Szenarios mit den Datenquellen B2B-Daten und OpenCorporates abgebildet. Dem Unternehmen-Matcher wurden 86.635 Unternehmensdatensätze aus der B2B Datenquelle und 122.184 Unternehmensdatensätze aus der OpenCorporates Datenquelle übergeben. Der Unternehmen-Matcher hat 7.117 Matches generiert. Von den 7.117 Matches wurden 416 sichere-Matches und 6.701 potenzielle-Matches generiert. Die manuelle Prüfung der 6.701 potenziellen-Matches ergab 4.982 TP. Die 416 sichere-Matches wurden zur Evaluation und Prüfung der *Hypothese 2 Reduktion manueller Prüfaufwand* geprüft. Alle sicheren-Matches sind TP, wodurch Hypothese 2 bestätigt wird. Mit dem Unternehmen-Matcher wurden 7.117 Matches generiert von denen 5.398 TP sind, was einer Precision von 75,8% entspricht. Der Anteil der sicheren-Matches, die nicht mehr manuell überprüft werden müssen, an den gesamten TP beträgt 7,7%. Für die übrigen 92,3% der TP wurde ein manueller Prüfaufwand von 7 PT erbracht, um die 6.701 potenziellen-Matches zu überprüfen und weitere 4.982 TP zu identifizieren (siehe Tab. 7.6).

Für das zweite illustrative Szenario wurden dem Unternehmen-Matcher die 86.635 Unternehmensdatensätze der B2B Datenquelle und die 5.305.727 Unternehmensdatensätze aus dem Handelsregister übergeben. Der Unternehmen-Matcher generierte 29.572 Matches. Die 29.572 Matches wurden durch das Regelwerk des Unternehmen-Matcher in 14.834 sichere-Matches und 14.738 potenzielle-Matches unterteilt. Die manuelle Prüfung der 14.738 potenziellen-Matches ergab 11.537 TP. Die sicheren-Matches wurden zur Evaluation und

Tabelle 7.6: Fall 3 CEWE - Ergebnisse des Unternehmen-Matcher in den illustrativen Szenarien
PT = Personentage

	B2B-Daten	
	OpenCorporates	**Handelsregister**
Matches	7.117	29.572
TP	5.398	26.371
Precision	75,8%	89,2%
Sichere-Matches	416	14.834
TP sichere-Matches	416	14.834
TP sichere-Matches / TP	7,7%	56,3%
Potenzielle-Matches	6.701	14.738
TP potenzielle-Matches	4.982	11.537
TP potenzielle-Matches / TP	92,3%	43,7%
Manueller Aufwand	7 PT	15 PT

Prüfung der *Hypothese 2 Reduktion manueller Prüfaufwand* geprüft. Alle sicheren-Matches sind TP, wodurch Hypothese 2 bestätigt wird. Mit dem Unternehmen-Matcher wurden 29.572 Matches generiert von denen 26.371 TP sind, was einer Precision von 89,2% entspricht. Der Anteil der sicheren-Matches, die nicht mehr manuell überprüft werden müssen, an den gesamten TP beträgt 56,3%. Für die übrigen 43,7% der TP wurde ein manueller Prüfaufwand von 15 PT erbracht, um die 14.738 potenziellen-Matches zu überprüfen und weitere 11.537 TP zu identifizieren (siehe Tab. 7.6).

Durch die illustrativen Szenarien mit den Datenquellen von CEWE wurde bewiesen, dass der Unternehmen-Matcher nicht angepasst werden musste und die sicheren Regeln korrekte Treffer geliefert haben. Somit wurden Hypothese 1 und 2 durch diese illustrativen Szenarien bestätigt.

7.4.2 Fokusgruppe

Die Ergebnisse der durchgeführten illustrativen Szenarien wurden den Data Scientisten von CEWE vorgestellt und mit diesen diskutiert. Die Ergebnisse sind in Tabelle 7.6 und 7.7 dargestellt. Das Ziel der illustrativen Szenarien war es, jedem Datensatz der B2B-Daten zu einem OpenCorporates Unternehmen und Handelsregister Unternehmen zuzuordnen.

Der Unternehmen-Matcher hat 5.398 Unternehmen, 6,2% der B2B-Daten mit einem OpenCorporates Datensatz verknüpft (siehe Tab. 7.6). Der manuelle Aufwand betrug 7 Personentage, um die potenziellen-Matches zu prüfen. Im Anschluss an die Vorstellung der Ergebnisse, wurde die Anzahl an korrekten Matches diskutiert. Zum einen Stand für das illustrative

Tabelle 7.7: Fall 3 CEWE - Gesamtergebnis der illustrativen Szenarien

	B2B-Daten	
	OpenCorporates	**Handelsregister**
Anzahl Unternehmen	86.635	86.635
Korrekte Matches	5.398 (6,2%)	26.371 (30,4%)
Kein Match	81.237 (93,8%)	60.264 (69,6%)

Szenario nicht die vollumfängliche kommerzielle OpenCorporates Datenquelle zur Verfügung. Die Datenquelle enthält insgesamt 203.215.384 Unternehmen, von denen 5.686.794 deutsche Unternehmen sind. Für das illustrative Szenario stand mit 122.184 Unternehmen ein Ausschnitt der Daten zur Verfügung, sodass nicht für jeden B2B-Datensatz ein OpenCorporates Datensatz vorhanden war. Zudem enthalten die CEWE internen B2B-Daten vornehmlich Personengesellschaften, die in der OpenCorporates Datenquelle nicht vorhanden sind. Weiterhin wurde von den Data Scientisten auch die Datenqualität der B2B-Datensätze in Frage gestellt, sodass keine Verwunderung über die geringe Anzahl an TP bei den Data Scientisten bestand. Die Data Scientisten haben positiv hervorgehoben, dass mit verhältnismäßig geringem Aufwand ein Ausschnitt der Datenquelle OpenCorporates integriert werden konnte, um eine erste Evaluation der angereicherten Informationen durchzuführen.

Die Ergebnisse, die den Data Scientisten für die Integration der B2B-Daten mit dem Handelsregister durch den Unternehmen-Matcher präsentiert wurden, sind in Tabelle 7.7 dargestellt. Von den 86.635 B2B Datensätzen wurden 26.371, 30,4% mit einem Handelsregister Unternehmen verknüpft. Es wurden 14.738 potenzielle-Matches generiert. Der Aufwand der manuellen Prüfung der potenziellen-Matches beträgt ca. 15 Personentage. In der Diskussion mit den Data Scientisten relativierte sich die Anzahl der korrekten Matches auch für dieses illustrative Szenario, da viele Personengesellschaften in der B2B Datenquelle existieren und zu diesen kein Treffer in der Handelsregister Datenquelle existiert. Bei der Durchsicht der Ergebnisse ist aufgefallen, dass vor allem mittelständische und große Unternehmen verknüpft worden sind. Eine Einschränkung der B2B Datenquelle auf Kapitalgesellschaften würde die Ergebnisse besser darstellen, ist aber aufgrund eines fehlenden Selektionskriteriums nicht möglich gewesen. Auch in der Diskussion dieser Ergebnisse wurde die schnelle und mit geringem Aufwand verbundene Integration der externen Datenquelle Handelsregister positiv hervorgehoben. Durch den Fokus auf die sichere-Matches kann der manuelle Prüfaufwand für die Proof-of-Concpet Erstellung mit einer externen Datenquellen deutlich reduziert werden. Daher sehen die Data Scientisten Potenzial im Unternehmen-Matcher, um externe Datenquellen mit wenig Aufwand zu evaluieren.

7.5 Fall 4: Oldenburgisch-Ostfriesischer Wasserverband

Abbildung 7.13: OOWV Logo - Oldenburgisch Ostfriesischer Wasserverband (2020)

Der vierte Fall wurde mit dem OOWV durchgeführt. Der OOWV ist eine Körperschaft öffentlichen Rechts und versorgt rund 1 Mio. Menschen, Industrie, Gewerbe und Landwirtschaft im Nordwesten Niedersachsens mit ca. 84 Mio. m3/a Trinkwasser. Hierfür betreibt der OOWV 15 Wasserwerke mit 267 Förderbrunnen zur Grundwasserförderung sowie 46 Kläranlagen für die Abwasseraufbereitung. Der OOWV beteiligt sich zudem an einer Vielzahl von Forschungsvorhaben mit Förderung durch Land, Bund und Europäischer Union (vgl. Oldenburgisch Ostfriesischer Wasserverband, 2020).

Innerhalb der Fallstudie mit dem OOWV sind Business-Intelligence Entwickler und Fachexperten des OOWV für die Datenbereitstellung und die Ergebnisdiskussion beteiligt. Für die Evaluation des Unternehmen-Matcher beim OOWV wurde ein Auszug der internen Großkundendatenbank und ein Auszug der Datenbank mit Wasserrechten aus der niedersächsischen Landesdatenbank für wasserwirtschaftliche Daten[51] zur Verfügung gestellt (siehe Abb. 7.14). Für beide Datenquellen sind in Abbildung 7.14 Beispieldatensätze abgebildet. Zusätzlich wurden in der Fallstudie noch die externen Datenquellen Databyte und Handelsregister genutzt. Ein Beispieldatensatz für die Databyte Datenquelle wurde bereits in Abbildung 7.8 und für das Handelsregister in Abbildung 7.11 dargestellt.

Für die illustrativen Szenarien wurde definiert, dass die internen Großkunden mit den Wasserrechten, Databyte und Handelsregister integriert werden. Die Motivation der Business-Intelligence Entwickler und Fachexperten des OOWV an der Integration der Wasserrechte, Databyte und Handelsregister mit den internen Großkundendaten liegt darin begründet, dass sie mehr Informationen über die Kunden erhalten wollen, um internen Prozesse zu unterstützen und zu verbessern.

7.5.1 Illustratives Szenario

Das erste illustrative Szenario wurde mit der Großkunden Datenquelle des OOWV und der externen Datenquelle Databyte durchgeführt. Für die Großkunden wurden dem Unternehmen-Matcher die Attribute *Geschäftspartner* als Unternehmensname und die Attribute *Adresse_einzeilig_st*, *Straße*, *Hausnummer*, *Ort_Text*, *Ortsteil_Text* und *Land* als Adressdaten übergeben (siehe Abb. 7.14). Für die Datenquelle Databyte wurden dem Unternehmen-Matcher dieselben Attribute übergeben, wie bereits in Abschnitt 7.3.1 beschrieben worden ist.

In Abbildung 7.15 ist der Ablauf des illustrativen Szenarios dargestellt. Dem Unterneh-

[51] `http://www.wasserdaten.niedersachsen.de/cadenza/`

Großkunden

1.159 Unternehmen

Attribut	Inhalt
ID	1206343
Geschäftspartner	OOWV
Adresszeile_0	OOWV
Adresszeile_1	Brake
Adresszeile_2	Georgstr. 4
Adresszeile_3	D-26919 Brake
Adresszeile_4	Deutschland
Adresszeile_5	Null
Adresszeile_6	Null
Adresse_einzeilig	OOWV / D-26919 Brake
Adresse_einzeilig_st	OOWV / Georgstr. 4 / D-26919 Brake
Hausnummer	4
Land	Deutschland
Ort_Text	Brake
Ortsteil_Text	Brake
Postleitzahl	26919
Straße	Georgstr.

Wasserrechte

4.062 Unternehmen

Attribut	Inhalt
waterRightNo	100000778
Bailee	Oldenburgisch-Ostfriesischer Wasserverband
validTo	2039-12-21
validFrom	2009-12-21
legalTitle	Bewilligung
County	Ammerland
localSubDistrict	Westerstede

Abbildung 7.14: Fall 4 OOWV - Beispieldatensätze der Datenquellen Großkunden und Wasserrechte

men-Matcher wurden die 1.159 Unternehmensdatensätze aus der Großkunden Datenquelle des OOWV und die 14.651 aus Oldenburg stammenden Unternehmensdatensätze aus der Databyte Datenquelle übergeben. Der Unternehmen-Matcher hat 100 Matches generiert. Die 100 Matches wurden durch das Regelwerk des Unternehmen-Matcher in 25 sicheren-Matches und 75 potenzielle-Matches unterteilt (siehe Tab. 7.6). Die 25 sicheren-Matches wurden zur Evaluation und Prüfung der *Hypothese 2 Reduktion manueller Prüfaufwand* geprüft. Alle sicheren-Matches sind TP, wodurch Hypothese 2 bestätigt wird. Mit dem Unternehmen-Matcher wurden 100 Matches generiert von denen 45 TP sind, was einer Precision von 45,0% entspricht. Der Anteil der sicheren-Matches, die nicht mehr manuell überprüft werden müssen, an den gesamten TP beträgt 55,6%. Für die übrigen 44,4% der TP wurde ein manueller Prüfaufwand von 0,1 PT erbracht, um die 75 potenziellen-Matches zu überprüfen und weitere 20 TP zu identifizieren (siehe Tab. 7.6).

Das zweite illustrative Szenario wurde mit den Großkunden und der externen Datenquelle Handelsregister durchgeführt. Für die Datenquelle Handelsregister wurden dem Unternehmen-Matcher dieselben Attribute wie im illustrativen Szenario aus Abschnitt 7.4.1 übergeben. Der Ablauf des illustrativen Szenario ist in Abbildung 7.15 dargestellt. Dem Un-

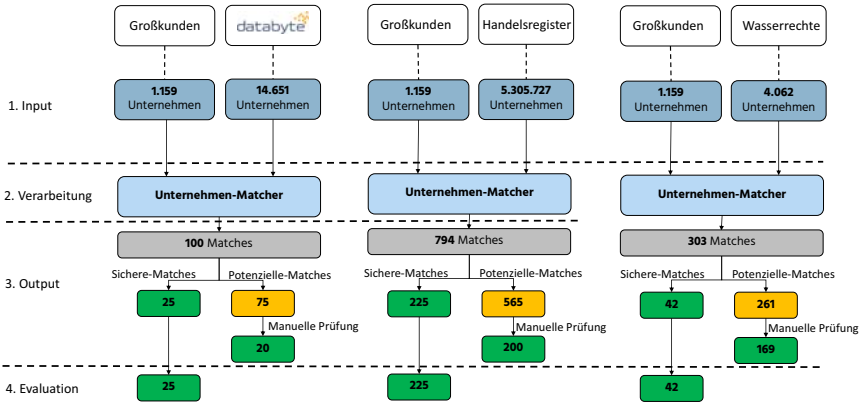


Abbildung 7.15: Fall 4 OOWV - Ablauf der illustrativen Szenarien

ternehmen-Matcher wurden die 1.159 Datensätze aus der Großkunden Datenquelle und die 5.305.727 Datensätze aus der Handelsregister Datenquelle übergeben. Der Unternehmen-Matcher hat 794 Matches generiert. Die 794 Matches wurden durch das Regelwerk des Unternehmen-Matcher in 225 sichere-Matches und 565 potenzielle-Matches unterteilt (siehe Tab. 7.6). Die 225 sicheren-Matches wurden zur Evaluation und Prüfung der *Hypothese 2 Reduktion manueller Prüfaufwand* geprüft. Alle sicheren-Matches sind TP, wodurch Hypothese 2 bestätigt wird. Mit dem Unternehmen-Matcher wurden 794 Matches generiert von denen 425 TP sind, was einer Precision von 53,5% entspricht. Der Anteil der sicheren-Matches, die nicht mehr manuell überprüft werden müssen, an den gesamten TP beträgt 52,9%. Für die übrigen 47,1% der TP wurde ein manueller Prüfaufwand von 0,5 PT erbracht, um die 565 potenziellen-Matches zu überprüfen und weitere 200 TP zu identifizieren (siehe Tab. 7.6).

Im dritten und letzten illustrativen Szenario wurde die Großkunden Datenquelle mit der Wasserrechte Datenquelle verknüpft. Für die Datenquelle Wasserrechte wurden dem Unternehmen-Matcher das Attribut *Bailee* als Unternehmensname und die Attribute *County* und *localSubDistrict* als Adressdaten übergeben. Der Ablauf des illustrativen Szenario ist in Abbildung 7.15 dargestellt. Dem Unternehmen-Matcher wurden die 1.159 Datensätze aus der Großkunden Datenquelle und die 4.062 Datensätze aus der Wasserrechte Datenquelle übergeben. Der Unternehmen-Matcher hat 303 Matches generiert. Die 303 Matches wurden durch das Regelwerk des Unternehmen-Matcher in 42 sichere-Matches und 261 potenzielle-Matches unterteilt (siehe Tab. 7.6). Die 42 sicheren-Matches wurden zur Evaluation und Prüfung der *Hypothese 2 Reduktion manueller Prüfaufwand* geprüft. Alle sicheren-

Tabelle 7.8: Fall 4 OOWV - Ergebnisse des Unternehmen-Matcher in den illustrativen Szenarien
PT = Personentage

	Großkunden		
	Databyte	**Handelsregsiter**	**Wasserrechte**
Matches	100	794	303
TP	45	425	211
Precision	45,0%	53,5%	69,6%
Sichere-Matches	25	225	42
TP sichere-Matches	25	225	42
TP sichere-Matches/TP	55,6%	52,9%	19,9%
Potenzielle-Matches	75	565	261
TP potenzielle-Matches	20	200	169
TP potenzielle-Matches/TP	44,4%	47,1%	80,1%
Manueller Aufwand	0,1 PT	0,5 PT	0,3 PT

Matches sind TP, wodurch Hypothese 2 bestätigt wird. Mit dem Unternehmen-Matcher wurden 303 Matches generiert von denen 211 TP sind, was einer Precision von 69,6% entspricht. Der Anteil der sicheren-Matches, die nicht mehr manuell überprüft werden müssen, an den gesamten TP beträgt 19,9%. Für die übrigen 80,1% der TP wurde ein manueller Prüfaufwand von 0,3 PT erbracht, um die 261 potenziellen-Matches zu überprüfen und weitere 169 TP zu identifizieren (siehe Tab. 7.6).

Um die Güte der Ergebnisse der drei illustrativen Szenarien durch den Unternehmen-Matcher vollständig bewerten zu können, muss bewertet werden, ob die Anzahl der TP allen existierenden korrekten Matches entspricht, oder ob FN in den illustrativen Szenarien existieren. Daher werden die Großkunden, für die kein Match generiert wurde, stichprobenhaft überprüft. Für die Datenquelle Databyte wurde für 1.114 Großkunden (96,1%), für die Datenquelle Handelsregister für 734 Großkunden (63,3%) und für die Datenquelle Wasserrechte für 948 Großkunden (81,8%) kein Match generiert (siehe Tab. 7.9).

Die Databyte Datenquelle enthält ausschließlich Unternehmen mit Sitz in Oldenburg, sodass auch nur Großkunden des OOWV mit Sitz in Oldenburg verknüpft werden können. Die manuelle Einzelfallprüfung aller Oldenburger Großkunden in der Datenquelle des OOWV hat ergeben, dass lediglich drei Großkunden mit keinem Databyte Datensatz verknüpft worden sind. Die manuelle Suche der drei Unternehmen in der Databyte Datenquelle führte zu keinem Treffer, sodass davon ausgegangen wird, dass die drei Unternehmen nicht in der Databyte Datenquelle vorhanden sind und keine FN existieren. Der Unternehmen-Matcher hat alle existierenden Matches identifiziert und verknüpft.

Für das illustrative Szenario der Handelsregister Datenintegration sind 50 Großkundendatensätze des OOWV ausgewählt worden, denen kein Datensatz der Handelsregister Datenquellen zugeordnet worden ist. Für jeden Datensatz wurde manuell geprüft, ob ein korrekter Match in der Handelsregister Datenquelle existiert. Zehn der 50 Datensätze sind FN, da ein korrekter Match in der Handelsregister Datenquelle identifiziert wurde. Für 40 Datensätze wurde kein korrekter Match identifiziert, sodass diese TN sind.

Für das illustrative Szenario zur Integration der Wasserrechte sind 50 Großkundendatensätze des OOWV zufällig ausgewählt worden. Für jeden der 50 Großkunden wurde manuell geprüft, ob ein korrekter Match in der Wasserrechte Datenquelle existiert. Für keinen der 50 Datensätze existiert ein korrekter Match, sodass kein FN identifiziert wurde.

Durch die illustrativen Szenarien mit den Datenquellen des OOWV wurde bewiesen, dass der Unternehmen-Matcher nicht angepasst werden musste und die sicheren Regeln korrekte Treffer geliefert haben. Somit wurden Hypothese 1 und 2 durch diese illustrativen Szenarien bestätigt.

7.5.2 Fokusgruppe

Die Ergebnisse der durchgeführten illustrativen Szenarien innerhalb des OOWV wurden den Business-Intelligence Entwicklern und Fachexperten des OOWV präsentiert und mit diesen diskutiert. Die Ergebnisse sind in Tabelle 7.8 und 7.9 abgebildet.

Tabelle 7.9: Fall OOWV - Gesamtergebnis der illustrativen Szenarien

		Großkunden	
	Databyte	**Handelsregsiter**	**Wasserrechte**
Anzahl Unternehmen	1.159	1.159	1.159
Korrekte Matches	45 (3,9%)	425 (36,7%)	211 (18,2%)
Kein Match	1114 (96,1%)	734 (63,3%)	948 (81,8%)

Der Unternehmen-Matcher hat 45 Unternehmen, 3,9%, mit einem Databyte Unternehmen verknüpft (siehe Tab. 7.8). Der manuelle Aufwand betrug 0,1 PT, um die 75 potenziellen-Matches zu prüfen. Die manuelle Prüfung der kein Match Datensätze ergab, dass sämtliche korrekte Matches identifiziert worden sind.

Die Ergebnisse für die Integration der Großkunden Datenquelle mit dem Handelsregister sind ebenfalls in Tabelle 7.8 dargestellt. Von den 1.159 Unternehmensdatensätzen der Großkunden Datenquelle wurden 425 korrekte Matches zu Unternehmensdatensätzen aus der Handelsregister Datenquelle generiert. Mit einem manuelle Aufwand von 0,5 PT wurden die 565 potenziellen-Matches geprüft.

Die Ergebnisse des Unternehmen-Matcher für die Integration der Großkunden mit den Wasserrechten ist in Tabelle 7.8 dargestellt. Von den 1.159 Unternehmensdatensätzen wurden 211, 18,2%, mit einem Unternehmensdatensatz aus der Wasserrechte Datenquelle verknüpft. Für die manuelle Prüfung der 261 potenzielle-Matches war ein Aufwand von 0,3 PT notwendig.

Im Fokusgruppen-Workshop wurden die Ergebnisse des Unternehmen-Matcher von den Business-Intelligence-Entwicklern und Fachexperten des OOWV positiv bewertet. Vor allem die durch den Unternehmen-Matcher ermöglichten Anwendungsfälle rund um die 360-Grad Kundensicht wurden durch die Fachexperten hervorgehoben. Die angereicherten Informationsmehrwerte wie bspw. Brancheninformationen aus der Datenquelle Databyte für die OOWV Großkunden kann die OOWV internen Analysen verbessern. Weiterhin sehen die Fachexperten und Business-Intelligence-Entwickler den Vorteil des Unternehmen-Matcher in der Datenquellenunabhängigkeit, sodass weitere intern verwendete Datenquellen oder neue externe Datenquelle wie bspw. von HitHorizons[52] zu den OOWV-Großkunden CRM Datensätzen ergänzt werden können. Für die Analyse der existierenden Wasserrechte und welche OOWV Großkunden ein eigenes Wasserrecht besitzen, entsteht dem OOWV laut der Fachexperten aktuell ein großer manueller Aufwand. Durch den Unternehmen-Matcher wird dieser Aufwand deutlich reduziert, sodass die Datenquelle vollständig und kontinuierlich nutzbar gemacht werden kann. Alle Teilnehmer des Fokusgruppen-Workshop sehen im Unternehmen-Matcher eine Lösung mit Einsatzpotenzialen im OOWV, um Zeit und damit Kosten bei der Datenintegration einzusparen sowie neue Datenquellen und damit unternehmerische Mehrwerte zu erschließen. Die Automatisierung durch den Unternehmen-Matcher ermöglicht es den Fachexperten des OOWV kontinuierlich die unternehmerisch relevanten Datenquellen zu integrieren.

7.6 Gesamtanalyse der Fallstudie

In den Abschnitten 7.2, 7.3, 7.4 und 7.5 wurden die einzelnen Fälle der Fallstudie beschrieben. In diesem Abschnitt sollen die gewonnenen Erkenntnisse aus der Evaluation innerhalb der einzelnen Fälle zusammengeführt werden.

Durch die illustrativen Szenarien in den vier Fällen der Fallstudie sollten die in der Planung der Fallstudie aufgestellten Hypothesen untersucht werden. Mit *Hypothese 1 Datenquellenunabhängiger Ansatz* wird geprüft, ob der Unternehmen-Matcher ohne weitere Anpassungen auf neue Datenquellen mit der Realwelt-Entität Unternehmen übertragen werden kann. Mit *Hypothese 2 Reduktion manueller Prüfaufwand* wird geprüft, ob die sicheren Regeln des Unternehmen-Matcher bei neuen Datenquellen ausschließlich TP liefern und damit den manuellen Prüfaufwand reduzieren. In Abbildung 7.16 ist das Gesamtergebnis

[52] `https://www.hithorizons.com/`

der Prüfung der Hypothesen über alle Fälle dargestellt. In allen vier Fällen wurde *Hypothese 1 Datenquellen-unabhängiger Ansatz* und *Hypothese 2 Reduktion manueller Prüfaufwand* bestätigt.

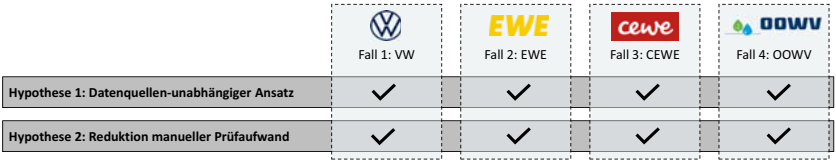


Abbildung 7.16: Gesamtauswertung der Hypothesen der Fallstudie

Mit den in Abbildung 7.16 dargestellten Evaluationsergebnissen wurde gezeigt, dass die prototypische Implementierung des Unternehmen-Matcher dieser Arbeit neue Erkenntnisse zur Forschung zu Datenquellen-unabhängigen und automatisierten Datenintegrationsprozessen beiträgt.

Durch die Fokusgruppen-Workshops innerhalb der einzelnen Fälle konnten weitere qualitative Erkenntnisse zum Einsatzpotenzial des Unternehmen-Matcher im Unternehmenskontext gewonnen werden. Diese qualitativen Erkenntnisse sind in einer Pro und Contra Liste zusammengefasst worden und in Abbildung 7.17 aufgeführt.

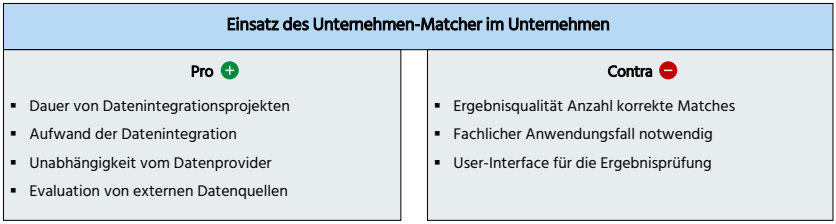


Abbildung 7.17: Gesamtauswertung der Einsatzpotenziale des Unternehmen-Matcher in Unternehmen

Im Folgenden werden die aus den Fokusgruppen-Workshops ermittelten Pro-Erkenntnisse für den Einsatz des Unternehmen-Matcher im Unternehmen erläutert (siehe Abb. 7.17):

Dauer von Datenintegrationsprojekten: Mit dem Einsatz des Unternehmen-Matcher kann der bisher benötigte Zeitraum für Datenintegrationsprojekte reduziert werden. Vor allem für die Kernanwendungen der Unternehmen ist es von hoher Bedeutung, dass diese in einem Datenintegrationsprojekt möglichst schnell in einen konsistenten Zustand überführt

werden.

Aufwand der Datenintegration: Durch die Automatisierung des Datenintegrationsprozesses mit dem Unternehmen-Matcher werden weniger Personalressourcen für ein Datenintegrationsprojekt benötigt. Dadurch wird die Dauer der Projekte verkürzt und die Mitarbeiter können ihre Ressource für die wichtigen, nachgelagerten Analysen der integrierten Daten einsetzen.

Unabhängigkeit vom Datenprovider: Einige Datenprovider, wie Echobot[53], Databyte[54] oder S&P Global[55] bieten ihren Kunden neben dem Kauf der Daten auch die Möglichkeit ihre Daten mit den Daten des Datenproviders zu integrieren. Der Unternehmen-Matcher ist durch den Datenquellen-unabhängigen Ansatz in der Lage, verschiedene Datenquellen mit der Realwelt-Entität Unternehmen zu integrieren, sodass keine Abhängigkeit zu Datenprovidern besteht, um die Daten zu integrieren.

Evaluation externer Datenquellen: Bisher wurden externe Datenquellen oftmals losgelöst von internen Datenquellen evaluiert, da eine Integration für ein Proof-of-Concept zu aufwändig und kostenintensiv ist. Durch den Unternehmen-Matcher wird es Unternehmen ermöglicht effizient verschiedene externe Datenquellen für ein Proof-of-Concept zu integrieren, zu analysieren und zu evaluieren. Denn für ein Proof-of-Concept reichen die sicheren-Matches, die nicht manuell geprüft werden müssen, sodass der manuelle Aufwand reduziert wird.

In den Fokusgruppen-Workshops wurden ebenfalls Contra-Erkenntnisse für den Einsatz des Unternehmen-Matcher im Unternehmen ermittelt, die im Folgenden erläutert werden (siehe Abb. 7.16):

Ergebnisqualität Anzahl korrekte Matches: Die Ergebnisqualität des prototypisch implementierten Unternehmen-Matcher erzielte in den vier Fällen gute und vielversprechende Ergebnisse. Dennoch wurde in den Workshops mit den Experten deutlich, dass die Anzahl der korrekten Matches für einen Einsatz im Unternehmen weiter erhöht werden sollte. Bereits während der illustrativen Szenarien konnten Optimierungspotenziale identifiziert werden, um die Anzahl der korrekten Matches zu erhöhen. Bei der zukünftigen Optimierung des Unternehmen-Matcher muss allerdings berücksichtigt werden, dass er nach wie vor auf neue Datenquellen übertragen werden kann. Vor allem im RL-Prozessschritt Blocking und in der Berücksichtigung weiterer Attribute für den Vergleich bieten sich Optimierungspotenziale für den Unternehmen-Matcher.

[53] `https://www.echobot.de/datacare`
[54] `https://www.databyte.de/`
[55] `https://www.marketplace.spglobal.com/en/solutions`

Fachlicher Anwendungsfall ist notwendig: Das technische Potenzial des Unternehmen-Matcher wurde in allen Fällen bestätigt. Jedoch wurde in den Workshops deutlich, dass der Einsatz und die Einführung des Unternehmen-Matcher ohne den konkreten Bedarf eines fachlichen Anwendungsfalls, wie er in Fall 1 mit VW und Fall 2 mit der EWE TEL vorhanden war, schwierig ist.

Benutzeroberfläche für die Ergebnisprüfung: Die Ergebnisse aus den illustrativen Szenarien mit dem Unternehmen-Matcher wurden über Power-Point- und Excel-Dateien präsentiert. Die manuelle Prüfung der Treffer-Vorschläge erfolgte über das prototypisch angepasste Open-Source Tool RecrdLinkage Annotator[56]. Für den produktiven Einsatz im Unternehmen benötigt der Unternehmen-Matcher eine Benutzeroberfläche mit größerem Funktionsumfang, wie bspw. einer höheren Stabilität und Mehrbenutzer-Funktionalität.

[56] `https://github.com/yamanzein/recordlinkage-annotator`

8 Zusammenfassung und Ausblick

Diese Kapitel fasst die Ergebnisse und Erkenntnisse in Abschnitt 8.1 zusammen. Anschließend wird in Abschnitt 8.2 auf die theoretischen und praktischen Implikationen dieser Arbeit eingegangen. Danach werden in Abschnitt 8.3 die Limitationen dieser Arbeit beschrieben. Abschließend wird in Abschnitt 8.4 der aus den gewonnenen Erkenntnissen und den Limitationen abgeleitete weitere Forschungsbedarf diskutiert.

8.1 Zusammenfassung

Qualitativ hochwertige und vollständige Kundendaten sind die entscheidende Grundlage für die erfolgreiche Digitalisierung und den erfolgreichen Einsatz von KI in Unternehmen. Um diese notwendige Grundlage herzustellen, benötigen Unternehmen den Datenintegrationsprozess. Mit dem Datenintegrationsprozesses werden Duplikate in einzelnen Datenquellen identifiziert. Weiterhin werden mit dem Datenintegrationsprozess die verteilt vorliegenden internen und externen Datenquellen, die keine gemeinsame ID besitzen, integriert, um eine 360-Grad-Sicht über u.a. die Kunden, Produkte und Zulieferer der Unternehmen zu erhalten.

Allerdings bedeutet die Durchführung des Datenintegrationsprozesses für Unternehmen nach wie vor hohen manuellen Aufwand und erfordert hohes Know-how von IT-Fachkräften. Dies ist einer der Hauptgründe, weshalb die existierende Forschung zum Datenintegrationsprozess bisher selten in Unternehmen eingesetzt wird (vgl. Barlaug & Atle Gulla, 2020).

Daher wird in dieser Arbeit der Datenintegrationsprozess End-to-End betrachtet und der Fokus auf die Integration einzelner Realwelt-Entitäten, wie die Realwelt-Entität Unternehmen, gelegt, um den Datenintegrationsprozess für beliebige Datenquellen zu automatisieren.

Um dieses Ziel zu erreichen, wurde zuerst in Kapitel 3 die Teilforschungsfrage 1 *„Welche Record Linkage-Verfahren und -Vorgehensweisen existieren“* durch ein qualitatives Literaturreview beantwortet. Das qualitative Literaturreview lieferte einen Überblick über den State-of-the-Art in der RL-Forschung. Das Literaturreview zeigt, dass für jeden RL-Prozessschritt eine Vielzahl von Algorithmen existiert. Die Mehrheit der RL-Forschung wird mit den häufig verwendeten Benchmark-Datensets durchgeführt und hat das Ziel, die Ergebnisqualität für diese zu optimieren. Die aktuellsten RL-Systeme sind Magellan und JedAI. Der RL-Prozessschritt Data Preparation wird von keiner Forschungsarbeit und keinem RL-System betrachtet und unterstützt. Die Ergebnisse des Literaturreviews und der verwandten Arbeiten aus Kapitel 3 zeigen, dass keine Forschungsarbeit den Fokus auf die Datenquellenunabhängigkeit und die Reduzierung des manuellen Prüfaufwands des RL-Prozesses legt.

In Kapitel 4 wurde der Datenintegrationsprozess durch diese Arbeit um den Prozessschritt

der Datenquellenauswahl erweitert. Die Datenquellenauswahl ist entscheidend für den Erfolg von Data Science Projekten. Daher wird in Kapitel 4 Teilforschungsfrage 2 *„Wie kann bei der Auswahl zu integrierender Datenquellen unterstützt werden“* beantwortet. Die in dieser Arbeit entwickelten Taxonomie zur Unterstützung der Datenquellenauswahl, die mit Experten aus der Wissenschaft und Wirtschaft evaluiert wurde, beantwortet Teilforschungsfrage 2. Mit Hilfe der Taxonomie können Data Scientisten und Fachexperten die relevanten Datenquellen für den jeweiligen Anwendungsfall auswählen, die integriert werden sollen.

In Kapitel 5 wird das Konzept zur Entwicklung von Datenquellen-unabhängigen RL-Prozessen beschrieben. In diesem Kapitel wird Teilforschungsfrage 3 *„Welche Datenintegrationsprobleme existieren in Datenquellen mit der Realwelt-Entität Unternehmen“* beantwortet. Im entwickelten Konzept wird eine Realwelt-Entität fokussiert, um einen Datenquellen-unabhängigen RL-Prozess für diese zu entwickeln. Das Konzept basiert auf der Annahme, dass häufig vorkommende Datenintegrationsprobleme für dieselbe Realwelt-Entität über verschiedene Datenquellen existieren. In dieser Arbeit wurde das Konzept für die Realwelt-Entität Unternehmen angewendet. Dazu wurden mit Hilfe der zuvor entwickelten Taxonomie 18 praxisrelevante Datenquellen, die die Realwelt-Entität enthalten, ausgewählt. Durch die datengetriebeneinduktive Forschung wurde ein Informationsprofil mit den häufig vorkommenden Informationen über ein Unternehmen aus den 18 Datenquellen erstellt. Aus den häufig vorkommenden Informationen wurden dann die häufig vorkommenden Datenintegrationsprobleme für die Realwelt-Entität Unternehmen abgeleitet.

In Kapitel 6 wird ein RL-Prozess für die Integration der Realwelt-Entität Unternehmen, der Unternehmen-Matcher, prototypisch implementiert. Der Unternehmen-Matcher löst die zuvor in Kapitel 5 identifizierten Datenintegrationsprobleme der Realwelt-Entität Unternehmen. Um die Datenquellenunabhängigkeit und die Reduktion des manuellen Prüfaufwands des Unternehmen-Matcher zu gewährleisten, wurden neun Feldexperimente mit den identifizierten praxisrelevanten Datenquellen durchgeführt. Durch die Feldexperimente wurde der Unternehmen-Matcher iterativ entwickelt. Der Unternehmen-Matcher besitzt für die Data Preparation klassische Datenaufbereitungsverfahren, den RechtsformService und den AdressService. Für den Schritt Blocking wurde das Sorted Neighborhood Verfahren ausgewählt und implementiert. Im RL-Prozessschritt Comparsion werden zur Ähnlichkeitsberechnung die Distanzmaße Levenshtein, Jaccard, Jaro-Winkler, Monge-Elkan und Haversine genutzt. Für den RL-Prozessschritt Classification wurde innerhalb der Feldexperimente ein parametrisiertes Regelwerk entwickelt, dass die Matches in sichere-Matches und potenzielle-Matches unterteilt. Da die sicheren-Matches nicht mehr manuell geprüft werden müssen, wird der manuelle Prüfaufwand zur Bewertung der Ergebnisqualität reduziert. Lediglich die potenziellen-Matches müssen manuell geprüft werden. Dieses Kapitel beantwortet Teilforschungsfrage 4 *„Welche Algorithmen und Verfahren eigenen sich, um die Datenintegra-*

tionsprobleme für die Realwelt-Entität Unternehmen zu lösen“. Mit dem Unternehmen-Matcher wurde ein RL-Prozess implementiert, der beliebige Datenquellen mit der Realwelt-Entität Unternehmen integriert und den manuellen Prüfaufwand reduziert.

In Kapitel 7 wird die Evaluation des Unternehmen-Matcher beschrieben. Das Ziel der Evaluation ist es die Hypothesen zu bestätigen, dass der Unternehmen-Matcher beliebige Datenquellen mit der Realwelt-Entität Unternehmen integriert und den manuellen Prüfaufwand des RL-Prozesses für Unternehmen reduziert. Zur Durchführung der Evaluation wurde daher eine Fallstudie mit den Unternehmen Volkswagen AG, EWE TEL GmbH, CEWE Stiftung & Co. KGaA und Oldenburgisch-Ostfriesischen Wasserverband durchgeführt. Jedes Unternehmen stellt einen Fall der gesamten Fallstudie dar. In jedem Fall wurden illustrative Szenarien durchgeführt. In den illustrativen Szenarien wurde der Unternehmen-Matcher auf Datenquellen, die durch das Partnerunternehmen zur Verfügung gestellt wurden, angewendet und die Ergebnisse wurden bewertet. Zusätzlich wurden in jedem Fall Fokusgruppen-Workshops durchgeführt, um die Ergebnisse und die Einsatzpotenziale des Unternehmen-Matcher im jeweiligen Unternehmen qualitativ zu evaluieren. Die gesamte Fallstudie hat die Datenquellenunabhängigkeit und die Reduktion des manuellen Prüfaufwands durch den Unternehmen-Matcher bestätigt. Zudem wurden das Einsatzpotenzial des Unternehmen-Matcher bestätigt, da er die Dauer und den Aufwand von Datenintegrationsprojekten reduziert sowie die Unabhängigkeit und Evaluation von Datenprovidern unterstützt.

8.2 Theoretische und praktische Implikationen

Die in dieser Forschungsarbeit gewonnenen Erkenntnisse liefern theoretische und praktische Implikationen, die im Folgenden beschrieben werden.

Die theoretischen Implikationen der Arbeit beziehen sich auf die Forschung im Bereich Datenintegration und RL. Grundsätzlich leistet diese Forschungsarbeit einen Beitrag und Ansätze, wie die Erkenntnisse der aktuellen Datenintegrations- und RL-Forschung Einsatz in die Praxis finden. Mit dem durchgeführten Literaturreview wurde ein aktueller Überblick über den Stand der Forschung geschaffen der einen Beitrag zur RL-Forschung darstellt. Weiterhin wurde der Datenintegrationsprozess um den Prozessschritt der Datenquellenauswahl erweitert. Zusätzlich wurde mit der Datenquellen-Taxonomie ein Artefakt geschaffen, das den Prozessschritt der Datenquellenauswahl unterstützt. Die Erweiterung des Datenintegrationsprozesses und die Datenquellen-Taxonomie stellen einen Beitrag zur Datenintegrations-Forschung dar. Mit dem entwickelten Konzept zur Entwicklung Datenquellen-unabhängiger RL-Prozesse und der prototypischen Implementierung des Unternehmen-Matcher wurde erstmalig in der RL-Forschung aufgezeigt, wie ein RL-Prozess End-to-End betrachtet und implementiert werden sollte, um die Datenquellenunabhängigkeit und die Reduzierung des manuellen

Prüfaufwands zu gewährleisten. Zudem liefert diese Forschungsarbeit mit dem Konzept für die Entwicklung Datenquellen-unabhängiger RL-Prozesse eine neue Vorgehensweise für die Entwicklung von RL-Systemen, die die Automatisierung der Datenintegration unterstützen.

Durch diese Forschungsarbeit konnten ebenfalls praktische Implikationen für Unternehmen aufgezeigt werden. Bereits der Einsatz des prototypisch implementierten Unternehmen-Matcher zeigte das Potenzial zur Reduktion des manuellen Implementierungs- und Prüfaufwands von IT-Fachkräften sowie der Dauer von Datenintegrationsprojekten. Dadurch wird es Unternehmen ermöglicht neue bisher ungenutzte Datenquellen in ihre Digitalisierungs- und KI-Projekte einzubeziehen und damit einen entscheidenden Wettbewerbsvorteil zu erzielen. Weiterhin bestehen Überlegungen auf Basis der generierten Erkenntnisse dieser Forschungsarbeit den entwickelten Prototyp zu einer marktreifen Software weiterzuentwickeln, um ein zukünftiges KI-Startup zu gründen. Denn durch die Evaluation mit den Partnerunternehmen wurde der Bedarf der Überführung des prototypisch implementierten Unternehmen-Matcher zu einem marktfähigen Produkt identifiziert.

8.3 Limitationen

Diese Forschungsarbeit betrachtet die Datenintegrationsprozessschritte (1) Datenquellenauswahl, (2) Schema Matching und (3) RL. Zur Eingrenzung wurde der Prozessschritt (4) Data Fusion nicht betrachtet, da dieser erst für die nachgelagerte Weiterverarbeitung und Analyse der integrierten Daten notwendig ist.

Für die Entwicklung und Anwendung des Konzepts zur Entwicklung von Datenquellen-unabhängigen RL-Prozessen wurde in dieser Arbeit die Realwelt-Entität Unternehmen fokussiert. Weitere Realwelt-Entitäten wie bspw. Personen oder Produkte wurden in dieser Arbeit nicht betrachtet.

Bei der Entwicklung des Unternehmen-Matcher wurde der Fokus darauf gelegt, einen ersten Prototyp zu entwickeln, der den RL-Prozess End-to-End abbildet und die Datenquellenunabhängigkeit und Reduktion des manuellen Prüfaufwands beweist. Daher haben die entwickelten Algorithmen und Verfahren innerhalb der einzelnen RL-Prozessschritte weiteres Optimierungspotenzial.

8.4 Weiterer Forschungsbedarf

Weitere Forschungsarbeiten sollten untersuchen, wie der Unternehmen-Matcher um Prozesse sowie Algorithmen und Verfahren erweitert werden sollte, um den Datenintegrationsprozessschritt Data Fusion zu unterstützten.

Das Konzept zur Entwicklung von Datenquellen-unabhängigen RL-Prozessen sollte in weiteren Forschungsarbeiten auf Übertragbarkeit überprüft werden. Da in dieser Arbeit die Realwelt-Entität Unternehmen genutzt wurde, sollte das Konzept mit Realwelt-Entitäten wie Person und Produkt, angewandt werden, um die Übertragbarkeit aufzuzeigen. Durch die Anwendung des Konzepts auf andere Realwelt-Entitäten werden neue Datenintegrationsprobleme identifiziert, für die Algorithmen und Verfahren erforscht und entwickelt werden müssen.

Durch die Erweiterung des bisher prototypisch implementierten Unternehmen-Matcher um weitere Realwelt-Entitäten sollten weitere Forschungsarbeiten die Konzeption und Umsetzung von geeigneten Softwarearchitekturen für RL-Systeme untersuchen.

Der Unternehmen-Matcher liefert weiteren Forschungsbedarf entlang der einzelnen RL-Prozessschritte. Weitere Forschungsarbeiten sollten untersuchen, ob weitere Attribute, wie bspw. Branchen, Beschreibungstexte oder Angaben zur Website integriert werden sollten. Für den RL-Prozessschritt Blocking sollte ein Benchmark der aktuell existierenden Algorithmen und Verfahren durchgeführt werden, um diesen zu optimieren. Auch der RL-Prozessschritt Classification bietet Potenzial für weitere Forschungsarbeiten. Das parametrisierte Regelwerk könnte in weiteren Forschungsarbeiten optimiert werden, um mehr sichere-Matches als potenzielle-Matches zu generieren. Durch diese Forschungsarbeit wurden mehrere tausend Trainingsdatensätze über verschiedene Datenquellen erzeugt, sodass der Einsatz von supervised Learning Verfahren erneut untersucht werden sollte. Ebenfalls könnten Active Learning Verfahren für die Menge der potenziellen-Matches untersucht werden.

A Relevante Publikationen der qualitativen Inhaltsanalyse

Tabelle A.1: Identifizierte relevante Publikationen durch die Suchstrategie

ID	Autor	Jahr	Titel
ID_1	Koudas, Nick; Sarawagi, Sunita; Srivastava, Divesh	2006	Record Linkage: Similarity Measures and Algorithms
ID_2	Bhattacharya, Indrajit; Getoor, Lise	2007	Collective entity resolution in relational data
ID_3	Christen, Peter	2007	A Two-Step ClassificationApproach to Unsupervised Record Linkage
ID_4	Elmagarmid, Ahmed K.; Ipeirotis, Panagiotis G.; Verykios, Vassilios S.	2007	Duplicate Record Detection: A Survey
ID_5	Leitão, Luís; Calado, Pável; Weis, Melanie	2007	Structure-based inference of xml similarity for fuzzy duplicate detection
ID_6	Köpcke, Hanna; Rahm, Erhard	2008	Training Selection for Tuning Entity Matching
ID_7	Song, Min; Rudniy, Alex	2008	Detecting Duplicate Biological Entities Using Markov Random Field-Based Edit Distance
ID_8	Zhao, Huimin; Ram, Sudha	2008	Entity matching across heterogeneous data sources: An approach based on constrained cascade generalization
ID_9	Gómez-Bao, Jordi; Larriba-Pey, Josep-L.; Ribes Puig, Josepa	2009	Record linkage performance for large data sets
ID_10	Shu, Liangcai; Long, Bo; Meng, Weiyi	2009	A Latent Topic Model for Complete Entity Resolution
ID_11	Vries, Timothy de; Ke, Hui; Chawla, Sanjay; Christen, Peter	2009	Robust record linkage blocking using suffix arrays
ID_12	DuVall, Scott L.; Kerber, Richard A.; Thomas, Alun	2010	Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators
ID_13	Köpcke, Hanna; Rahm, Erhard	2010	Frameworks for entity matching: A comparison
ID_14	Köpcke, Hanna; Thor, Andreas; Rahm, Erhard	2010	Evaluation of entity resolution approaches on real-world match problems
ID_15	Ramesh Babu, Jagannathan Srinivasan\n	2010	Subsequent Patient Visit Detection in a High Volume OPD using Record Linkage Techniques

ID_16	Conrad, Jack G.; Dozier, Christopher; Molina-Salgado, Hugo; Thomas, Merine; Veeramachaneni, Sriharsha	2011	Public record aggregation using semi-supervised entity resolution
ID_17	Tromp, Miranda; Ravelli, Anita C.; Bonsel, Gouke J.; Hasman, Arie; Reitsma, Johannes B.	2011	Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage
ID_18	Hogan, Aidan; Zimmermann, Antoine; Umbrich, Jürgen; Polleres, Axel; Decker, Stefan	2012	Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora
ID_19	Köpcke, Hanna; Thor, Andreas; Thomas, Stefan; Rahm, Erhard	2012	Tailoring entity resolution for matching product offers
ID_20	Shu, Liangcai; Lin, Can; Meng, Weiyi; Han, Yue; Yu, Clement T.; Smalheiser, Neil R.	2012	A framework for entity resolution with efficient blocking
ID_21	Bakker, Marnix de; Vandic, Damir; Frasincar, Flavius; Kaymak, Uzay	2013	Model Words-Driven Approaches for Duplicate Detection on the Web
ID_22	Bellare, Kedar; Iyengar, Suresh; Parameswaran, Aditya; Rastogi, Vibhor	2013	Active Sampling for Entity Matching with Guarantees
ID_23	Panse, Fabian; van Keulen, Maurice; Ritter, Norbert	2013	Indeterministic Handling of Uncertain Decisions in Deduplication
ID_24	Dalvi, Nilesh; Olteanu, Marian; Raghavan, Manish; Bohannon, Philip	2014	Deduplicating a Places Database
ID_25	Gruenheid, Anja; Dong, Xin Luna; Srivastava, Divesh	2014	Incremental record linkage
ID_26	Jupin, Joseph; Shi, Justin Y.	2014	Identity Tracking in Big Data: Preliminary Research Using In-Memory Data Graph Models for Record Linkage and Probabilistic Signature Hashing for Approximate String Matching in Big Health and Human Services Databases
ID_27	Lee, Sanghoon; Lee, Jongwuk; Hwang, Seung-won	2014	Efficient entity matching using materialized lists
ID_28	Schewe, Klaus-Dieter; Wang, Qing	2014	A theoretical framework for knowledge-based entity resolution
ID_29	Sukharev, Jeffrey; Zhukov, Leonid; Popescul, Alexandrin	2014	Parallel Corpus Approach for Name Matching in Record Linkage

ID_30	Wandelt, Sebastian; Wang, Jiaying; Leser, Ulf; Deng, Dong; Gerdjikov, Stefan; Mishra, Shashwat; Mitankin, Petar; Patil, Manish; Siragusa, Enrico; Tiskin, Alexander; Wang, Wie	2014	State-of-the-art in string similarity search and join
ID_31	Dharavath, Ramesh; Kumar, Chiranjeev	2015	Entity resolution based EM for integrating heterogeneous distributed probabilistic data
ID_32	Enríquez, J. G.; Domínguez Mayo, Francisco José; Escalona Cuaresma, María José; Garcia-Garcia, J.A.; Lee, Vivian; Goto, Masatomo	2015	Entity Identity Reconciliation based Big Data Federation - A MDE approach\n
ID_33	Liu, Hong; Kumar, T. AshwinK.; Thomas, Johnson P.	2015	Cleaning Framework for Big Data - Object Identification and Linkage
ID_34	Lu, Chang; Wang, Hongzhi; Zhang, Yan; Gao, Hong	2015	Euclidean-Based Entity Resolution for Evolving Data
ID_35	Medhat, Doaa; Hassan, Ahmed; Salama, Cherif	2015	A hybrid cross-language name matching technique using novel modified Levenshtein Distance
ID_36	Yang, Yang; Sun, Yizhou; Tang, Jie; Ma, Bo; Li, Juanzi	2015	Entity Matching across Heterogeneous Sources
ID_37	Bhoskar, Umesh S.; Manjaramkar, Arati	2016	Generalized Classificationrules for entity identification
ID_38	Colin Conrad, Naureen Ali, Vlado Keselj, Qigang Gao	2016	ELM: An Extended Logic Matching Method on Record Linkage Analysis of Disparate Databases for Profiling Data Mining
ID_39	Enríquez, J. G.; Blanco, Raquel; Domínguez-Mayo, F. J.; Tuya, Javier; Escalona, M. J.	2016	Towards an MDE-Based Approach to Test Entity Reconciliation Applications\n
ID_40	Gottapu, Ram Deepak; Dagli, Cihan; Ali, Bharami	2016	Entity Resolution Using Convolutional Neural Network
ID_41	Karapiperis, Dimitrios; Gkoulalas-Divanis, Aris; Verykios, Vassilios S.	2016	LSHDB: a parallel and distributed engine for record linkage and similarity search
ID_42	Kim, Kunho; Giles, C. Lee	2016	Financial Entity Record Linkage with Random Forests
ID_43	Kong, Chao; Gao, Ming; Xu, Chen; Qian, Weining; Zhou, Aoying	2016	Entity Matching Across Multiple Heterogeneous Data Sources
ID_44	Mann, Willi; Augsten, Nikolaus; Bouros, Panagiotis	2016	An empirical evaluation of set similarity join techniques

ID_45	Mishra, Sumit; Saha, Sriparna; Mondal, Samrat	2016	An automatic framework for entity matching in bibliographic databases
ID_46	Nentwig, Markus; GroB, Anika; Rahm, Erhard	2016	Holistic Entity Clustering for Linked Data
ID_47	Peled, Olga; Fire, Michael; Rokach, Lior; Elovici, Yuval	2016	Matching entities across online social networks
ID_48	Simonini, Giovanni; Bergamaschi, Sonia; Jagadish, H.V.	2016	BLAST: a Loosely Schema-aware Meta-blocking Approach for Entity Resolution
ID_49	V, Dhivyabharathi G.; Kumaresan, S.	2016	A survey on duplicate record detection in real world data
ID_50	van Dam, Iris; van Ginkel, Gerhard; Kuipers, Wim; Nijenhuis, Nikki; Vandic, Damir; Frasincar, Flavius	2016	Duplicate Detection in Web Shops using LSH to Reduce the Number of Computations
ID_51	Wang, Fei; Wang, Hongbo	2016	Record Linkage Using the Combination of Twice Iterative SVM Training and Controllable Manual Review
ID_52	Bahmani, Zeinab; Bertossi, Leopoldo; Vasiloglou, Nikolaos	2017	ERBlox : Combining matching dependencies with machine learning for entity resolution
ID_53	El-Ghafar, Randa M. Abd; Gheith, Mervat H.; El-Bastawissy, Ali H.; Nasr, Eman S.	2017	Record linkage approaches in big data: A State Of The Art Study
ID_54	Enríquez, J. G.; Domínguez-Mayo, F. J.; Escalona, M. J.; Ross, M.; Staples, G.	2017	Entity reconciliation in big data sources
ID_55	Hogan, Aidan; Zimmermann, Antoine; Umbrich, Jürgen; Polleres, Axel; Decker, Stefan	2017	A novel ensemble learning approach to unsupervised record linkage
ID_56	Ma, Bo; Jiang, Tonghai; Zhou, Xi; Zhao, Fan; Yang, Yating	2017	A Novel Data Integration Framework Based on Unified Concept Model
ID_57	Marple, Tim; Desmarais, Bruce; Young, Kevin L.	2017	Collapsing corporate confusion: Leveraging network structures for effective entity resolution in relational corporate data
ID_58	Blanco, Raquel; Enriquez, Jose G.; Dominguez-Mayo, Francisco J.; Escalona, M. J.; Tuya, Javier	2018	Early Integration Testing for Entity Reconciliation in the Context of Heterogeneous Data Sources
ID_59	Dong, Yongquan; Dragut, E. C.; Meng, Weiyi	2018	Normalization of Duplicate Records from Multiple Sources
ID_60	Fan, Fengfeng; Li, Zhanhuai; Chen, Qun; Chen, Lei	2018	Reasoning about attribute value equivalence in relational data

ID_61	Ferguson, John; Hannigan, Ailish; Stack, Austin	2018	A new computationally efficient algorithm for record linkage with field dependency and missing data imputation
ID_62	Fier, Fabian; Augsten, Nikolaus; Bouros, Panagiotis; Leser, Ulf; Freytag, Johann-Christoph	2018	Set similarity joins on mapreduce
ID_63	Kobayashi, Fumiko; Eram, Aziz; Talburt, John	2018	Entity Resolution Using Logistic Regression as an extension to the Rule-Based Oyster System
ID_64	Kooli, Nihel; Allesiardo, Robin; Pigneul, Erwan	2018	Deep Learning Based Approach for Entity Resolution in Databases
ID_65	Mudgal, Sidharth; Li, Han; Rekatsinas, Theodoros; Doan, AnHai; Park, Youngchoon; Krishnan, Ganesh; Deep, Rohit; Arcaute, Esteban; Raghavendra, Vijay	2018	Deep Learning for Entity Matching: A Design Space Exploration
ID_66	Prabhu, T.; Gnana Dhas, C. Suresh	2018	Improved scalability in mining using ontology record linkage algorithm
ID_67	Schneider, Andrew T.; Mukherjee, Arjun; Dragut, Eduard C.	2018	Leveraging Social Media Signals for Record Linkage
ID_68	Simonini, Giovanni; Papadakis, George; Palpanas, Themis; Bergamaschi, Sonia	2018	Schema-agnostic Progressive Entity Resolution

B Informationsprofil Unternehmen

B.1 Informationsprofil Unternehmensname

	Name	Alternative Namen	Rechtsform
AlphaVantage	x		x
Bureau van Dijk	x		x
CapitalIQ	x	x	x
Crunchbase ODM	x		x
Crunchbase Snapshot	x		x
DataByte	x		x
DeepMatcher Company			
Enigma NASDAQ	x		x
Enigma Nike	x		x
German AI Startups	x		x
GLEIF	x		x
Handelsregister	x		x
OpenCorporates	x	x	x
Owler	x		x
Uscompanylist - Business	x		x
Uscompanylist - Company	x		x
USPTO	x		x
Wikidata	x	x	x
Anzahl	17	3	17

B.2 Informationsprofil Adresse

	Haus	Nummer	Einheit	Ebene	Straße	Ortsteil	Stadt	Staat	Land	PLZ	Latitude	Longitude
AlphaVantage		x			x		x	x	x	x		
Bureau van Dijk		x			x		x	x	x	x		
CapitalIQ	x	x	x	x	x	x	x	x	x	x		
Crunchbase ODM							x	x	x			
Crunchbase Snapshot						x	x	x	x			
DataByte		x			x		x	x	x	x		
DeepMatcher Company												
Enigma NASDAQ												
Enigma Nike		x	x	x	x		x	x	x	x		
German AI Startups							x					
GLEIF	x	x	x	x	x		x	x	x	x		
Handelsregister		x			x		x	x	x	x		
OpenCorporates		x	x	x	x		x	x	x	x		
Owler		x	x		x		x	x	x	x		
Uscompanylist - Business		x	x		x		x	x		x	x	x
Uscompanylist - Company		x	x		x	x	x	x		x	x	x
USPTO							x	x	x			
Wikidata		x			x		x		x	x		
Anzahl	2	12	7	4	12	3	16	14	13	12	2	2

B.3 Informationsprofil weitere Informationen

	Branche	Beschreibungstext	Homepage
AlphaVantage	x	x	
Bureau van Dijk		x	x
CapitalIQ	x	x	x
Crunchbase ODM		x	x
Crunchbase Snapshot		x	x
DataByte	x	x	x
DeepMatcher Company		x	
Enigma NASDAQ	x		
Enigma Nike			
German AI Startups		x	
GLEIF	x		
Handelsregister			
OpenCorporates	x		
Owler	x	x	x
Uscompanylist - Business	x		x
Uscompanylist - Company	x		x
USPTO			
Wikidata	x	x	x
Anzahl	10	12	9

C RechtsformService

C.1 Deutsche Rechtsformen für das Feldexperimente

Rechtsform
AG & Co. KG
AG & Co. KGaA
AG & Co. OHG
Aktiengesellschaft
EG
EK
EV
GbR
gGmbH
GmbH
GmbH & Co. KG
GmbH & Co. KGaA
GmbH & Co. OHG
KG
KGaA
Limited & Co. KG
No legal form
OHG
PartG
SE
SE & Co. KG
SE & Co. OHG
SE Co. KGaA
Stiftung
Stiftung & Co. KG
UG
UG & Co. KG
VVaG

Literaturverzeichnis

Abbasi, A., Sarker, S. & Chiang, R. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, *17* (2), I–XXXII. doi: 10.17705/1jais.00423

AlphaVantage. (o. J.). Zugriff auf `https://www.alphavantage.co/`

Appanion. (o. J.). Zugriff auf `https://de.appanion.com/startups`

Assaf, A., Senart, A. & Troncy, R. (2016). Towards An Objective Assessment Framework for Linked Data Quality. *International Journal on Semantic Web and Information Systems*, *12* (3), 111–133. doi: 10.4018/IJSWIS.2016070104

Barlaug, N. (2020). Tailoring Entity Matching for Industrial Settings. In M. d'Aquin, S. Dietze, C. Hauff, E. Curry & P. Cudre Mauroux (Hrsg.), *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (S. 3217–3220). New York, NY, USA: ACM. doi: 10.1145/3340531.3418514

Barlaug, N. & Atle Gulla, J. (2020). Neural Networks for Entity Matching: A Survey.

Behnen, P., Kruse, F. & Marx Gómez, J. (2021). Enhancement of Record Linkage by Using Attributes containing Natural Language Text. In A. Martin et al. (Hrsg.), *AAAI-MAKE 2021 Combining Machine Learning and Knowledge Engineering* (S. 1–14).

Blanco, R., Enriquez, J. G., Dominguez-Mayo, F. J., Escalona, M. J. & Tuya, J. (2018). Early Integration Testing for Entity Reconciliation in the Context of Heterogeneous Data Sources. *IEEE Transactions on Reliability*, 1–19. doi: 10.1109/TR.2018.2809866

Blazquez, D. & Domenech, J. (2018a). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*, 99–113. doi: 10.1016/j.techfore.2017.07.027

Blazquez, D. & Domenech, J. (2018b). WEB DATA MINING FOR MONITORING BUSINESS EXPORT ORIENTATION. *Technological and Economic Development of Economy*, *24* (2), 406–428. doi: 10.3846/20294913.2016.1213193

Bleiholder, J. & Schmid, J. (2018). Datenintegration und Deduplizierung. In K. Hildebrand, M. Gebauer, H. Hinrichs & M. Mielke (Hrsg.), *Daten- und Informationsqualität: Auf dem Weg zur Information Excellence* (S. 121–140). Wiesbaden: Springer Fachmedien Wiesbaden. doi: 10.1007/978-3-658-21994-9{\textunderscore}7

Bureau van Dijk. (o. J.). Zugriff auf `https://www.bvdinfo.com/de-de/`

Capital IQ. (o. J.). Zugriff auf https://www.spglobal.com/marketintelligence/en/

Cato, P. (2016). *Einflüsse auf den Implementierungserfolg von Big Data Systemen* (Dissertation). Verlag Dr. Kovač.

CEWE. (o. J.). *CEWE Fotoservice.* Zugriff am 11.02.2022 auf https://www.cewe.de/index.html

Christen, P. (2005). Probabilistic Data Generation for Deduplication and Data Linkage. In M. Gallagher, J. P. Hogan & F. Maire (Hrsg.), *Intelligent Data Engineering and Automated Learning - IDEAL 2005* (S. 109–116). Berlin, Heidelberg: Springer Berlin Heidelberg.

Christen, P. (Hrsg.). (2012a). *Data Matching.* Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-31164-2

Christen, P. (2012b). A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering*, *24* (9), 1537–1555. doi: 10.1109/TKDE.2011.127

Christen, P. (2019). Data Linkage: The Big Picture. *Harvard Data Science Review*, *1* (2). Zugriff auf https://hdsr.mitpress.mit.edu/pub/8fm8lo1e doi: 10.1162/99608f92.84deb5c4

Christen, P., Churches, T. & Hegland, M. (2004). Febrl – A Parallel Open Source Data Linkage System. In H. Dai, R. Srikant & C. Zhang (Hrsg.), *Advances in Knowledge Discovery and Data Mining* (S. 638–647). Berlin, Heidelberg: Springer Berlin Heidelberg.

Christen, P. & Winkler, W. E. (2016). Record Linkage. In C. Sammut & G. I. Webb (Hrsg.), *Encyclopedia of Machine Learning and Data Mining* (Bd. 19, S. 1–10). Boston, MA: Springer US. doi: 10.1007/978-1-4899-7502-7{\textunderscore}712-1

Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G. & Stefanidis, K. (2021). An Overview of End-to-End Entity Resolution for Big Data. *ACM Computing Surveys*, *53* (6), 1–42. doi: 10.1145/3418896

Comber, S. & Arribas-Bel, D. (2019). Machine learning innovations in address matching: A practical comparison of word2vec and CRFs. *Transactions in GIS*, *23* (2), 334–348. doi: 10.1111/tgis.12522

Conrad, J. G., Dozier, C., Molina-Salgado, H., Thomas, M. & Veeramachaneni, S. (2011). Public record aggregation using semi-supervised entity resolution. In T. van Engers & K. Ashley (Hrsg.), *Proceedings of the 13th International Conference on Artificial Intelligence and Law - ICAIL '11* (S. 239–248). New York, New York, USA: ACM Press. doi: 10.1145/2018358.2018392

Crunchbase Open Data Map. (o. J.). Zugriff auf https://data.crunchbase.com/docs/open-data-map

Crunchbase Snapshot 2013. (o. J.). Zugriff auf https://data.crunchbase.com/docs/2013-snapshot

Cuffe, J. & Goldschlag, N. (2018). Squeezing More Out of Your Data: Business Record Linkage with Python..

Dalvi, N., Olteanu, M., Raghavan, M. & Bohannon, P. (2014). Deduplicating a places database. In C.-W. Chung, A. Broder, K. Shim & T. Suel (Hrsg.), *Proceedings of the 23rd international conference on World wide web - WWW '14* (S. 409–418). New York, New York, USA: ACM Press. doi: 10.1145/2566486.2568034

Databyte. (o. J.). Zugriff auf https://www.databyte.de/

DeepMatcher Company. (o. J.). Zugriff auf https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md

Deloitte. (2018). *Szenarioanalyse: So managen Sie Unsicherheiten planvoll: Wie künstliche Intelligenz und menschliche Fantasie zukunftssichere Entscheidungen ermöglichen.* Zugriff am 14.10.2018 auf https://www2.deloitte.com/de/de/pages/trends/szenarioanalysen.html

de Vries, T., Ke, H., Chawla, S. & Christen, P. (2009). Robust record linkage blocking using suffix arrays. In D. Cheung, I.-Y. Song, W. Chu, X. Hu & J. Lin (Hrsg.), *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09* (S. 305). New York, New York, USA: ACM Press. doi: 10.1145/1645953.1645994

Doan, A. (2017). *How-To Guide to Entity Matching.* Zugriff am 04.05.2020 auf http://pradap-www.cs.wisc.edu/magellan/how-to-guide/how_to_guide_magellan.pdf

Doan, A., Halevy, A. & Ives, Z. G. (2012). *Principles of data integration.* Amsterdam: Elsevier Morgan Kaufmann.

Doan, A., Konda, P., Suganthan, P., Ardalan, A., Ballard, J. R., Das, S., ... Zhang, H. (2018). Toward a System Building Agenda for Data Integration (and Data Science). *IEEE Data Eng. Bull.*, *41* (2), 35–46.

Doan, A., Konda, P., Suganthan G. C., P., Govind, Y., Paulsen, D., Chandrasekhar, K., ... Christie, M. (2020). Magellan: Toward Building Ecosystems of Entity Matching Solutions. *Communications of the ACM*, *63* (8), 83–91. doi: 10.1145/3405476

Doan, A., Suganthan, G. C. P., Zhang, H., Ardalan, A., Ballard, J., Das, S., ... Paulson, E. (2017). Human-in-the-Loop Challenges for Entity Matching. In Unknown (Hrsg.), *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics - HILDA'17* (S. 1–6). New York, New York, USA: ACM Press. doi: 10.1145/3077257.3077268

Dong, X. L. & Rekatsinas, T. (2018). Data Integration and Machine Learning. In G. Das, C. Jermaine & P. Bernstein (Hrsg.), *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18* (S. 1645–1650). New York, New York, USA: ACM Press. doi: 10.1145/3183713.3197387

Dong, X. L., Saha, B. & Srivastava, D. (2012). Less is more. *Proceedings of the VLDB Endowment*, *6* (2), 37–48. doi: 10.14778/2535568.2448938

Dong, X. L. & Srivastava, D. (2013). Big data integration. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)* (S. 1245–1248). IEEE. doi: 10.1109/ICDE.2013.6544914

Dong, X. L. & Srivastava, D. (2015). Big Data Integration. *Synthesis Lectures on Data Management*, *7* (1), 1–198. doi: 10.2200/S00578ED1V01Y201404DTM040

Dunn, H. L. (1946). Record linkage. *American Journal of Public Health and the Nations Health*, *36* (12), 1412–1416.

Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M. & Tang, N. (2018). Distributed representations of tuples for entity resolution. *Proceedings of the VLDB Endowment*, *11* (11), 1454–1467. doi: 10.14778/3236187.3236198

El-Ghafar, R. M. A., Gheith, M. H., El-Bastawissy, A. H. & Nasr, E. S. (2017). Record linkage approaches in big data: A state of art study. In *2017 13th International Computer Engineering Conference (ICENCO)* (S. 224–230). IEEE. doi: 10.1109/ICENCO.2017.8289792

Elmagarmid, A. K., Ipeirotis, P. G. & Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *19* (1), 1–16. doi: 10.1109/TKDE.2007.250581

Enigma NASDAQ. (o. J.). *: zum Zeitpunkt des Datenabrufs noch kostenfrei und offen.* Zugriff auf `https://enigma.com/`

Enigma Nike. (o. J.). *: zum Zeitpunkt des Datenabrufs noch kostenfrei und offen.* Zugriff auf `https://enigma.com/`

Enríquez, J. G., Blanco, R., Domínguez-Mayo, F. J., Tuya, J. & Escalona, M. J. (2016). Towards an MDE-based approach to test entity reconciliation applications. In T. Vos, S. Eldh

& W. Prasetya (Hrsg.), *Proceedings of the 7th International Workshop on Automating Test Case Design, Selection, and Evaluation - A-TEST 2016* (S. 74–77). New York, New York, USA: ACM Press. doi: 10.1145/2994291.2994303

Enríquez, J. G., Domínguez Mayo, F. J., Escalona Cuaresma, M. J., Garcia-Garcia, J., Lee, V. & Goto, M. (2015). Entity Identity Reconciliation based Big Data Federation - A MDE approach.

Enríquez, J. G., Domínguez-Mayo, F. J., Escalona, M. J., Ross, M. & Staples, G. (2017). Entity reconciliation in big data sources: A systematic mapping study. *Expert Systems with Applications*, *80*, 14–27. doi: 10.1016/j.eswa.2017.03.010

EWE. (o. J.). *Strom, Erdgas, Internet, Mobilfunk aus einer Hand.* Zugriff auf `https://www.ewe.de/`

Fasel, D. & Meier, A. (Hrsg.). (2016). *Big Data: Grundlagen, Systeme und Nutzungspotenziale*. Wiesbaden: Springer Vieweg. Zugriff auf `http://dx.doi.org/10.1007/978-3-658-11589-0` doi: 10.1007/978-3-658-11589-0

Fellegi, I. P. & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *64* (328), 1183–1210. doi: 10.1080/01621459.1969.10501049

Ferguson, J., Hannigan, A. & Stack, A. (2018). A new computationally efficient algorithm for record linkage with field dependency and missing data imputation. *International journal of medical informatics*, *109*, 70–75. doi: 10.1016/j.ijmedinf.2017.10.021

Fier, F., Augsten, N., Bouros, P., Leser, U. & Freytag, J.-C. (2018). Set similarity joins on mapreduce. *Proceedings of the VLDB Endowment*, *11* (10), 1110–1122. doi: 10.14778/3231751.3231760

Gartner. (2021). *Gartner Report - Gartner Magic Quadrant for Data Integration Tools 2021.* Zugriff auf `https://www.talend.com/de/lp/gartner-magic-quadrant-data-integration/`

GLEIF. (o. J.). Zugriff auf `https://www.gleif.org/en`

Gluchowski, P. & Chamoni, P. (2016). *Analytische Informationssysteme*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-662-47763-2

Golshan, B., Halevy, A., Mihaila, G. & Tan, W.-C. (2017). Data Integration: After the Teenage Years. In J. van den Bussche, F. Geerts & E. Sallinger (Hrsg.), *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - PODS '17* (S. 101–106). New York, New York, USA: ACM Press. doi: 10.1145/3034786.3056124

Gómez-Bao, J., Larriba-Pey, J.-L. & Ribes Puig, J. (2009). Record linkage performance for large data sets. In V. Muntés-Mulero & J. Nin (Hrsg.), *Proceeding of the ACM first international workshop on Privacy and anonymity for very large databases - PAVLAD '09.* New York, New York, USA: ACM Press. doi: 10.1145/1651449.1651453

González Enríquez, J. (2017). *A model-driven engineering approach for the uniquely identity reconciliation of heterogeneous data sources* (Dissertation). Universidad de Sevilla, Sevilla.

Gottapu, R. D., Dagli, C. & Ali, B. (2016). Entity Resolution Using Convolutional Neural Network. *Procedia Computer Science*, *95*, 153–158. doi: 10.1016/j.procs.2016.09.306

Govind, Y., Konda, P., Suganthan G C, P., Martinkus, P., Nagarajan, P., Soundararajan, A., ... Doan, A. (2019). Entity Matching Meets Data Science: A Progress Report from the Magellan Project.

Govind, Y., Sun, M., Paulson, E., Nagarajan, P., C., P. S. G., Doan, A., ... Carter, M. (2018). Cloudmatcher: a hands-off cloud/crowd service for entity matching. *Proceedings of the VLDB Endowment*, *11* (12), 2042–2045. doi: 10.14778/3229863.3236255

Grover, V. & Lyytinen, K. (2015). New State of Play in Information Systems Research: The Push to the Edges. *MIS Quarterly*, *39* (2), 271–296. doi: 10.25300/MISQ/2015/39.2.01

Gschwind, T., Miksovic, C., Minder, J., Mirylenka, K. & Scotton, P. (2019). Fast Record Linkage for Company Entities. In *2019 IEEE International Conference on Big Data (Big Data)* (S. 623–630). IEEE. doi: 10.1109/BigData47090.2019.9006095

Han, J., Kamber, M. & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed. Aufl.). Amsterdam and Boston: Elsevier/Morgan Kaufmann. Zugriff auf `https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=377411`

Handelsregister. (o. J.). Zugriff auf `https://offeneregister.de/`

Heinrich, C. & Stühler, G. (2018). Die Digitale Wertschöpfungskette: Künstliche Intelligenz im Einkauf und Supply Chain Management. In *Fallstudien zur Digitalen Transformation : Case Studies für die Lehre und praktische Anwendung* (S. 77–88). Wiesbaden, Germany: Springer Gabler.

Hildebrand, K., Gebauer, M., Hinrichs, H. & Mielke, M. (Hrsg.). (2018). *Daten- und Informationsqualität: Auf dem Weg zur Information Excellence* (4. Aufl.). Wiesbaden: Springer Fachmedien Wiesbaden.

Hildebrand, K., Gebauer, M. & Mielke, M. (2021). *Daten- und Informationsqualität* (5. Aufl.). Wiesbaden: Springer Fachmedien Wiesbaden. doi: 10.1007/978-3-658-30991-6

Iffert, L. & Derwisch, S. (2018). *Datenhandelsplätze: Wissenstankstellen des 21. Jahrhunderts.* Zugriff am 06.11.2019 auf `https://barc.de/Artikel/datenhandelsplatze-wissenstankstellen-des-21-jahrhunderts`

Jupin, J. & Shi, J. Y. (2014). Identity Tracking in Big Data: Preliminary Research Using In-Memory Data Graph Models for Record Linkage and Probabilistic Signature Hashing for Approximate String Matching in Big Health and Human Services Databases. In A. Chin, J. Zhan, W. Ding, J. Wu, W. Xu & F. Wang (Hrsg.), *Proceedings of the 2014 International Conference on Big Data Science and Computing - BigDataScience '14* (S. 1–8). New York, New York, USA: ACM Press. doi: 10.1145/2640087.2644170

Karapiperis, D., Gkoulalas-Divanis, A. & Verykios, V. S. (2016). LSHDB: a parallel and distributed engine for record linkage and similarity search. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (S. 1–4). IEEE. doi: 10.1109/ICDMW.2016.7867099

Kessler, R. & Marx Gómez, J. (2020). Implikationen von Machine Learning auf das Datenmanagement in Unternehmen. *HMD Praxis der Wirtschaftsinformatik*, *57* (1), 89–105. doi: 10.1365/s40702-020-00585-z

Kobayashi, F., Eram, A. & Talburt, J. (2018). Entity Resolution Using Logistic Regression as an extension to the Rule-Based Oyster System. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (S. 146–151). IEEE. doi: 10.1109/MIPR.2018.00033

Kölbl, L., Mühlroth, C., Wiser, F., Grottke, M. & Durst, C. (2019). Big Data im Innovationsmanagement: Wie Machine Learning die Suche nach Trends und Technologien revolutioniert. *HMD Praxis der Wirtschaftsinformatik.* doi: 10.1365/s40702-019-00528-3

Konda, P., Naughton, J., Prasad, S., Krishnan, G., Deep, R., Raghavendra, V., ... Zhang, H. (2016a). Magellan: Toward Building Entity Matching Management Systems. *Proceedings of the VLDB Endowment*, *9* (12), 1197–1208. doi: 10.14778/3007263.3007314

Konda, P., Naughton, J., Prasad, S., Krishnan, G., Deep, R., Raghavendra, V., ... Zhang, H. (2016b). Magellan: Toward Building Entity Matching Management Systems: Technical Report.

Konda, P., Subramanian Seshadri, S., Segarra, E., Hueth, B. & Doan, A. (2019). Executing Entity Matching End to End: A Case Study. In Melanie Herschel, Helena Galhardas, Berthold Reinwald, Irini Fundulaki, Carsten Binnig & Zoi Kaoudi (Hrsg.), *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019* (S. 489–500). OpenProceedings.org. Zugriff auf `https://doi.org/10.5441/002/edbt.2019.45` doi: 10.5441/002/edbt.2019.45

Konda, P., Zhang, H., Naughton, J., Prasad, S., Krishnan, G., Deep, R., ... Panahi, F. (2018). Magellan: Toward Building Entity Matching Management Systems. *ACM SIGMOD Record*, *47* (1), 33–40. doi: 10.1145/3277006.3277015

Kong, C., Gao, M., Xu, C., Qian, W. & Zhou, A. (2016). Entity Matching Across Multiple Heterogeneous Data Sources. In S. B. Navathe, W. Wu, S. Shekhar, X. Du, X. S. Wang & H. Xiong (Hrsg.), *Database Systems for Advanced Applications* (Bd. 9642, S. 133–146). Cham: Springer International Publishing. doi: 10.1007/978-3-319-32025-0{\textunderscore}9

Kooli, N., Allesiardo, R. & Pigneul, E. (2018). Deep Learning Based Approach for Entity Resolution in Databases. In N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham & B. Trawiński (Hrsg.), *Intelligent Information and Database Systems* (Bd. 10752, S. 3–12). Cham: Springer International Publishing. doi: 10.1007/978-3-319-75420-8{\textunderscore}1

Köpcke, H. (2014). *Object Matching on real-world problems* (Dissertation). Universität Leipzig, Leipzig.

Köpcke, H. & Rahm, E. (2008). Training Selection for Tuning Entity Matching.

Köpcke, H. & Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, *69* (2), 197–210. doi: 10.1016/j.datak.2009.10.003

Köpcke, H., Thor, A. & Rahm, E. (2010). Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, *3* (1-2), 484–493. doi: 10.14778/1920841.1920904

Köpcke, H., Thor, A., Thomas, S. & Rahm, E. (2012). Tailoring entity resolution for matching product offers. In E. Rundensteiner, V. Markl, I. Manolescu, S. Amer-Yahia, F. Naumann & I. Ari (Hrsg.), *Proceedings of the 15th International Conference on Extending Database Technology - EDBT '12* (S. 545). New York, New York, USA: ACM Press. doi: 10.1145/2247596.2247662

Koudas, N., Sarawagi, S. & Srivastava, D. (2006). Record Linkage Similarity Measures and Algorithms.

Koumarelas, I., Jiang, L. & Naumann, F. (2020). Data Preparation for Duplicate Detection. *Journal of Data and Information Quality (JDIQ)*, *1* (1), 1–24.

Koumarelas, l., Papenbrock, T. & Naumann, F. (2020). MDedup: Duplicate Detection with Matching Dependencies. *Proceedings of the VLDB Endowment*, *13* (5), 712–725. doi: 10.14778/3377369.3377379

Kruse, F. (2019). Towards a Record Linkage Layer to Support Big Data Integration. In W. Abramowicz & R. Corchuelo (Hrsg.), *Business Information Systems Workshops* (Bd. 373, S. 625–636). Cham: Springer International Publishing. doi: 10.1007/978-3-030-36691-9{\textunderscore}52

Kruse, F., Awick, J.-P., Marx Gómez, J. & Loos, P. (2021). Developing a Legal Form Classification and Extraction Approach for Company Entity Matching. *Business Information Systems*, 13–26. doi: 10.52825/bis.v1i.44

Kruse, F., Dmitriyev, V. & Marx Gómez, J. (2018). Building a Connection Between Decision Maker and Data-Driven Decision Process. *Archives of Data Science, Series A (Online First)*, *4* (1), 16 S. online. doi: 10.5445/KSP/1000085951/03

Kruse, F., Hassan, A. P., Awick, J.-P. & Marx Gómez, J. (2020). A Qualitative Literature Review on Linkage Techniques for Data Integration. In Tung Bui (Hrsg.), *53nd Hawaii International Conference on System Sciences, HICSS 2020, Grand Wailea, Maui, Hawaii, USA, January 7-10, 2020* (S. 1063–1073). ScholarSpace / AIS Electronic Library (AISeL). Zugriff auf `http://hdl.handle.net/10125/63871`

Kruse, F., Schröer, C. & Marx Gómez, J. (2021). Data Source Selection Support in the Big Data Integration Process - Towards a Taxonomy. In F. Ahlemann, R. Schütte & S. Stieglitz (Hrsg.), *Internationale Tagung Wirtschaftsinformatik (WI)*.

Laudon, K. C., Laudon, J. P. & Schoder, D. (2016). *Wirtschaftsinformatik: Eine Einführung* (3., vollständig überarbeitete Auflage Aufl.). Hallbergmoos: Pearson. Zugriff auf `http://lib.myilibrary.com?id=838570`

Leitão, L., Calado, P. & Weis, M. (2007). Structure-based inference of xml similarity for fuzzy duplicate detection. In M. J. Silva et al. (Hrsg.), *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07.* New York, New York, USA: ACM Press. doi: 10.1145/1321440.1321483

Li, L., Li, J., Wang, H. & Gao, H. (2011). Context-based entity description rule for entity resolution. In I. Ounis, I. Ruthven & C. Macdonald (Hrsg.), *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11* (S. 1725). New York, New York, USA: ACM Press. doi: 10.1145/2063576.2063825

Lin, Y., Wang, H., Li, J. & Gao, H. (2016). Data Source Selection for Information Integration in Big Data Era.

Liu, H., Kumar, T. A. & Thomas, J. P. (2015). Cleaning Framework for Big Data - Object Identification and Linkage. In *2015 IEEE International Congress on Big Data* (S. 215–221). IEEE. doi: 10.1109/BigDataCongress.2015.38

Maass, W., Parsons, J., Purao, S., Storey, V. C. & Woo, C. (2018). Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research. *Journal of the Association for Information Systems*, 1253–1273. doi: 10.17705/1jais.00526

Marple, T., Desmarais, B. & Young, K. L. (2017). Collapsing corporate confusion: Leveraging network structures for effective entity resolution in relational corporate data. In *2017 IEEE International Conference on Big Data (Big Data)* (S. 2637–2643). IEEE. doi: 10.1109/BigData.2017.8258224

Mayring, P. (2000). Qualitative Content Analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1* (2). Zugriff auf `http://www.qualitative-research.net/index.php/fqs/article/view/1089`

Mayring, P. (2014). *Qualitative content analysis: theoretical foundation, basic procedures and software solution*.

Mayring, P. (2015). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (12., überarb. Aufl. Aufl.). Weinheim: Beltz.

Medhat, D., Hassan, A. & Salama, C. (2015). A hybrid cross-language name matching technique using novel modified Levenshtein Distance. In *2015 Tenth International Conference on Computer Engineering & Systems (ICCES)* (S. 204–209). IEEE. doi: 10.1109/ICCES.2015.7393046

Mertens, P. (2019). *Wirtschaftsinformatik*. Zugriff am 03.11.2019 auf `http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/uebergreifendes/Disziplinen%20der%20WI/Wirtschaftsinformatik`

Meyn, A., Sock, W., Adan, J. & Stüben, J. (2019). Daten schaffen Business Value: Data-Management als Unternehmensfundament. *BI-Spektrum : Fachzeitschrift für Business Intelligence und Data Warehousing*, *14* (3), 12–16.

Mishra, S., Saha, S. & Mondal, S. (2016). An automatic framework for entity matching in bibliographic databases. In *2016 IEEE Congress on Evolutionary Computation (CEC)* (S. 271–278). IEEE. doi: 10.1109/CEC.2016.7743805

Monge, A. E. & Elkan, C. P. (1996). The Field Matching Problem: Algorithms and Applications. In *KDD*.

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., … Raghavendra, V. (2018). Deep Learning for Entity Matching. In G. Das, C. Jermaine & P. Bernstein (Hrsg.), *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18* (S. 19–34). New York, New York, USA: ACM Press. doi: 10.1145/3183713.3196926

Mühlroth, C. & Grottke, M. (2018). A systematic literature review of mining weak signals and trends for corporate foresight. *Journal of Business Economics*, *37* (6), 3. doi: 10.1007/s11573-018-0898-4

Müllerleile, T. (Hrsg.). (2019). *Prozessakzeptanz*. Wiesbaden: Springer Fachmedien Wiesbaden. doi: 10.1007/978-3-658-27103-9

Nentwig, M., GroB, A. & Rahm, E. (2016). Holistic Entity Clustering for Linked Data. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (S. 194–201). IEEE. doi: 10.1109/ICDMW.2016.0035

Newcombe, H. B., Kennedy, J. M., Axford, S. J. & James, A. P. (1959). Automatic linkage of vital records. *Science (New York, N.Y.)*, *130* (3381), 954–959. doi: 10.1126/science.130.3381.954

Nickerson, R. C., Varshney, U. & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22* (3), 336–359. doi: 10.1057/ejis.2012.26

Oldenburgisch Ostfriesischer Wasserverband. (2020). *Unser Wasser. Unser Element: Geschäftsbericht 2020*. Zugriff auf `https://www.oowv.de/fileadmin/user_upload/oowv/content_pdf/geschaeftsbericht/OOWV_GB_2020_final.pdf`

OpenCorporates. (o. J.). Zugriff auf `https://opencorporates.com/`

Owler. (o. J.). Zugriff auf `https://corp.owler.com/`

Papadakis, G., Alexiou, G., Papastefanatos, G. & Koutrika, G. (2015). Schema-agnostic vs schema-based configurations for blocking methods on homogeneous data. *Proceedings of the VLDB Endowment*, *9* (4), 312–323. doi: 10.14778/2856318.2856326

Papadakis, G., Ioannou, E. & Palpanas, T. (2020). Entity Resolution: Past, Present and Yet-to-Come.

doi: 10.5441/002/edbt.2020.85

Papadakis, G., Ioannou, E., Thanos, E. & Palpanas, T. (2021). The Four Generations of Entity Resolution. *Synthesis Lectures on Data Management*, *16* (2), 1–170. doi: 10.2200/S01067ED1V01Y202012DTM064

Papadakis, G., Mandilaras, G., Gagliardelli, L., Simonini, G., Thanos, E., Giannakopoulos, G., … Koubarakis, M. (2020). Three-dimensional Entity Resolution with JedAI. *Information Systems*, *93*, 101565. doi: 10.1016/j.is.2020.101565

Papadakis, G., Skoutas, D., Thanos, E. & Palpanas, T. (2020). Blocking and Filtering Techniques for Entity Resolution. *ACM Computing Surveys*, *53* (2), 1–42. doi: 10.1145/3377455

Papadakis, G., Tsekouras, L., Thanos, E., Giannakopoulos, G., Palpanas, T. & Koubarakis, M. (2018). The return of jedAI. *Proceedings of the VLDB Endowment*, *11* (12), 1950–1953. doi: 10.14778/3229863.3236232

Papadakis, G., Tsekouras, L., Thanos, E., Giannakopoulos, G., Palpanas, T. & Koubarakis, M. (2020). Domain- and Structure-Agnostic End-to-End Entity Resolution with JedAI. *ACM SIGMOD Record*, *48* (4), 30–36. doi: 10.1145/3385658.3385664

Peled, O., Fire, M., Rokach, L. & Elovici, Y. (2016). Matching entities across online social networks. *Neurocomputing*, *210*, 91–106. doi: 10.1016/j.neucom.2016.03.089

Peng, T., Li, L. & Kennedy, J. (2012). A Comparison of Techniques for Name Matching. *GSTF INTERNATIONAL JOURNAL ON COMPUTING*, *2* (1).

Pershina, M. (2016). *Graph-Based Approaches to Resolve Entity Ambiguity* (Dissertation). New York University, New York, New York.

Prabhu, T. & Gnana Dhas, C. S. (2018). Improved scalability in mining using ontology record linkage algorithm. *Computers & Electrical Engineering*. doi: 10.1016/j.compeleceng.2018.01.026

Rahm, E. (2016). The Case for Holistic Data Integration. In J. Pokorný, M. Ivanović, B. Thalheim & P. Šaloun (Hrsg.), *Advances in Databases and Information Systems* (Bd. 9809, S. 11–27). Cham: Springer International Publishing. doi: 10.1007/978-3-319-44039-2{\textunderscore}2

Rahm, E. & Hai Do, H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Eng. Bull.*, *23*, 3–13.

Rahm, E., Nagel, W. E., Peukert, E., Jäkel, R., Gärtner, F., Stadler, P. F., ... Lehner, W. (2019). Big Data Competence Center ScaDS Dresden/Leipzig: Overview and selected research activities. *Datenbank-Spektrum*, *19* (1), 5–16. doi: 10.1007/s13222-018-00303-6

Randall, S. M., Ferrante, A. M., Boyd, J. H. & Semmens, J. B. (2013). The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making*, *13*, 1–10. doi: 10.1186/1472-6947-13-64

Rekatsinas, T., Dong, X. L. & Srivastava, D. (2014). Characterizing and selecting fresh data sources. In C. Dyreson, F. Li & M. T. Özsu (Hrsg.), *Proceedings of the 2014 ACM SIGMOD*

international conference on Management of data - SIGMOD '14 (S. 919–930). New York, New York, USA: ACM Press. doi: 10.1145/2588555.2610504

Ridder, H.-G. (2017). The theory contribution of case study research designs. *Business Research*, *10* (2), 281–305. doi: 10.1007/s40685-017-0045-z

Robra-Bissantz, S. & Strahringer, S. (2020). Wirtschaftsinformatik-Forschung für die Praxis. *HMD Praxis der Wirtschaftsinformatik*, *57* (2), 162–188. doi: 10.1365/s40702-020-00603-0

Roeder, J., Muntermann, J. & Kneib, T. (2020). Towards a Taxonomy of Data Heterogeneity. In N. Gronau, M. Heine, K. Poustcchi & H. Krasnova (Hrsg.), *WI2020 Zentrale Tracks* (S. 293–308). GITO Verlag. doi: 10.30844/wi{\textunderscore}2020{\textunderscore}c6 -roeder

Safhi, H. M., Frikh, B. & Ouhbi, B. (2019). Data Source Selection in Big Data Context. In M. Indrawan-Santiago, E. Pardede, I. L. Salvadori, M. Steinbauer, I. Khalil & G. Anderst-Kotsis (Hrsg.), *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services* (S. 611–616). New York, NY, USA: ACM. doi: 10.1145/3366030.3366121

Saluja, C. (2018). *Data Preparation - A crucial step in Data Mining.* Zugriff am 04.03.2019 auf `https://medium.com/@chhavi.saluja1401/data-preparation-a-crucial-step-in -data-mining-dba35772f281`

Sayers, A., Ben-Shlomo, Y., Blom, A. W. & Steele, F. (2016). Probabilistic record linkage. *International Journal of Epidemiology*, *45* (3), 954–964. doi: 10.1093/ije/dyv322

Schild, C.-J. & Schultz, S. (2017). Linking Deutsche Bundesbank Company Data using Machine-Learning-Based Classification.

Schildgen, J. & Deßloch, S. (2016). Heterogenität überwinden mit der Datentransformationssprache NotaQL. *Datenbank-Spektrum*, *16* (1), 5–15. doi: 10.1007/s13222-015-0207-0

Schmidt, A. (2010). *Entwicklung Einer Methode Zur Stammdatenintegration.* Berlin: Logos Verlag Berlin.

Schneider, A. T., Mukherjee, A. & Dragut, E. C. (2018). Leveraging Social Media Signals for Record Linkage. In P.-A. Champin, F. Gandon, M. Lalmas & P. G. Ipeirotis (Hrsg.), *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18* (S. 1195–1204). New York, New York, USA: ACM Press. doi: 10.1145/3178876.3186018

Shu, L., Lin, C., Meng, W., Han, Y., Yu, C. T. & Smalheiser, N. R. (2012). A framework for entity resolution with efficient blocking. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)* (S. 431–440). IEEE. doi: 10.1109/IRI.2012.6303041

Simonini, G., Bergamaschi, S. & Jagadish, H. (2016). BLAST: a Loosely Schema-aware Meta-blocking Approach for Entity Resolution. Zugriff auf http://www.vldb.org/pvldb/vol9/p1173-simonini.pdf

Simonini, G., Papadakis, G., Palpanas, T. & Bergamaschi, S. (2018). Schema-agnostic Progressive Entity Resolution. *IEEE Transactions on Knowledge and Data Engineering*, 1. doi: 10.1109/TKDE.2018.2852763

Stonebraker, M. & Ilyas, I. (2018). Data Integration: The Current Status and the Way Forward. *IEEE Data Eng. Bull.*, *41* (2), 3–9.

Szopinski, D., Schoormann, T. & Kundisch, D. (2019). Because your taxonomy is worth it: Towards a framework for taxonomy evaluation. In *Proceedings of the Twenty-Seventh European Conference on Information Systems (ECIS).*

Talburt, J. R. (2011). *Entity Resolution and Information Quality.* Elsevier. doi: 10.1016/C2009-0-63396-1

Theodoros I. Rekatsinas, Xin Dong, Lise Getoor & Divesh Srivastava. (2015). Finding Quality in Quantity: The Challenge of Discovering Valuable Sources for Integration. In *CIDR.*

Tsang, E. W. (2014). Case studies and generalization in information systems research: A critical realist perspective. *The Journal of Strategic Information Systems*, *23* (2), 174–186. doi: 10.1016/j.jsis.2013.09.002

UScompanylist Business. (o. J.). Zugriff auf https://www.uscompanieslist.com/

UScompanylist Company. (o. J.). Zugriff auf https://www.uscompanieslist.com/

USPTO. (o. J.). Zugriff auf https://developer.uspto.gov/

van Dam, I., van Ginkel, G., Kuipers, W., Nijenhuis, N., Vandic, D. & Frasincar, F. (2016). Duplicate detection in web shops using LSH to reduce the number of computations. In S. Ossowski (Hrsg.), *Proceedings of the 31st Annual ACM Symposium on Applied Computing - SAC '16* (S. 772–779). New York, New York, USA: ACM Press. doi: 10.1145/2851613.2851861

Volkswagen. (o. J.). *Volkswagen Konzern.* Zugriff auf https://www.volkswagenag.com/de.html

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R. & Cleven, A. (2015). Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the Association for Information Systems*, *37*, 205–224. doi: 10.17705/1CAIS.03709

Wandelt, S., Wang, J., Leser, U., Deng, D., Gerdjikov, S., Mishra, S., ... Wang, W. (2014). State-of-the-art in string similarity search and join. *ACM SIGMOD Record*, *43* (1), 64–76. doi: 10.1145/2627692.2627706

Wang, F. & Wang, H. (2016). Record Linkage Using the Combination of Twice Iterative SVM Training and Controllable Manual Review. In *2016 IEEE 14th Intl Conf 2016* (S. 31–38). doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.21

Wang, R. Y. & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12* (4), 5–33. doi: 10.1080/07421222.1996.11518099

Webster, J. & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26* (2), xiii–xxiii. Zugriff auf `http://www.jstor.org/stable/4132319`

Wikidata. (o. J.). Zugriff auf `https://www.wikidata.org/`

Wilde, T. & Hess, T. (2006). *Methodenspektrum der Wirtschaftsinformatik: Überblick und Portfoliobildung: Arbeitsbericht Nr. 2/2006*. München.

Wilde, T. & Hess, T. (2007). Forschungsmethoden der Wirtschaftsinformatik. *WIRTSCHAFTSINFORMATIK*, *49* (4), 280–287. doi: 10.1007/s11576-007-0064-z

Wirth, R. & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining.*

Witte, J.-H., Gerberding, J., Melching, C. & Marx Gómez, J. (2021). Evaluation of Deep Learning Instance Segmentation Models for Pig Precision Livestock Farming. *Business Information Systems*, 209–220. doi: 10.52825/bis.v1i.59

Yin, C.-Y. (2018). Measuring organizational impacts by integrating competitive intelligence into executive information system. *Journal of Intelligent Manufacturing*, *29* (3), 533–547. doi: 10.1007/s10845-015-1135-4

Yin, R. K. (2018). *Case study research and applications: Design and methods* (Sixth edition Aufl.). Los Angeles and London and New Dehli and Singapore and Washington DC and Melbourne: SAGE.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan & Wang-Chiew Tan. (2020). Deep Entity Matching with Pre-Trained Language Models. *ArXiv*, *abs/2004.00584*.

Zecchini, L., Simonini, G. & Bergamaschi, S. (2020). Entity Resolution on Camera Records Without Machine Learning. In *DI2KG@VLDB*.

Zrenner, J., Hassan, A. P., Otto, B. & Marx Gómez, J. (2017). Data source taxonomy for supply network structure visibility..

Eidesstattliche Erklärung

Ich versichere hiermit, dass ich meine Dissertation „End-to-End-Datenintegration von Realwelt-Entitäten“ selbständig und ohne fremde Hilfe angefertigt habe, und dass ich alle von anderen Autoren wörtlich übernommenen Stellen wie auch die sich an die Gedankengänge anderer Autoren eng anlegenden Ausführungen meiner Arbeit besonders gekennzeichnet und die Quellen zitiert habe.

Oldenburg, den 18. Oktober 2022 Felix Kruse