

MetaOpenData

Projektabschlussbericht

Förderkennzeichen mundialis GmbH & Co KG: 19F1031A

Förderkennzeichen Sopra Steria SE: 19F1031B

Bundesministerium für Verkehr und Digitale Infrastruktur
mFUND

Ansprechpartner:

Till Adams	adams@mundialis.de	Tel. +49 228 96289952
Sebastian Görke	sebastian.goerke@soprasteria.com	Tel. +49 151 40625945

mundialis GmbH & Co. KG

Sitz u. Registergericht: Bonn
Amtsgericht Bonn HRA 8528

Komplementärin:
mundialis Verwaltungsgesellschaft mbH

vertreten durch:
Dr. Markus Neteler,
Hinrich Paulsen, Till Adams

Sopra Steria SE

Sitz der Gesellschaft: Hamburg
Handelsregister: HRB151350
Amtsgericht Hamburg

Vertreten durch den Vorstand:
Urs M. Krämer

© mundialis 2019

Dieses Dokument ist als geistiges Eigentum des Urhebers (mundialis GmbH & Co. KG) geschützt.
Die Weitergabe von Informationen daraus ist ohne vorherige Rücksprache nicht erlaubt.

Informationen über Ihre gespeicherten Daten finden Sie auf unserer Homepage unter folgendem Link:
<https://www.mundialis.de/datenschutzerklaerung/>

Inhaltsverzeichnis

1	Einleitung.....	6
1.1	Ausgangssituation und Aufgabenstellung.....	6
1.2	Innovationsbezug und Gesamtziel des Vorhabens.....	6
1.3	Beschreibung des Vorhabens.....	7
1.4	Arbeitspakete.....	9
2	AP I: Datentypisierung.....	10
2.1	Durchführung einer Umfrage zum Thema Metadaten.....	10
2.1.1	Zielsetzung.....	10
2.1.2	Zielgruppe.....	10
2.1.3	Technische Realisierung.....	11
2.1.4	Aufbau der Umfrage.....	11
2.1.5	Ergebnisse der Umfrage.....	12
2.2	Dateninfrastrukturen.....	13
2.2.1	Übersicht in Frage kommender Dateninfrastrukturen.....	13
2.2.2	Betrachtung der mCLOUD.....	14
2.2.3	Betrachtung COPERNICUS.....	17
2.3	Metadatenmodelle.....	18
2.3.1	Betrachtung des DCAT-AP.de Metadatenprofils.....	19
2.3.2	Betrachtung des INSPIRE-Metadatenprofils.....	20
2.3.3	Vergleich der beiden Metadatenmodelle.....	22
2.4	Fazit der Datentypisierung.....	22
3	AP II: Erarbeitung und Evaluation der Machbarkeit der verschiedenen Ansätze zur Metadatenerneuerung.....	23
3.1	Zielformat für die Metadaten.....	23
3.2	Datenbetrachtung zur Metadatenerstellung.....	23
3.3	KI-basierter Ansatz zur Metadatenerstellung.....	24
3.4	Fazit zu den Ansätzen zur Metadatenerneuerung.....	26
4	AP III: Deep Learning.....	27
4.1	Einführung in das Konzept des Deep Learnings.....	27
4.2	Data Analytics.....	28
4.3	NLP.....	30
4.4	Machine Learning.....	32
4.5	Deep Learning.....	33
5	AP IV: Entwicklung eines Proof of Concept.....	37
5.1	Data-Interpreter.....	38
5.2	GNOS-Harvester.....	39
5.3	Deep-Learning-Interpreter.....	39
5.3.1	Daten sammeln.....	40

5.3.2 Daten aufbereiten.....	40
5.3.3 Daten Analysieren.....	42
5.3.4 Modell erstellen.....	46
5.4 ActiniaGDI-CSW-Writer.....	48
6 AP V: Evaluierung der Ergebnisse und Herausstellung des Verwertungspotentials.	49
6.1 Prototyp für die automatisierte Metadatenerzeugung.....	49
6.2 Potenziale zum Einsatz KI-basierter Verfahren bei der Metadatenerzeugung.....	50
6.2.1 Decision Tree.....	51
6.2.2 Naive Bayes.....	52
6.2.3 Random Forest.....	52
6.2.4 Deep Learning.....	52
6.2.5 Zusammenfassung der Projektergebnisse im Bereich KI-Verfahren.....	53
6.2.6 Ausblick KI-Verfahren in der Metadatenautomatisierung.....	54
6.2.7 Fazit.....	55
6.3 Ergebnispräsentationen und Folgeaktivitäten.....	55
6.3.1 Vorträge.....	55
6.4 Projektflyer.....	56

Abbildungsverzeichnis

Abbildung 1: Datenformate in der mCLOUD.....	14
Abbildung 2: Datentypisierung der in mCLOUD referenzierten Datenquellen.....	15
Abbildung 3: Daten-Metadaten-Kopplung in der mCLOUD.....	16
Abbildung 4: Verlinkte Metadaten nach Datenmodellen in der mCLOUD.....	16
Abbildung 5: Auszug der angezeigten Metadaten des CMEMS.....	18
Abbildung 6: Disziplinen der künstlichen Intelligenz.....	27
Abbildung 7: Bereiche des Data Analytics.....	28
Abbildung 8: Grundsätzliches Vorgehen bei Data Analytics.....	29
Abbildung 9: Klassische NLP Verfahren.....	32
Abbildung 10: Entscheidungsbaum.....	32
Abbildung 11: Schematische Darstellung von NLP mit Deep Learning.....	33
Abbildung 12: Ein Perzeptron.....	34
Abbildung 13: Multi Layered Perception (MLP).....	34
Abbildung 14: Arten von künstlichen neuronalen Netzen (KNN).....	36
Abbildung 15: Abläufe zur automatisierten Metadatenerzeugung.....	37
Abbildung 16: Architektur Prototyp.....	38
Abbildung 17: Erzeugung einer übergreifenden Datenstruktur.....	41
Abbildung 18: Daten der .gml-Dateien mit TopicCategoryCode.....	42
Abbildung 19: Aufkommen der Begriffe aus den .gml-Dateien samt TopicCategoryCode-Zuordnung	43
Abbildung 20: Keine Clusterung der Begriffe aus den .gml-Dateien erkennbar.....	44
Abbildung 21: Verteilung der TopicCategoryCode- Werte.....	45
Abbildung 22: RapidMiner Workflow zum Vergleich von ML und Deep Learning.....	46
Abbildung 23: RapidMiner Workflow – Validation-Komponente.....	48
Abbildung 24: Startseite des Prototypen MetaOpenData.....	49
Abbildung 25: Darstellung des berechneten Decision Tree.....	51
Abbildung 26: Testergebnisse Decision Tree.....	51
Abbildung 27: Testergebnisse Naive Bayes.....	52
Abbildung 28: Testergebnisse Random Forest.....	52
Abbildung 29: Testergebnisse Deep Learning.....	53
Abbildung 30: Testergebnisse aller KI-Verfahren.....	53

Tabellenverzeichnis

Tabelle 1: Änderungshistorie des Dokumentes.....	5
Tabelle 2: Dateninfrastrukturen, die im Projekt MetaOpenData untersucht wurden.....	14
Tabelle 3: Pflichtfelder DCAT-AP.de.....	20
Tabelle 4: Pflichtfelder INSPIRE-Metadatenmodell.....	21
Tabelle 5: Metadatenattribute und ihre Eignung für die Ableitung mittels KI (in fett) und mittels Datenbetrachtung.....	26
Tabelle 6: Natural Language Processing.....	30
Tabelle 7: Beispiele zum NLP.....	31
Tabelle 8: INSPIRE- Attribute.....	40
Tabelle 9: Dokumentenstruktur.....	40
Tabelle 10: Elemente der GML-Datei.....	41
Tabelle 11: Dateien mit TopicCategoryCode.....	43
Tabelle 12: Vergleich DCAT-AP.de und INSPIRE Metadatenmodelle.....	60

Änderungen:

Version	Datum	Autor	Beschreibung
0.1	25.10.2018	C. Eberz	Tabelle einfügen, neu Benennung des Dokumentes, Vermerk: Logo ändern
0.2	02.11.2018	E. Leonhardt	Kapitel zu AP III
0.3	05.11.2018	S. Goerke	Überarbeitung der Texte und Struktur
0.3	02.12.2019	C. Eberz	Überarbeitung der Texte und Struktur
0.4	07.11.2018	S. Goerke & E. Leonhardt	Überarbeitung und Kapitelstruktur
0.5	13.11.2018	S. Goerke, O. Bildesheim	QS Texte und Überarbeitungen
0.6	25.05.2019	T. Adams	Überarbeitung des Dokumentes
1.0	28.05.2019	T. Adams	Finalisierung Dokument v 1.0
1.1	25.06.2019	T. Adams	Finalisierung Dokument v 1.1
1.2	27.06.2019	C. Eberz	Finalisierung Dokument v 1.2

Tabelle 1: Änderungshistorie des Dokumentes

1 Einleitung

Das einleitende Kapitel stellt die Zielsetzungen und die Motivation hinter der Studie vor und beschreibt unsere Aufgabenstellung sowie unser Vorgehen.

1.1 Ausgangssituation und Aufgabenstellung

Der freie und unbeschränkte Zugang zu Daten ist das zentrale Element des Open-Data-Gedankens. Dieses Ziel ist aber alleine durch eine Bereitstellung von Daten nicht zu erreichen, denn die bereitgestellten Daten müssen für interessierte Nutzer auch auffindbar sein. Gerade im Kontext des Bundesministeriums für Verkehr und Digitale Infrastruktur (BMVI) ist dies eine zentrale Anforderung in Bezug auf die großen und wertvollen Datenbestände des Ressorts. Die beiden Unternehmen mundialis GmbH & Co KG und Sopra Steria Consulting haben ihre Expertise in den Bereichen Geo-IT, Open Data, Big Data und Deep Learning gebündelt und im Zuge des Vorhabens „MetaOpenData“, welches im Rahmen des Modernitätsfonds mFUND gefördert wurde, die Machbarkeit von Maßnahmen zur Verbesserung der Auffindbarkeit von Daten untersucht.

Im Fokus stand hierbei die Erarbeitung von Lösungsansätzen zur Automatisierung der Erfassung von Daten beschreibenden Metadaten.. Diese Metadaten sollen durch auf Deep-Learning basierenden Verfahren angereichert werden und letztlich zur Verbesserung der Datenauffindbarkeit – auch hinsichtlich unterschiedlicher Suchkontexte – führen. In derartigen Maßnahmen sehen die Projektbeteiligten insgesamt ein sehr großes Einsparpotential hinsichtlich der bisher für Metadatenerstellung und -pflege notwendigen personellen, finanziellen und zeitlichen Aufwände. Ebenfalls erwarten die Beteiligten von den erarbeiteten Lösungsansätzen, dass sie geeignet sind, die Hürden für die Datenbereitstellung auf der Anbieterseite zu senken und dadurch die öffentlich zur Verfügung stehende Datenbasis quantitativ wie auch qualitativ aufzuwerten. Insbesondere der Einsatz von Deep-Learning-Verfahren, welchen bislang in den Bereichen Geo-IT und Open Data noch zu geringe Bedeutung zukommt, verfügen über das Potential, auf disruptive Art und Weise dazu beizutragen, bisherige Verfahren der Metadatenbereitstellung zu verbessern und neue Möglichkeiten der Datenrecherche zu ermöglichen.

1.2 Innovationsbezug und Gesamtziel des Vorhabens

Das Gesamtziel des Vorhabens im Rahmen einer Machbarkeitsstudie ist die Verbesserung der Auffindbarkeit von Daten aus dem Geschäftsbereich des BMVI. Dieses lässt sich in die folgenden Teilziele untergliedern:

- Vereinfachung und Optimierung der Metadatenerfassung, -pflege und –bereitstellung und damit Vereinfachung und Verbesserung des Zugriffs auf freie Daten und freie Geodaten durch automatisierte Anreicherung dieser Daten mit Metadaten.
- Verbesserung der Auffindbarkeit von Daten.
- Optimierung der Metadatenzuordnung durch Verfahren des Deep Learnings.
- Evaluation der Machbarkeit möglicher Ansätze zur Metadatenautomatisierung mittels einer Machbarkeitsstudie („Proof of Concept“).

- Nachnutzung der Ergebnisse durch Entwicklung einer standardisierten Open Source Lösung zur Metadatenautomatisierung in einer zweiten Projektphase.

1.3 Beschreibung des Vorhabens

Der freie und unbeschränkte Zugang zu Daten ist das zentrale Element des Open Data Gedankens. Um dieses zu erreichen, ist eine möglichst einfache Auffindbarkeit der offengelegten Datensätze erforderlich. „Ein Leitgedanke des mFUNDs besteht daher auch darin, einen breiten Zugang zu den Daten des BMVI und seines Geschäftsbereichs zu gewähren und damit Innovationen und umsetzungsnahe Anwendungsfälle für die Datennutzung zu ermöglichen“, so wird eines der Kernziele von mFUND in der Präambel der Förderrichtlinie formuliert.

Ein Hemmnis für die angestrebte volle Ausschöpfung des Potenciales der Daten besteht in deren Auffindbarkeit und damit einhergehend auch in der Abhängigkeit von der Qualität der beschreibenden Metadaten. Zu diesem Zweck haben sich insbesondere im Bereich der Geo-IT Suchdienste auf Basis standardisierter Metadaten etabliert. Die Erstellung und vor allem die Pflege der hierzu benötigten Metadaten ist jedoch erfahrungsgemäß sehr kosten-, arbeits- und zeitintensiv wie u.a. auch das Beispiel des Aufbaus der europäischen Geodateninfrastruktur INSPIRE zeigt.

Von immenser Wichtigkeit ist, dass die Metadaten eine hohe Qualität und damit verbunden auch eine hohe Aktualität besitzen. Nur so kann gewährleistet werden, dass die Fachdaten im jeweiligen Suchkontext auch gefunden werden. Dieser unbedingt notwendige, jedoch auch hohe Aufwand führt häufig dazu, dass Kapazitäten für die Pflege und den Aufbau der Metadatenbestände der datenhaltenden Stellen gebunden werden, oder auch, dass Daten überhaupt nicht bzw. nur mit qualitativ unzureichenden Metadaten veröffentlicht werden. Damit einher geht oft eine mangelhafte Aktualität der Metadaten. Diese Faktoren schränken sowohl die themenspezifische Verfügbarkeit als auch die Auffindbarkeit ggf. interessanter Datenbestände und damit deren Nachnutzung (in welcher das eigentliche Potenzial offener Daten besteht) ein oder verhindert sie sogar. Gerade im Geschäftsbereich des „Datenministeriums“ BMVI mit den umfangreichsten Fachdatenbeständen auf Bundesebene sind dies wichtige Argumente dafür, weniger aufwendige, einheitliche Lösungen, Standards und Regeln zur Metadatenpflege zu erarbeiten, die auch zur besseren bzw. intensiveren Datennutzung führen können.

Ähnliche Herausforderungen zeigen sich auch bei den Sentinel Satellitendaten aus dem Europäischen Erdbeobachtungsprogramm Copernicus. Diese Satellitendaten werden beispielsweise durch das Deutsche Zentrum für Luft- und Raumfahrt (DLR) deutschlandweit über die „Copernicus Data and Exploitation Platform - Deutschland“ (CODE-DE) bereitgestellt. Die Datenbestände sind in zwei Jahren auf weit über eine Million Szenen angewachsen und ihre Anzahl wächst täglich. Hier ist es notwendig, dass die Bilddaten mit ausreichend detaillierten Metadaten beschrieben werden, damit relevante Szenen im Suchkontext bestimmter Fragestellungen aufgefunden werden können und so die Nutzung dieser offenen Daten auch für eine breite Masse von Anwendungen ermöglicht wird.

Vor diesem Hintergrund hatte das Projekt „MetaOpenData“ zum Ziel, Wege aufzuzeigen, wie ein dynamisches und automatisiertes Ableiten von Metadateninformationen aus vorhandenen Daten ermöglicht werden kann.

Auf Basis einer Machbarkeitsstudie („Proof of Concept“) sollte eine prototypische Lösung erarbeitet werden, die in einer möglichen weiteren Projektphase zu einem produktiv einsetzbaren System weiterentwickelt werden soll. Eine solche Lösung würde Ressourcen freimachen, die ansonsten in Erfassung und Pflege von Metadaten gebunden sind. Darüber hinaus könnten somit Fachdaten erstmalig mit quantitativ und qualitativ ausreichenden und vereinheitlichten Metadaten beschrieben und als offene Daten (engl. Open Data) publiziert werden.

In der Studie sollte untersucht werden, wie eine automatisierte Anreicherung von Metadaten aus den vorliegenden Geodaten selbst sowie weiterer Metadaten, die sich aufgrund von räumlichen oder attributiven Zusammenhängen aus anderen Quellen extrahieren lassen, umgesetzt werden kann. Als Beispiel mag die Anreicherung der Metadaten von Sentinel Satellitenbildern um die entsprechende durchschnittliche Landhöhe oder die durchschnittlichen Niederschlagswerte innerhalb eines „Footprints“ (d.h. die vom Satelliten abgedeckte geografische Region) dienen. Ein weiteres Beispiel stellt die Beschreibung von sehr dynamischen Fachdaten aus dem Mobilitätsumfeld dar, die eines sehr dichten Aktualisierungszyklus bedürfen. Dabei sollten auch Aspekte berücksichtigt werden, die bislang nicht durch Metadatensuchen abgedeckt werden: Der Aktualitätsbezug von Fachdaten, die für einen bestimmten Zeitpunkt Fragestellungen beantworten helfen wie bspw. „Welche Objekte befinden sich wann an welchem Ort?“. Hier spielt die interne Verknüpfung der Fachdaten aus dem Geschäftsbereich des BMVI eine große Rolle, beispielsweise die von Wetter- mit Verkehrsdaten.

Darüber hinaus besteht eine weitere Problematik darin, dass Datenbereitsteller mit ihren Metadaten auch eine gewisse Nutzungsrichtung für die Daten vorgeben. Eine Arbeitshypothese der Studie war daher, dass diese Vorgehensweise möglicherweise die Nutzung von Daten in anderem Kontext verhindert, da sie bei Recherche in einem komplett anderen Kontext schlicht nicht auffindbar bzw. nur schwer als ggf. relevant erkennbar sind. Im Rahmen unseres Vorhabens sollten insbesondere Methoden und Verfahren des Deep Learnings eingesetzt werden. Darunter versteht man die Fähigkeit von Systemen, mittels Lernalgorithmen sinnhafte Entscheidungen zu treffen bzw. vorzuschlagen. In Bezug auf die Fragestellung, die Auffindbarkeit von Daten durch automatisierte Ableitung von Metadaten zu optimieren, sollte mit solcherlei Verfahren das System befähigt werden, selbstständig und objektiv Metadaten zuzuordnen. Als Eingangsdaten zur Systemkalibrierung könnten dazu bereits mit Metadaten versehene Datensätze verwendet werden. Im Rahmen der BMVI-Initiative rund um den mFUND (datenbasierte Förderprogramm des BMVI) und die mCLOUD (zentraler Zugangspunkt zu offenen Daten rund um Themen des BMVI Ressorts) wird ein Fokus auf die bessere Verfügbarkeit und Mehrwertgenerierung hinsichtlich der Daten aus dem Geschäftsbereich gesetzt. Die Studie sollte sowohl dazu beitragen die Datenverfügbarkeit als auch deren Auffindbarkeit zu verbessern als auch diese Ansätze auf ihre Umsetzbarkeit hin untersuchen. Dazu sollten verschiedene Lösungsansätze zur dynamischen, qualitativ hochwertigen sowie weitestgehend automatisierten Ableitung von Metadaten aus den eigentlichen Daten bzw. durch Verknüpfung mit weiteren Datensätzen erarbeitet werden. Diese Machbarkeitsstudie soll damit

Grundstein für weitere Projekte sein, die die Umsetzung der erarbeiteten Ansätze zum Inhalt haben und somit einen Beitrag dazu leisten können, dass die Daten im BMVI-Geschäftsbereich besser auffindbar werden und somit eine noch größere Wertschöpfung erreicht und Innovation gefördert werden kann. Insbesondere sehen wir hier die Weiterentwicklung unserer Ansätze hin zu einer Open-Source-Lösung zur Metadatenautomatisierung als wichtigen Treiber für diese weitergehende Wertschöpfung an.

1.4 Arbeitspakete

Insgesamt wurden 5 Arbeitspakete (APs) im Projekt MetaOpenData identifiziert. Diese wurden in agiler Herangehensweise in unterschiedlichen Konstellationen bearbeitet. Die 5 Arbeitspakete sowie deren Ergebnisse werden in den folgenden Kapitel vorgestellt. Im einzelnen wurden die folgenden Arbeitspakete bearbeitet:

- AP I: Datentypisierung (Kap. 2)
- AP II: Erarbeitung und Evaluation der Machbarkeit der verschiedenen Ansätze zur Metadatenautomatisierung (Kap. 3)
- AP III: Deep Learning (Kap. 4)
- AP IV: Entwicklung eines Proof of Concept (Kap. 5)
- AP V: Evaluierung der Ergebnisse und Herausstellung des Verwertungspotentials (Kap. 6)

2 AP I: Datentypisierung

Am Anfang dieser Machbarkeitsstudie steht die Frage, wie die Daten aussehen, deren Auffindbarkeit verbessert werden soll. Ohne Kenntnis darüber, welche Daten verfügbar sind, wie diese aussehen und strukturiert sind und wie die zugehörigen Metadaten aussehen, ist es nicht möglich, sinnvolle Ansätze zur Metadatenerarbeitung zu erarbeiten. Einen Kenntnisstand zu den vorhandenen Daten zu erarbeiten ist daher notwendig. . Somit ist es auch erforderlich, einen Überblick der zur Verfügung stehenden Dateninfrastrukturen zu erhalten und darüber hinaus die darin enthaltenen Daten zu typisieren. Um dies zu erreichen, wurde im Rahmen von MetaOpenData eine Online-Befragung zur Metadatennutzung durchgeführt. Die Ergebnisse dieser Umfrage lieferen wichtige Hinweise und Grundlagen für den weiteren Projektverlauf.

In Anhang I befindet sich ein ausgefüllter anonymer Beispelfragebogen mit den Fragen, die die Teilnehmer beantwortet haben.

2.1 Durchführung einer Umfrage zum Thema Metadaten

2.1.1 Zielsetzung

Ziel der Umfrage war es, Nutzungspräferenzen und Anforderungen an Metadaten abzufragen. Die Ergebnisse des Projektes MetaOpenData sollen dazu dienen, vollständigere Metadaten zu generieren, um ein besseres Auffinden und Beurteilen der Eignung von Daten zu ermöglichen - und zwar aus Nutzersicht. Um potentielle Nutzer eines Systems zur automatisierten Metadaten-Zuordnung frühzeitig in das Projekt einzubeziehen und möglichst umfassende Kenntnisse über deren Erfahrungen mit und Anforderungen an Metadaten zu erfahren (Nutzen Sie Metadaten? Welche Zusatzinformationen zu Daten interessieren Sie? Was fehlt?), wurde daher eine umfassende Online Befragung entwickelt. Die Umfrage ist ein wichtiger Beitrag zu AP-I, im Wesentlichen Erarbeitung der Datentypisierung. Dies sollte insbesondere dazu beitragen, eine Bestandsaufnahme hinsichtlich der Quantität und der Qualität der Metadaten aus Nutzersicht zu erstellen, um die Kriterien für die Datentypisierung zu definieren und diejenigen Datentypen auszuwählen, die am besten für eine Betrachtung im Rahmen der Machbarkeitsstudie relevant und geeignet erschienen.

2.1.2 Zielgruppe

Zielgruppe der Umfrage waren in erster Linie Nutzer von (Meta-)Daten jeder Art, mit einem Fokus auf Daten im Geschäftsbereich des BMVI. Darunter fallen unter anderem Daten und Dienste des Europäischen Erdbeobachtungsprogramms Copernicus (<http://www.copernicus.eu/>), insbesondere des darin enthaltenden Dienstes zur Überwachung der Meeresumwelt (<http://marine.copernicus.eu/>), für den das Bundesamt für Seeschifffahrt und Hydrographie (BSH) als Geschäftsbereichsbehörde des BMVI die Fachkoordination in Deutschland übernimmt. Auch Nutzer weiterer Dienste und Portale, wie bspw. mCLOUD (<https://www.mcloud.de/>), mit welchem das BMVI einen zentralen Zugangspunkt zu offenen Daten rund um Themen seines Ressorts bereitstellt, das Geoportal der Bundesanstalt für Gewässerkunde (BfG) (<https://geoportal.bafg.de/portal/Start.do>), das GeoSeaPortal des BSH (<https://www.geoseaportal.de/mapapps/?lang=de>, die Marine Daten Infrastruktur Deutschland (MDI-DE, <http://projekt.mdi-de.org/services/nokis.html>) oder auch Daten der Deutschen Bahn (<https://data.deutschebahn.com/>) standen grundsätzlich im Fokus.

Insbesondere sorgte die aktive Streuung der Umfrage in unterschiedlichen Branchen und Unternehmen für einen weitreichenderen Überblick für das Projekt MetaOpenData.

2.1.3 Technische Realisierung

Die Umfrage wurde mit Hilfe des Online Tools EUSurvey (<https://ec.europa.eu/eusurvey/>) der Europäischen Kommission erstellt. EUSurvey ist ein Instrument zur Verwaltung von Online-Umfragen. Damit lassen sich für die meisten Web-Browser geeignete Fragebögen und interaktive Formulare erstellen, veröffentlichen und verwalten. Die Umfrage zu MetaOpenData wurde für eine Bearbeitungszeit von etwa 10 bis 12 Minuten konzipiert und war über den Link <https://ec.europa.eu/eusurvey/runner/MetaOpenData> abrufbar. Die Umfrage war anonym (d.h. es können keine Rückschlüsse auf die Teilnehmenden gezogen werden) und war über einen nicht-personalisierten Zugang erreichbar. Die in deutscher Sprache verfasste Umfrage war sechs Wochen lang zugänglich.

2.1.4 Aufbau der Umfrage

Die Umfrage war in insgesamt fünf Teile gegliedert. Insgesamt bestand die Umfrage aus 18 Fragen: sechs Fragen mit Einfachantwortmöglichkeit, drei mit Mehrfachantwortmöglichkeit sowie acht freien Textantworten und einer numerischen Eingabe.

Der einleitende Teil informierte die Teilnehmenden über die Zielsetzung und den Zweck der Umfrage. Im zweiten Teil („Allgemeine Fragen“) konnten Angaben über die Organisation des Umfrageteilnehmers gemacht werden, um bei der Auswertung einen Überblick über die Nutzerkategorie und die jeweiligen Arbeitsfelder der Teilnehmenden zu erhalten.

Im dritten Themenblock („Fragen zur Nutzung von Daten“) wurden Fragen über die Nutzung von Daten im Allgemeinen ergänzt. Hierzu wurden insbesondere die genutzten Datenformate wie beispielsweise *.XLS oder *.CSV abgefragt. Die hier frei einzufügenden Formate konnten aus den verschiedensten Datenbereichen stammen (Geodaten, Wetterdaten u. v. m.). Auch wurden die unterschiedlichen Themenfelder erfragt, in welchen Daten verwendet werden, sowie die Häufigkeit der Nutzung dieser Daten in Projekten.

Im vierten Themenbereich („Fragen zur Nutzung und Anforderungen an Metadaten“) wurde konkret das Thema Metadaten behandelt. Hier sollte in Erfahrung gebracht werden, ob und wie Metadaten von den verschiedenen Teilnehmenden genutzt werden (bspw. zum Auffinden oder zur Beurteilung der Daten) und ob die Teilnehmenden auch selbst Metadaten erstellen (bspw. als datenführende Stelle). Zusätzlich spielten auch die Kriterien zur Beurteilung der Eignung von Daten eine wichtige Rolle, um weitere Erkenntnisse hinsichtlich des Designs von MetaOpenData zu gewinnen. Eine zentrale Frage in diesem Kontext war, welche Informationen (Metadaten) über Daten bislang aus Nutzersicht fehlen. Dazu konnten die Teilnehmenden sowohl über vordefinierte Antworten (Mehrfachauswahl) als auch durch freie Texteingabe Angaben machen.

Im letzten Umfrageabschnitt („Kommentare“) hatten die Teilnehmenden die Möglichkeit, durch freie Texteingabe weitere Hinweise oder Anregungen zu geben.

2.1.5 Ergebnisse der Umfrage

Der Rücklauf zur Umfrage (Sachstand 30.06.2018) betrug insgesamt 38 Teilnehmende. Im Folgenden werden einige wesentliche Erkenntnisse aus der Umfrage zusammengefasst. Die Prozentangaben (gerundet) beziehen sich stets auf den Anteil an der Gesamtzahl der Antwortenden (38). Die vollständige Auswertung der Umfrage befinden sich in Anhang II.

Von den insgesamt 38 Teilnehmenden war ein großer Teil (42%) bei Behörden, wissenschaftlichen Einrichtungen (37%) und in der freien Wirtschaft (16%) beschäftigt. 18% der Behördenvertreter arbeiten im Geschäftsbereich des BMVI.

Geodaten und andere Daten werden laut Umfrage hauptsächlich in den Bereichen Verkehr (39%), Wasser (47%) und Wetter/Klima (45%) genutzt. In der Kategorie „Weitere“ (47%) wurden Straßenverkehr (32%) und Bahnverkehr (21%) genannt.

Ein großer Teil der Teilnehmenden schätzt das eigene Vorwissen als zumindest umfangreich ein (umfangreich: 32%, sehr umfangreich: 32%). Nur 3% geben sehr geringes Wissen in diesem Themenbereich an.

63% der Teilnehmenden erstellt und / oder pflegt selbst Metadaten, wobei 32% diese Aufgaben sowohl für Dritte (bspw. als datenführende Stelle/Behörde) oder intern für Ihre eigene Organisation wahrnehmen. 18% erstellen Metadaten ausschließlich für die interne Nutzung, 13% ausschließlich für Dritte.

Bezüglich der Informationskategorien in den Metadaten waren Datenformat und Zeitraum der Aufzeichnung (jeweils 76%) sowie räumliche Ausdehnung (68%) und Nutzungsrechte (66%) die wichtigsten. Darüber hinaus wurden die thematische Kategorie und Ansprechpartner häufig genannt (55% bzw. 47%). Ein großer Teil der Teilnehmenden (39%) gab an, selten ausreichend Informationen, die interessant oder zur Beurteilung der Eignung der Daten für die eigene Anwendung benötigt würden, in den Metadaten zu finden. Hingegen geben 32% an, dass dies meist der Fall ist und 16% fanden stets alle erforderlichen Informationen.

Zu den am häufigsten verwendeten Datenformaten der Teilnehmenden gehört das XLS-Format (insgesamt 25 Nennungen). Somit wird am häufigsten das Office-Tool „Excel“ im Zusammenhang mit Datennutzung bei den Teilnehmern verwendet. Weitere stark genutzte Datenformate umfassen XML (22 Nennungen) sowie KML (15 Nennungen), die beide den Standards der OGC entsprechen und zur Beschreibung von Geodaten dienen. Weitere vereinzelt genannte Datenformate umfassen gängige Office-Formate aber auch übliche Datenformate zur Verarbeitung von Geodaten wie Geo-Tiff, GeoJson oder Shapefiles.

Die am häufigsten genannten Metadatenstandards waren das Data Catalog Vocabulary (DCAT) und die Infrastructure for Spatial Information in the European Community (INSPIRE). Auch zahlreiche weitere, z.T. sehr fachspezifische Standards wurden aufgeführt wie Exif, Rapidminer Annotations oder NetCDF CF1. Insgesamt wurden jedoch deutlich weniger genutzte Metadatenstandards (insgesamt 29) genannt als Datenformate.

Zu den am häufigsten angegebenen Portalen für die Datensuche wurden neben zahlreichen Einzelnennungen hauptsächlich mCloud und CODE-DE aufgezählt. Auch GovData und die Marine Dateninfrastruktur Deutschland (MDI-DE) werden häufiger zur Datensuche genutzt.

2.2 Dateninfrastrukturen

Im Rahmen der Machbarkeitsstudie sollte untersucht werden, wie Erfassung und Pflege von Metadaten automatisiert werden können. Ein besonderer Fokus lag dabei auf der Erprobung von Deep-Learning-Ansätzen und auf Geodaten, die über einen eindeutigen Raumbezug verfügen. Vor diesem Hintergrund war es notwendig, geeignete Dateninfrastrukturen zu identifizieren, die eine Erfassung der Vielfalt der Datentypen im Bereich von Geodaten und offenen Daten ermöglichen. Die folgende Tabelle zeigt eine Übersicht der grundsätzlich in Frage gekommener Infrastrukturen.

2.2.1 Übersicht in Frage kommender Dateninfrastrukturen

Tabelle 2 umfasst nicht alle vorhandenen Dateninfrastrukturen. Sie stellt aber eine Übersicht zu den für dieses Projekt relevanten Dateninfrastrukturen dar.

Dateninfrastruktur	Domäne	Beschreibung
mCLOUD	Daten aus dem BMVI Ressort	Die mCLOUD ist das Datenportal des BMVI und hält dazu entsprechende Metadaten zu offenen Daten aus den nachgeordneten Bereichen. Dazu gehören Geodaten, Wetterdaten, Verkehrsdaten, etc.
GovData	Open (Government) Data	Das GovData Portal ist das Open-Data-Portal des Bundes und der teilnehmenden Länder, welches die Recherche nach offenen Verwaltungsdaten aus allen Bereichen ermöglicht.
Geoportal.de	Geodaten	Das Geoportal.de ist das Suchportal für Geodaten innerhalb der Geodateninfrastruktur Deutschland.
EU INSPIRE Geoportal	Geodaten	Das EU INSPIRE Geoportal ermöglicht die europaweite Recherche nach INSPIRE-konformen Geodaten und schließt damit auch die GDI-DE ein.
European Data Portal	Open (Government) Data / PSI	Das European Data Portal ermöglicht die EU-weite Recherche nach offenen Daten und PSI (Public Sector Information) und stellt damit quasi das europäische Pendant zu GovData dar.
Copernicus	Fernerkundungsdaten und daraus abgeleitete Geodaten, Open Data	Copernicus ist das Erdbeobachtungsprogramm der Europäischen Kommission. Es liefert Erdbeobachtungsdaten für den Um-

		weltschutz, zur Klimaüberwachung, zur Einschätzung von Naturkatastrophen und für andere gesellschaftliche Aufgaben.
GeoSeaPortal des BSH	Offene und geschlossene Geobasis- und Geofachdaten des BSH	Zentraler Einstiegspunkt in die Geodateninfrastruktur des BSH (GDI-BSH) und zugleich zentrale Zuganger zu allen Geodaten des BSH für interne und externe Nutzerinnen und Nutzer.
MDI-DE	Geodaten	Marine Dateninfrastruktur Deutschland ist das Suchportal für Geodaten und Informationen aus dem Küstingenieurwesen, dem Küstengewässerschutz, dem Meeresumweltschutz und dem Meeresnaturschutz.
Geoportal der BfG	Geodaten	Das Geoportal der BfG erschließt neben BfG-Daten auch Datenbestände der Partner der BfG aus dem Verkehrs- und Umweltbereich.

Tabelle 2: Dateninfrastrukturen, die im Projekt MetaOpenData untersucht wurden

2.2.2 Betrachtung der mCLOUD

Für die Machbarkeitsstudie lag es nahe, zunächst das Datenportal des BMVI, also die mCLOUD näher zu betrachten. Die Untersuchung der mCLOUD daraufhin, ob diese eine ausreichende Datengrundlage zur Durchführung unserer Machbarkeitsstudie darstellte, erbrachte die im folgenden dargestellten Ergebnisse (Abbildung 1). Die mCLOUD beinhaltet Metadaten zu Daten in den in Abbildung 1 aufgezeigten Datenformaten.



Abbildung 1: Datenformate in der mCLOUD

Zum Zeitpunkt der Betrachtung (Dezember 2017) beinhaltete das Datenportal knapp 600 Metadatensätze zu verschiedenen Datenquellen innerhalb des Geschäftsbereiches des BMVI und darüber hinaus verschiedener kommunaler Stellen mit Verkehrsbezug.

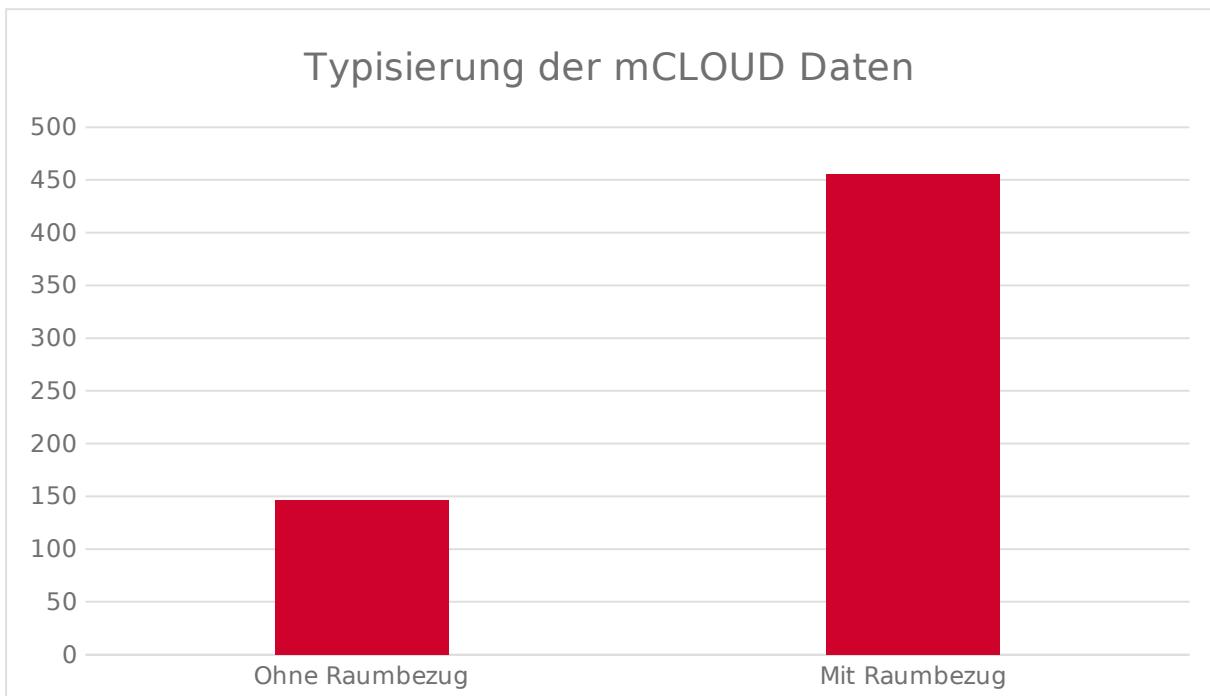


Abbildung 2: Datentypisierung der in mCLOUD referenzierten Datenquellen

Zur weiteren Typisierung wurden die Metadatensätze hinsichtlich der beschriebenen Daten kategorisiert. Hierzu wurde das Vorhandensein eines Raumbezugs der Daten als Kategorie gewählt.

Abbildung 2 zeigt, dass die mCLOUD zum überwiegenden Teil Metadaten zu Daten mit Raumbezug, also Geodaten beinhaltet. Für die Machbarkeitsstudie bedeutet dies, dass der Fokus bei der Betrachtung der zu behandelnden Daten auf den Bereich der Geodaten zu legen ist, da hier mit Abstand die meisten Daten vorhanden sind.

Ein wichtiges Merkmal bei der Betrachtung geeigneter Daten ist die Verbindung zwischen Daten und Metadaten. Ohne eine Verlinkung mit den eigentlichen Daten sind Metadaten grundsätzlich weniger für die Datenrecherche und -nutzung geeignet. Gleiches gilt für die Durchführung der Machbarkeitsstudie. Insbesondere für die Untersuchung KI-basierter Ansätze ist es erforderlich, dass eine technische Verbindung zwischen Daten und Metadaten besteht.

Hier ergibt sich aus der Betrachtung der mCLOUD ein ungenügendes Bild. Abbildung 3 zeigt, dass von den Metadaten der mCLOUD nicht einmal die Hälfte eine Referenz zum Datensatz enthält bzw. umgekehrt.

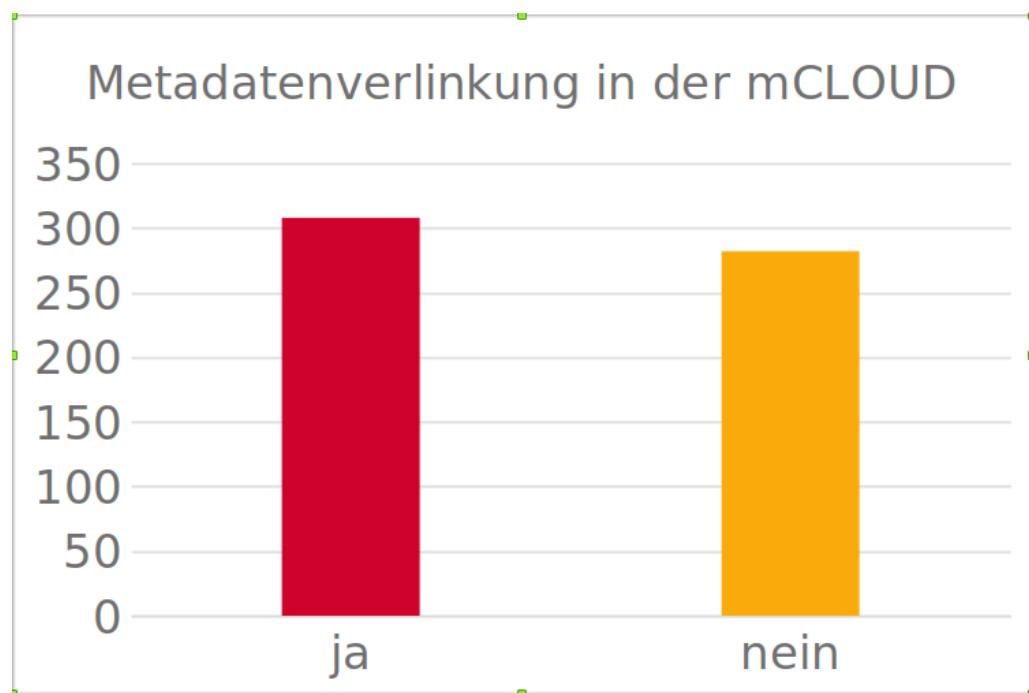


Abbildung 3: Daten-Metadaten-Kopplung in der mCLOUD

Eine nähere Betrachtung der Metadaten in der mCLOUD, welche eine Verlinkung mit den jeweiligen Datensatz besitzen zeigt, dass dies vor allem Geodaten sind. Ebenfalls fällt auf, dass insbesondere Metadatensätze, die dem INSPIRE-Standard entsprechen, solche Verlinkungen besitzen. Abbildung 4 gibt eine Übersicht über die Verteilung der Verlinkung auf die verschiedenen in der mCLOUD vorhandenen Datenmodellen.

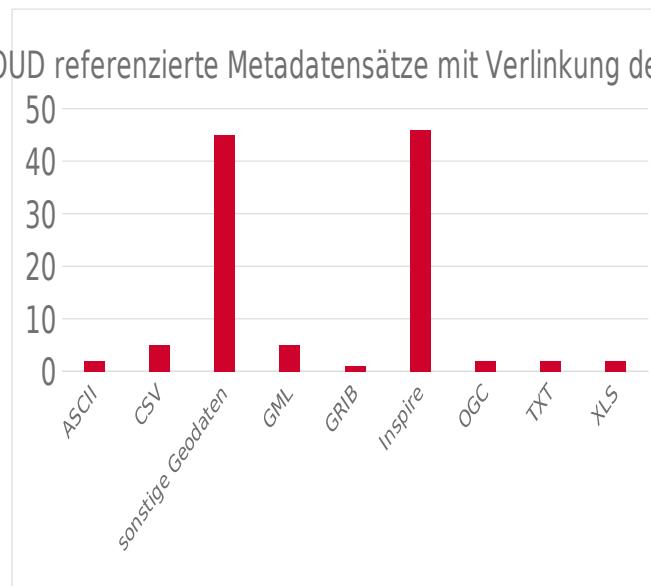


Abbildung 4: Verlinkte Metadaten nach Datenmodellen in der mCLOUD

Aus der Betrachtung der mCLOUD lassen sich die folgenden Erkenntnisse hinsichtlich der Verwendbarkeit der mCLOUD-Datenbasis für die Machbarkeitsstudie ableiten.

Zunächst ist festzuhalten, dass die Datenbasis von knapp 600 Metadatensätzen nicht ausreichend ist, um KI-basierte Ansätze näher zu betrachten. Hier arbeitet man in der Regel mit mindestens 1000 Datensätzen. Die Basis von brauchbaren Datensätzen reduziert sich durch die Betrachtung der unbedingt erforderlichen Daten-Metadatenkopplung, ohne die eine nähere Betrachtung der Daten-Metadatenpaare nicht mit angemessenem Aufwand möglich ist.

Das bedeutet, die mCLOUD ist für die Betrachtung der KI-gestützten Metadatenautomatisierung im Rahmen der Machbarkeitsstudie nicht brauchbar.

Eine weitere Erkenntnis aus der Betrachtung der mCLOUD ist jedoch, dass insbesondere Datensätze, die INSPIRE-Datenmodelle implementieren, über die Voraussetzungen verfügen, für einen KI-gestützten Ansatz zur Metadatenautomatisierung als Trainingsdaten zu dienen.

2.2.3 Betrachtung COPERNICUS

Copernicus Daten wurden im Rahmen der Studie deshalb betrachtet, weil diese – zumindest teilweise – auch in den Zuständigkeitsbereich des BMVI fallen.

Copernicus ist ein umfassendes Programm mit einer eigenen Satellitenflotte und beinhaltet In-situ-Daten und Daten nationaler und kommerzieller Satelliten. Copernicus ist ein Programm der Europäischen Union. Die Europäische Weltraumorganisation (ESA) ist mit der technischen Koordination der Weltraumkomponente beauftragt. Die ESA und die Europäische Organisation für die Nutzung Meteorologischer Satelliten (EUMETSAT) sind außerdem mit Betriebsaufgaben der Weltraumkomponente betraut. Die sechs thematischen Copernicus Kerndienste stellen umfangreiche Grundlageninformationen bereit, die für vielfältige Anwendungen weiter verarbeitet werden können. Die Informationsprodukte dieser Dienste stehen den Nutzern operationell zur Verfügung. Grundsätzlich werden die Daten und Informationsprodukte der Kerndienste kostenlos allen Nutzern zur Verfügung gestellt. Copernicus bietet gegenwärtig (Stand November 2018) 1.092 verschiedene Informationsprodukte in den sechs thematischen Diensten an (<https://services-portfolios.copernicus.eu/>). Dabei gilt zu beachten, dass ein Produkt weiter gefasst ist und verschiedenartige Daten, die sich teilweise auch thematisch überlappen können enthalten kann. So können bspw. Notfallkartierungen des Copernicus Emergency Management Service (CEMS) mehrere Datentypen enthalten. Als Formate handelt es sich in der Regel um Geotiffs und Shapefiles (<https://emergency.copernicus.eu/mapping>).

Neben den 6 thematischen Copernicus Services, die bereits aufgearbeitete und veredelte Daten, die auch aus anderen Quellen stammen können, zeichnen die zum Copernicus Erdbeobachtungsprogramm gehörenden Satelliten (die Sentinels) seit 2014 beinahe täglich Satellitenbilder der Erdoberfläche auf (<https://sentinels.copernicus.eu/web/sentinel/sentinel-data-access>).

Der Copernicus Service zur Überwachung der Meeressumwelt (CMEMS), für dessen Fachkoordination in Deutschland das BSH im Geschäftsbereich des BMVI zuständig ist, beinhaltet Metadaten zu insg. 164 Daten-Produkten. Diese Daten wurden entweder in situ, erfasst, modelliert oder aus Satellitendaten berechnet (Quelle: <http://marine.copernicus.eu/services-portfolio/access-to-products/>).

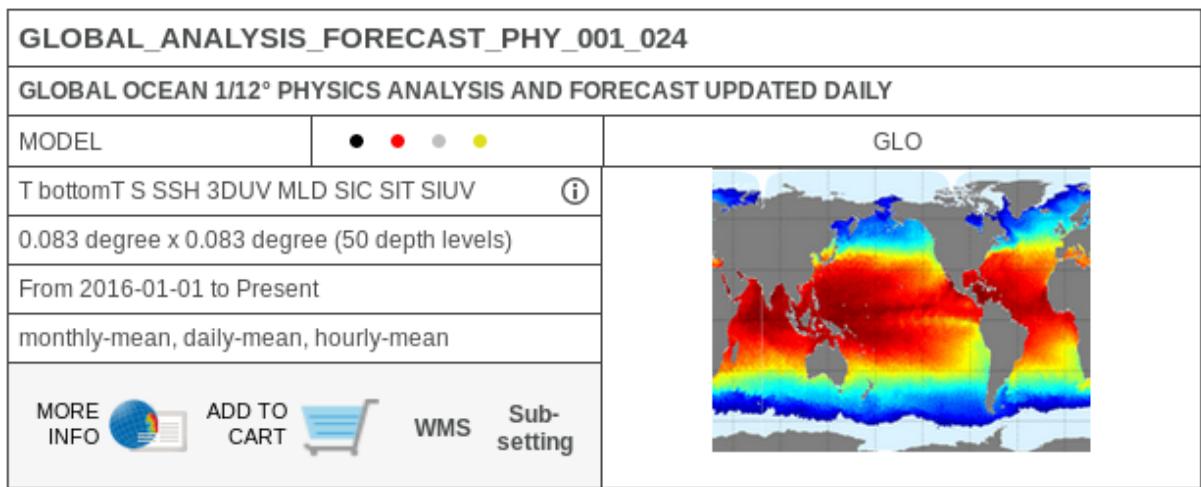


Abbildung 5: Auszug der angezeigten Metadaten des CMEMS

In Abbildung 5 sieht man den Button zu „More Info“, der zu einer Seite mit weiteren Metadaten führt. Diese stammen offensichtlich aus einem Metadatenkatalog. Auch hier zeigt sich der Nutzen der Metadaten aber auch der enorme Aufwand, der hinter Pflege und Bereitstellung steht.

Eine nähere Betrachtung der Metadaten im CMEMS, welche eine Verlinkung mit dem Datensatz besitzen zeigen, dass es sich dabei ausschließlich um Geodaten handelt. Eine Daten-Metadaten-kopplung ist bei den betrachteten gelisteten 164 Produkten durchgängig gegeben.

2.3 Metadatenmodelle

Metadaten als beschreibende Daten über die eigentlichen Daten beschreiben Herkunft, Ansprechpartner, Format u.v.m. der Daten. Metadaten werden seit jeher genutzt, um die Auffindbarkeit von Daten zu verbessern. Bereits in der rein „analogen“ Welt hatten sich Findbücher und Kataloge in Bibliotheken und Archiven etabliert. Nach bestimmten (Meta-)Datenmodellen wurden Bücher und Archivalien auffindbar gemacht, indem die Karteikarten mit den entsprechenden Metadaten nach bestimmten Systemen sortiert wurden.

Metadatenmodelle sind eine modellhafte Beschreibung bestimmter Aspekte der Erstellung von konzeptuellen Beschreibungsmodellen von Metadaten.

Mit der zunehmenden Digitalisierung wurden und werden digitale Metadatenmodelle und Metadatenformate entwickelt und eingeführt. Über sogenannte Katalogdienste sind über solche Metadaten erfasste Daten recherchierbar. In den verschiedenen Fachdomänen haben sich spezialisierte Metadatenmodelle etabliert. Entsprechend vielfältig sind die Modelle und Formate zur Erfassung und Bereitstellung von Metadaten. Daher war es erforderlich, diejenigen Profile zu identifizieren, die im Rahmen dieser Machbarkeitsstudie näher betrachtet werden konnten. Die Machbarkeitsstudie beschäftigte sich im Schwerpunkt mit Metadaten aus dem Geodatenumfeld, ohne dabei die allgemeinen offenen Daten auszulassen.

Das schränkt allerdings die Auswahl der relevanten Metadatenprofile ein. Zum Thema Metadaten und im speziellen Metadaten zu Geodaten gibt es eine Reihe von Standards der International Standardization Organization (ISO). Dies sind im Einzelnen die Normen ISO 19115, ISO 19119 und ISO 19139. Diese setzen die zur Beschreibung von Geodaten und Geodiensten notwendigen Informationen fest und definieren somit ein Metadatenmodell für Geodaten und -dienste. Gleichfalls definieren die ISO-Standards ein Anwendungsschema im XML-Format (Extensible Markup Language) anhand dessen Geo-Metadaten als XML-Dokumente erfasst und vorgehalten werden können. Dieses Metadatenmodell besitzt im Umfeld von Geodateninfrastrukturen eine weite Verbreitung. So findet das ISO-Metadatenmodell insbesondere im Rahmen der europäischen INSPIRE-Richtlinie Verwendung. Die Verordnungen der INSPIRE-Richtlinie verpflichten die Mitgliedstaaten rechtsverbindlich zur Nutzung eines speziellen INSPIRE-Profil des ISO-Metadatenmodells. Die europäische Geodateninfrastruktur verfügt dadurch über ein ausgereiftes und einheitliches Metadatenmodell. Durch den bereits weit fortgeschrittenen Umsetzungsstand, insbesondere bei der Umsetzung konformer Metadaten, bot es sich für die Machbarkeitsstudie an, die europäische Geodateninfrastruktur INSPIRE zumindest für die Teilregion Deutschland näher zu betrachten. Vor dem Hintergrund des bereits vorhandenen sehr umfangreichen und vor allem im Internet verfügbaren Datenbestandes stellte INSPIRE ein optimales Testfeld für die Machbarkeitsstudie dar.

Im Rahmen der Machbarkeitsstudie sollten allgemeine offene Daten neben der Fachdomäne der Geo-IT ebenfalls berücksichtigt werden. In der internationalen Open Data Community hat sich der vom W3C entwickelte Standard Data Catalog Vocabulary (DCAT) etabliert. Auch hier gibt es Initiativen auf europäischer Ebene, die zur Entwicklung des DCAT-Application Profile (DCAT-AP) geführt haben. In Deutschland wurde im Rahmen der Entwicklung des Portals für offene Verwaltungsdaten GovData ein auf DCAT-AP basierendes Profil entwickelt. Dieses Profil mit der Bezeichnung DCAT-AP.de erweitert die europäische Version von DCAT, DCAT-AP um Felder, die speziell in Deutschland zur Beschreibung von Verwaltungsdaten notwendig sind. Im Juni 2018 hat der IT-Planungsrat für den bestehenden Standardisierungsbedarf "einheitliche Metadatenstruktur für offene Verwaltungsdaten" den Standard DCAT-AP.de als verbindlich beschlossen.

Damit sind in Zukunft sämtliche Verwaltungsdaten mit einem dem Standard DCAT-AP.de entsprechenden Metadatensatz zu beschreiben. Deshalb war es sinnvoll, das Metadatenprofil DCAT-AP.de im Rahmen der Machbarkeitsstudie zu berücksichtigen. Hier bestand die Möglichkeit das Wertungspotential der Machbarkeitstudie erheblich zu erhöhen.

Mit der Betrachtung dieser beiden Metadatenprofile konnte der Umfang der zu betrachtenden Datendomänen bereits vollständig abgedeckt werden. Daher beschränkte sich die Studie darauf, die beiden beschriebenen Standards miteinander zu vergleichen und im Weiteren für das Proof of Concept zu berücksichtigen.

2.3.1 Betrachtung des DCAT-AP.de Metadatenprofils

Das Metadatenprofil DCAT-AP.de dient der Beschreibung von Verwaltungsdaten. Dementsprechend liegt der Fokus nicht in der Abbildung räumlicher Attribute, wie dies beim ISO-basierten Metadatenprofil von INSPIRE der Fall ist. DCAT-AP.de fällt allerdings durch einen schmal gehaltenen Anteil von Pflichtattributen auf und ist dadurch besonders attraktiv für die schnelle Erfassung von

Metadaten. Tabelle 3 zeigt die verpflichtenden Felder für einen DCAT-AP.de-konformen Metadatensatz. Dabei wurde bereits berücksichtigt, das XML-Encoding für DCAT-AP.de zu betrachten, da so eine bessere Vergleichbarkeit mit dem INSPIRE-Metadatenprofil möglich ist.

Xpath	Beschreibung
foaf:Agent/foaf:name	Verantwortliche Stelle
skos:Concept/skos:prefLabel	Kategorie
skos:ConceptScheme/dct:title	Kategorieschema
dcat:Catalog/dct:description	Katalogbeschreibung
dcat:Catalog/dct:publisher	Katalogbereitsteller
dcat:Catalog/dct:title	Katalogbezeichnung
dcat:CatalogRecord/dct:modified	Aktualisierungsdatum
dcat:CatalogRecord/foaf:primaryTopic	Verknüpfung mit der Datenstruktur
dcat:Dataset/dct:description	Datenstrukturbeschreibung
dcat:Dataset/dct:title	Datenstrukturbezeichnung
dcat:Distribution/dcat:accessURL	Distribution

Tabelle 3: Pflichtfelder DCAT-AP.de

2.3.2 Betrachtung des INSPIRE-Metadatenprofils

Die technischen Richtlinien zur Umsetzung von INSPIRE-Metadaten geben derzeit ausschließlich das XML-Encoding zur Implementierung vor. Daher betrachtet die folgende Tabelle entsprechend die XPath's der Attribute des XML-Formates für INSPIRE-Metadaten. Im Vergleich zum DCAT-AP.de-Profil fällt auf, dass die Tiefe der Pfade höher ist, wodurch das INSPIRE-Modell sich komplexer darstellt. Auch besitzt dieses Profil fast doppelt so viele Pflichtfelder als das DCAT-AP.de-Profil. Dadurch ist die Metadatenerfassung mittels dieses Formates aufwändiger. Die Pflichtfelder des INSPIRE-Metadatenmodells zeigt Tabelle 4.

XPath	Beschreibung
gmd:hierarchyLevel/gmd:MD_ScopeCode/@codeListValue	Typ der beschriebenen Ressource
gmd:distributionInfo//gmd:transferOptions//gmd:onLine/gmd:CI_OnlineResource/gmd:linkage/gmd:URL	URL zum Datensatz
gmd:citation/gmd:CI_Citation/gmd:identifier/*/gmd:code	Identifikator für Datensatz
gmd:language/gmd:LanguageCode	Sprache des Datensatzes
gmd:citation/gmd:CI_Citation/gmd:title	Titel der Ressource
gmd:abstract	Beschreibung der Ressource
gmd:topicCategory	Thema des Datensatzes

XPath	Beschreibung
gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:thesaurusName/gmd:CI_Citation	Katalog für Schlagwörter
gmd:descriptiveKeywords/gmd:MD_Keywords	Schlagwörter zum Datensatz
gmd:extent/gmd:EX_Extent/gmd:geographicElement/gmd:EX_GeographicBoundingBox	Räumliche Abdeckung des Datensatzes
gmd:referenceSystemInfo/gmd:MD_ReferenceSystem/gmd:referenceSystemIdentifier/gmd:RS_Identifier	Genutztes Koordinatenreferenzsystem
gmd:spatialRepresentationType/gmd:MD_SpatialRepresentationTypeCode	Geodatentyp
gmd:referenceSystemInfo/gmd:MD_ReferenceSystem/gmd:referenceSystemIdentifier/gmd:RS_Identifier	Genutztes temporales Referenzsystem
gmd:characterSet/gmd:MD_CharacterSetCode	Encoding
gmd:distributionFormat/gmd:MD_Format	Datenformat
gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date	Erstellungsdatum / letzte Revision
gmd:spatialResolution	Räumliche Auflösung
gmd:dataQualityInfo/gmd:DQ_DataQuality/gmd:report/gmd:DQ_DomainConsistency/gmd:result	Infos zur Datenqualität
gmd:lineage/gmd:LI_Lineage/gmd:statement	Qualitätssicherungsverfahren
gmd:accessConstraints/gmd:MD_RestrictionCode	Zugriffsbedingungen
gmd:otherConstraints/gmx:Anchor	Detaillierung der Zugriffsbedingungen
gmd:pointOfContact/gmd:CI_ResponsibleParty/gmd:organisationName	Verantwortliche Stelle
gmd:pointOfContact/gmd:CI_ResponsibleParty/gmd:contactInfo/gmd:CI_Contact/gmd:address/gmd:CI_Address/gmd:electronicMailAddress	Kontakt-E-Mailadresse
gmd:pointOfContact/gmd:CI_ResponsibleParty/gmd:role/gmd:CI_RoleCode	Festlegung der Rolle der verantwortlichen Stelle
gmd:MD_Metadata/gmd:dateStamp	Aktualisierungsdatum

Tabelle 4: Pflichtfelder INSPIRE-Metadatenmodell

2.3.3 Vergleich der beiden Metadatenmodelle

Die Fehler: Verweis nicht gefunden in Anhang III zeigt eine Zusammenschau der Pflichtelemente. Zusätzlich wurden die im jeweils anderen Profil optionalen Entsprechungen zu den Pflichtfeldern hinzugefügt. Hieraus ergibt sich ein mindestens umzusetzender Satz Metadatenfelder, die befüllt werden müssen, um im Rahmen der Machbarkeitsstudie standardkonforme Metadaten zu beiden Metadatenmodellen zu erzeugen.

2.4 Fazit der Datentypisierung

Die Zielsetzung des AP I: Datentypisierung war es, festzustellen, welche Datentypen im Bereich des BMVI genutzt werden und inwieweit sich diese eignen, auf eine Automatisierung der Metadatenherstellung und –pflege hin überprüft zu werden.

Es wurde festgestellt, dass die Daten aus der mCLOUD aufgrund Ihrer hohen Heterogenität quantitativ wie qualitativ für die nähere Betrachtung im Rahmen der Machbarkeitsstudie nur sehr begrenzt in Frage kommen. Es wurde außerdem festgestellt, dass Metadaten bislang vor allem für Geodaten vorhanden sind und dass insbesondere INSPIRE-konforme Datensätze die Anforderungen an die Daten-Metadatenkopplung erfüllen und somit besonders geeignet erscheinen, für KI-basierte Ansätze der Metadatenautomatisierung genutzt zu werden.

Speziell für die Metadaten des CMEMS wurde eine überdurchschnittlich gute Ausstattung mit Metadaten festgestellt. Damit würden sich diese Daten gut für einen automatisierten Vergleich eignen.

Im Rahmen des Arbeitspaketes wurden darüber hinaus verschiedene Metadatenmodelle auf ihre Relevanz und Eignung als Metadatenausgabeformat für den Proof of Concept untersucht. Dabei wurde festgestellt, dass insbesondere die Metadatenmodelle von DCAT-AP.de und INSPIRE in Frage kommen, da sie zum einen gesetzlich verpflichtend zu nutzen sind und zum anderen im Geodatenumfeld bereits eine entsprechende Verbreitung besitzen, was auch durch die durchgeführte Online-Befragung bestätigt wurde.

Die Ergebnisse aus AP I waren essentiell für die weiteren Überlegungen zur Umsetzung der verschiedenen Ansätze der Metadatenautomatisierung, welche im Rahmen des AP II näher betrachtet wurde.

3 AP II: Erarbeitung und Evaluation der Machbarkeit der verschiedenen Ansätze zur Metadatenerneuerung

Im Rahmen der Machbarkeitsstudie wurden verschiedene Ansätze zur Metadatenerneuerung untersucht. Allen gemein war dabei, dass am Ende ein standardkonformer Metadatensatz erzeugt wird. Entsprechend erfolgt hier vor der Beschreibung dieser Ansätze die Definition des Metadatenprofils, welches zur Ausgabe im Rahmen des Proof of Concept umgesetzt wurde. Im Anschluss werden dann die verschiedenen Ansätze zur Metadatenerneuerung näher betrachtet.

3.1 Zielformat für die Metadaten

Die Erkenntnisse aus dem AP I zeigen, dass ein sinnvolles Metadatenausgabeformat möglichst die Pflichtfelder der beiden Modelle aus DCAT-AP.de sowie INSPIRE abdecken soll.

3.2 Datenbetrachtung zur Metadatenerneuerung

In diesem Arbeitspaket stand die Aufgabe, Informationen über einen Datensatz automatisiert in einen Metadatensatz durch reine Betrachtung der Daten zu übertragen. Damit wurden zwei Ziele verfolgt; zum einen die grundsätzliche technische Möglichkeit, Metadaten automatisiert in neue Metadatenprofile zu übertragen, zum anderen bestimmte Metadaten wie Ausdehnung, Encoding, Format (u.a.) durch die Betrachtung eines unbekannten Datensatzes abzuleiten.

Als Beispiel sollte ein Server mit entsprechender, im wesentlichen bereits vorhandener Open Source Software eingerichtet werden. Als Zielverzeichnis wurde ein Geonetwork Open Source Katalog eingerichtet, wobei dieser nur exemplarisch für einen OGC-CSW konformen Metadaten-Katalog stehen sollte und keine Einschränkung der Methode auf diese Software mit sich bringt. Im Prototypen sollte man dann Geodaten über eine Weboberfläche hochladen können. Mit dem Hochladen soll automatisiert ein Metadatensatz im Katalog erzeugt und bereits ein Titel des Datensatzes angelegt werden. Auch dieses Vorgehen ist eher exemplarisch zu verstehen, im Praxisbetrieb einer solchen Lösung muss natürlich ein andersartiger Zugriff auf die mit Metadaten zu versehenden Daten realisiert werden.

Die hochgeladenen Geodaten sollen dann unter Zuhilfenahme verschiedener Open Source Bibliotheken untersucht werden, um exemplarisch Informationen daraus zu ziehen. Diese sollten dann automatisch in den oben genannten Metadatensatz geschrieben werden, der entsprechende Metadatensatz ist nun anschließend direkt im Geonetwork Katalog verfügbar und kann dort über die Oberfläche, aber auch über den OGC CSW Standard abgerufen werden.

Konkrete Beispiele für Informationen, die aus Datensätzen über die Komponente Datenbetrachtung entnommen werden können, sind unter anderem die Bounding Box (bezeichnet das südwestliche und nordöstliche Koordinatenpaar des Gültigkeitsbereiches des Datensatzes), Koordinatenreferenzsystem, das Encoding des Datensatzes und das Dateiformat. Für Daten und Darstellungsdiene ist das Auslesen weiterer Parameter wie GetCapabilities- und GetMap-URL, angebotene Koordinatensysteme usw. denkbar.

Im Rahmen des Prototypes wurden ebenfalls Versuche durchgeführt, um aus den Daten automatisiert einen sogenannten Feature-Katalog zu erzeugen. Ein Feature Katalog beschreibt die im Da-

tensatz enthaltenen Attributfelder sowie deren Datentypen. Unsere Annahme war es, das gerade hierin ein großes Potential der Zeiter sparnis bei der Metadaten-Erstellung sowie bei der Qualitätssteigerung von vorhandenen Metadaten, da diese Feature-Kataloge oft nicht vorhanden sind.

3.3 KI-basierter Ansatz zur Metadatenerstellung

Für Metadatenattribute, die nicht mit einfachen Regeln aus dem zu veröffentlichten Datensatz ausgelesen bzw. abgeleitet werden können, sollten KI-Modelle erstellt werden. Für die Erstellung solcher KI-Modelle müssen entsprechende Daten vorhanden sein. Tabelle 5 zeigt anhand des INSPIRE-Metadatenmodells, ob und wie Metadatenattribute mittels Einsatz von KI automatisiert werden können.

XPath	Beschreibung	Ansatz für automatische Erzeugung
gmd:hierarchyLevel/ gmd:MD_ScopeCode/ @codeListValue	Typ der beschriebenen Res- source	Einfache Ableitung aus Datensatz
gmd:distributionInfo// gmd:transferOptions// gmd:onLine/gmd:CI_OnlineResource/gmd:link- age/gmd:URL	URL zum Datensatz	Einfache Ableitung aus Datensatz
gmd:citation/gmd:CI_Cita- tion/gmd:identifier/*/ gmd:code	Identifikator für Datensatz	Einfache Ableitung aus Datensatz
gmd:language/gmd:Lang- uageCode	Sprache des Datensatzes	KI-Modell möglich: Analyse notwendig
gmd:citation/gmd:CI_Cita- tion/gmd:title	Titel der Ressource	KI-Modell unwahrscheinlich: Für die notwen- dige Kreativität wird mit den vorhandenen Da- ten kein Modell möglich sein.
gmd:abstract	Beschreibung der Ressource	KI-Modell unwahrscheinlich: Für die notwen- dige Kreativität wird mit den vorhandenen Da- ten kein Modell möglich sein.
gmd:topicCategory	Thema des Datensatzes	KI-Modell möglich: Analyse notwendig
gmd:descriptiveKey- words/gmd:MD_Key- words/gmd:thesaurus- Name/gmd:CI_Citation	Katalog für Schlagwörter	KI-Modell möglich: Analyse notwendig
gmd:descriptiveKey- words/gmd:MD_Key- words	Schlagwörter zum Datensatz	KI-Modell möglich: Analyse notwendig

XPath	Beschreibung	Ansatz für automatische Erzeugung
gmd:extent/gmd:EX_Extent/gmd:geographicElement/gmd:EX_GeographicBoundingBox	Räumliche Abdeckung des Datensatzes	Einfache Ableitung aus Datensatz
gmd:referenceSystem-Info/gmd:MD_ReferenceSystem/gmd:referenceSystemIdentifier/gmd:RS_Identifier	Genutztes Koordinatenreferenzsystem	Einfache Ableitung aus Datensatz
gmd:spatialRepresentationType/gmd:MD_SpatialRepresentationType-Code	Geodatentyp	Einfache Ableitung aus Datensatz
gmd:referenceSystem-Info/gmd:MD_ReferenceSystem/gmd:referenceSystemIdentifier/gmd:RS_Identifier	Genutztes temporales Referenzsystem	Einfache Ableitung aus Datensatz
gmd:characterSet/gmd:MD_CharacterSet-Code	Encoding	Einfache Ableitung aus Datensatz
gmd:distributionFormat/gmd:MD_Format	Datenformat	Einfache Ableitung aus Datensatz
gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date/gmd:date	Erstellungsdatum / letzte Revision	Einfache Ableitung aus Datensatz
gmd:spatialResolution	Räumliche Auflösung	Einfache Ableitung aus Datensatz
gmd:dataQualityInfo/gmd:DQ>DataQuality/gmd:report/gmd:DQ_DomainConsistency/gmd:result	Infos zur Datenqualität	Muss manuell ausgefüllt werden
gmd:lineage/gmd:LI_Lineage/gmd:statement	Qualitätssicherungsverfahren	Einfache Ableitung aus Datensatz
gmd:accessConstraints/gmd:MD_RestrictionCode	Zugriffsbedingungen	Einfache Ableitung aus Datensatz
gmd:otherConstraints/gmx:Anchor	Detaillierung der Zugriffsbedingungen	Einfache Ableitung aus Datensatz

XPath	Beschreibung	Ansatz für automatische Erzeugung
gmd:pointOfContact/gmd:CI_ResponsibleParty/gmd:organisationName	Verantwortliche Stelle	Lookup in zentraler Datenbank notwendig
gmd:pointOfContact/gmd:CI_ResponsibleParty/gmd:contactInfo/gmd:CI_Contact/gmd:address/gmd:CI_Address/gmd:electronicMailAddress	Kontakt-E-Mailadresse	Lookup in zentraler Datenbank notwendig
gmd:pointOfContact/gmd:CI_ResponsibleParty/gmd:role/gmd:CI_RoleCode	Festlegung der Rolle der verantwortlichen Stelle	Lookup in zentraler Datenbank notwendig
gmd:MD_Metadata/gmd:dateStamp	Aktualisierungsdatum	Einfache Ableitung aus Datensatz

Tabelle 5: Metadatenattribute und ihre Eignung für die Ableitung mittels KI (in fett) und mittels Datenbetrachtung

3.4 Fazit zu den Ansätzen zur Metadatenerneuerung

Im AP II wurden die Grundlagen dafür geschaffen, welche Möglichkeiten zur Ableitung der Informationen bestehen, Metadatensätze zu Datensätzen zu erstellen, für die es noch keine entsprechende Metadatendokumentation gibt. Dabei wurde deutlich, dass es für die automatisierte Metadatenerneuerung nicht „den einen“ Lösungsansatz gibt, sondern mehrere Ansätze kombiniert werden sollten, um eine umfassende Abbildung der erforderlichen Metadateninformationen zu gewährleisten. In Frage kommt dabei eine Kombination aus der einfachen Ableitung von Informationen aus dem eigentlichen Datensatz und einer auf der Nutzung von Trainingsdaten basierenden KI-Methodik: dem sogenannten „Deep Learning“.

Speziell birgt die automatisierte Erstellung der Feature-Kataloge ein großes Potential, die Nutzung der Daten über beschreibende Feature-Kataloge zu verbessern.

Die Ergebnisse in Bezug auf die Feature-Kataloge sind vielversprechend, eine tatsächliche Realisierung wurde aber im Rahmen der Prototyp-Erstellung nicht umgesetzt.

4 AP III: Deep Learning

4.1 Einführung in das Konzept des Deep Learnings

Deep Learning steht für eine Reihe von verschiedenen Ansätzen im Bereich der KI für die Erzeugung von Modellen, um gewaltige Datenmengen und komplexe Sachverhalte zu verarbeiten. Dabei sind die Deep-Learning-Ansätze eine spezielle Form des maschinellen Lernens (engl. Machine Learning), also von Methoden und technischen Mitteln in den Anwendungsbereichen Data Analytics sowie Natural Language Processing (NLP) (deutsch: natürliche Sprachverarbeitung) (siehe Abbildung 6). Sie ermöglichen erst innovative Anwendungsszenarien wie eine ausgereifte Spracherkennung, semantische Bilderkennung sowie autonomes Fahren.

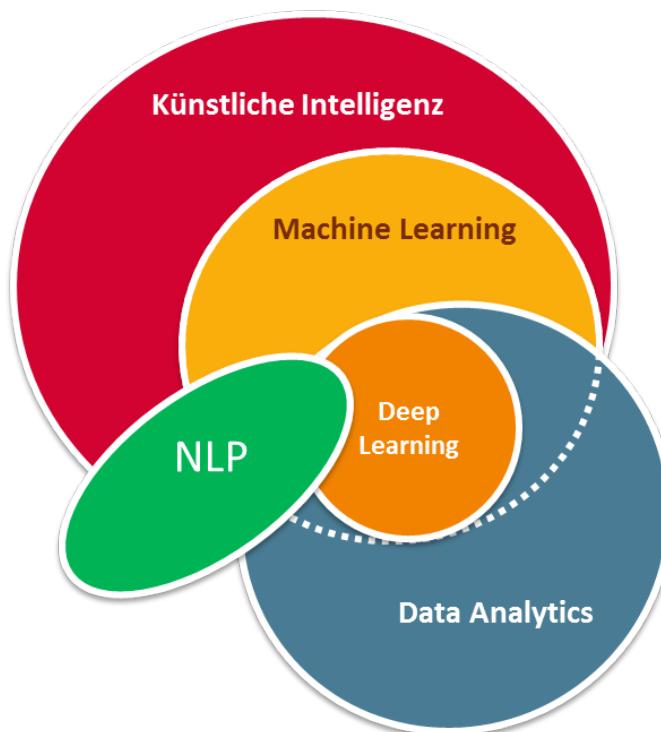


Abbildung 6: Disziplinen der künstlichen Intelligenz

Die Anwendungsbereiche Data Analytics und NLP nutzen eine Reihe von Methoden, wobei nicht alle Machine-Learning-Ansätze verfolgen. Solche Ansätze eignen sich jedoch für komplexere Sachverhalte und haben sich entsprechend in den letzten Jahren etabliert.

Beim Machine Learning geht es im weitesten Sinne darum, mit vorhandenen Daten ein System zu trainieren bzw. ein Modell zu erstellen, welches dann Schlussfolgerungen auf Basis von neuen Daten ziehen kann. Beim NLP geht es darum, Systeme zu befähigen, natürlichsprachliche Texte zu analysieren und zu verarbeiten. Die bekanntesten Anwendungen sind Googles Suchmaschine oder Sprachassistenten wie Siri (Apple) und Alexa (Amazon).

Da es bei der Erzeugung von Metadaten im Kontext von MetaOpenData darum geht, die Ausgangsdaten in Textform zu verstehen, werden hier auch kurz einige Techniken aus dem Bereich des NLP betrachtet.

4.2 Data Analytics

Bei der Verarbeitung von großen Datenmengen werden verschiedene Methoden angewandt, um entsprechende Ergebnisse zu erhalten. Angestrebte Ergebnisse können laut Gartner (Quelle: <https://www.gartner.com/it-glossary/predictive-analytics/>) entweder die Analyse und das Verständnis von Daten aus der Vergangenheit, die Vorhersage von Ereignissen oder die Identifikation künftigen Handlungsbedarfs sein, um ein gewisses Ziel zu erreichen bzw. den Zielerreichungsgrad zu optimieren (siehe Abbildung 7).

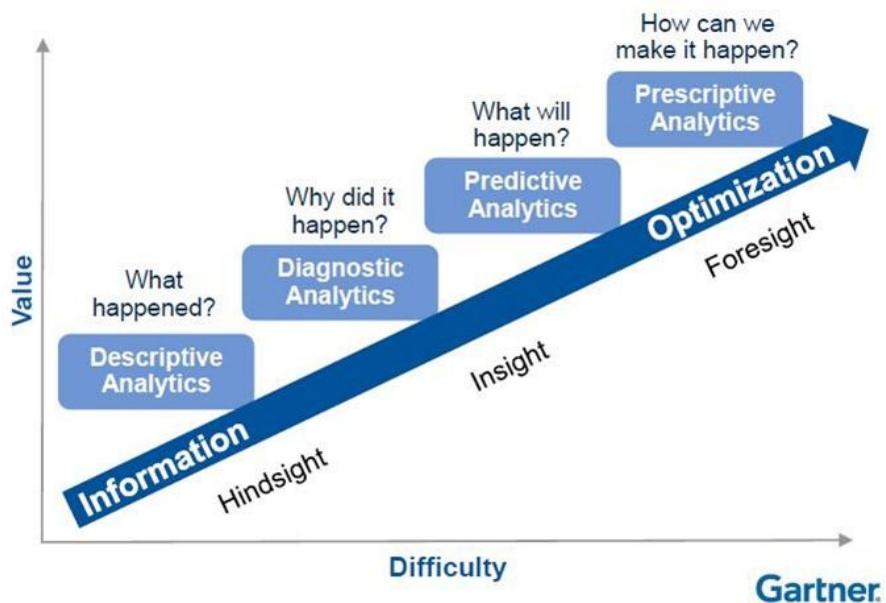


Abbildung 7: Bereiche des Data Analytics

Predictive Analytics ist ein Verfahren, um konkrete Geschäftsziele zu unterstützen. Datenanalysen sollen Verschwendungen identifizieren und dabei helfen, diese zu reduzieren, Zeit zu sparen oder Kosten zu senken. In einem solchen Prozess werden heterogene, häufig sehr große Datenmengen in Modellen abgebildet, die klare, handlungsrelevante Ergebnisse erzeugen können, um die Erreichung der Ziele zu unterstützen, wie z. B. weniger Materialverschwendungen, weniger Lagerbestand oder eine bessere Qualität des hergestellten Produktes.

Predictive Analytics kann in unterschiedlichen Branchen wie der Automobilbranche, Luft- und Raumfahrt, Finanzen und Fertigung eingesetzt werden:

- **Automobilbranche** - Innovationen mit autonomen Fahrzeugen:

Unternehmen, die Fahrerassistenztechnologien und neue autonome Fahrzeuge entwickeln, verwenden Predictive Analytics, um Sensordaten von vernetzten Fahrzeugen zu analysieren und Fahrerassistenz-Algorithmen zu erstellen.

- **Luft- und Raumfahrt** - Zustandsüberwachung für Flugzeugtriebwerke:

Um die Betriebszeit von Flugzeugen zu verbessern und Wartungskosten zu verringern, hat ein Triebwerkshersteller eine Echtzeit-Analyseanwendung erstellt, die die Leistung der Teilsysteme für Öl, Treibstoff, Flugzeugstart, mechanischen Zustand und Steuerung vorher sagt.

- **Finanzdienstleistungen** - Entwicklung von Kreditrisikomodellen:

Finanzinstitute verwenden Machine-Learning-Techniken und -Tools, um Kreditrisiken vorherzusagen.

- **Automatisierung und Maschinenbau** - Vorhersage von Maschinenausfällen:

Ein Kunststoff- und Folienhersteller spart monatlich 50.000 Euro mit einer Anwendung für die Zustandsüberwachung und vorausschauende Instandhaltung, die Ausfallzeiten reduziert und Verschwendungen minimiert.

Prescriptive Analytics kann in bestimmten Szenarien basierend auf den Vorhersagen entsprechende Handlungs- oder Entscheidungsempfehlungen geben. Prescriptive Analytics verwendet prädiktive Modelle, um Aktionen vorzuschlagen, die im Interesse optimaler Ergebnisse durchgeführt werden sollten. Solche Vorschläge beruhen auf Optimierungszielen und regelbasierten Entscheidungsfindungstechniken.

Das grundsätzliche Vorgehen bei Data-Analytics-Ansätze ist wie folgt (siehe Abbildung 8)

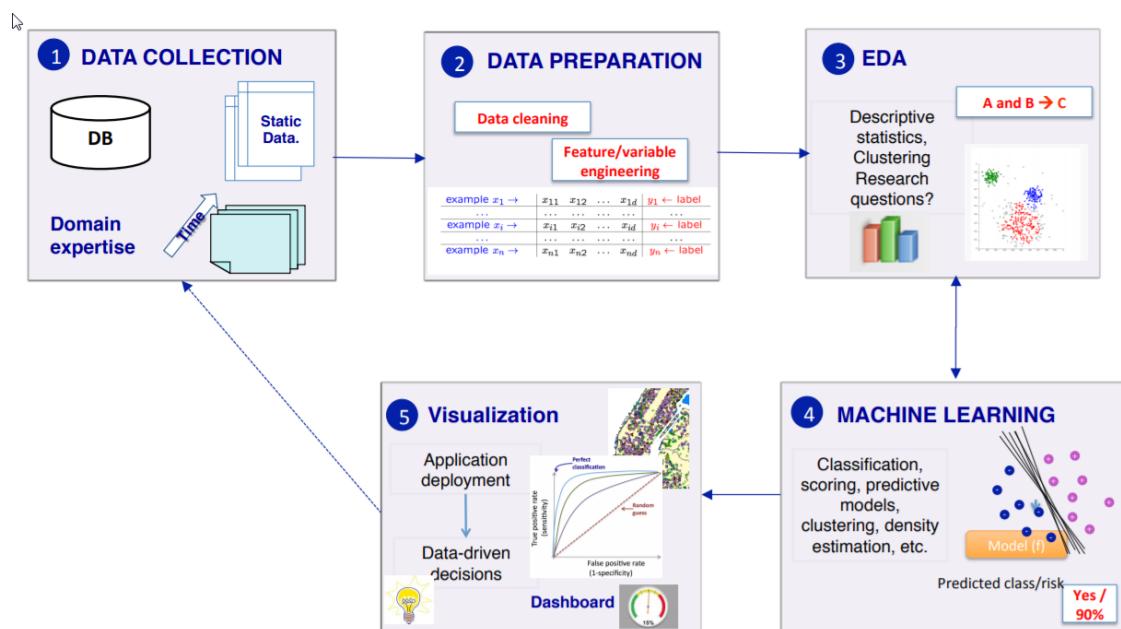


Abbildung 8: Grundsätzliches Vorgehen bei Data Analytics

1. **Daten sammeln:** Je mehr Daten vorhanden sind, desto höher ist die Wahrscheinlichkeit für gute Ergebnisse.

2. **Daten aufbereiten:** Um ein System zu trainieren, müssen Daten entsprechend aufbereitet und strukturiert werden. Die angewandten Methoden sind von den Inhalten der Datenbasis abhängig. Wenn es sich hierbei um keine reinen Zahlen, sondern Texte handelt, kommen NLP-Techniken zum Einsatz.
3. **Daten analysieren:** In strukturierten Daten werden durch Gruppieren oder Klassifizieren Zusammenhänge identifiziert, die bei der Erstellung eines Modells unterstützen. Hierbei werden verschiedene Cluster-Techniken auf die vorstrukturierten Daten angewendet wie z.B.
 - a. Partitionierende Verfahren
 - b. Hierarchische Verfahren
4. **Modell erstellen:** Ein Modell wird aus den vorhandenen Daten abgeleitet. Bei einigen Ansätzen wird auch ein Teil der Daten (z. B. 30%) zum Testen und Verifizieren des Modells verwendet. Hierbei kommt eine Vielzahl von Ansätzen zur Erstellung solcher Modelle in Frage, für die es viele verschiedene Gruppierungen gibt. Sie können auch grob eingeteilt werden in:
 - a. Regression
 - b. Machine Learning
5. **Anwendung und Visualisierung:** Das Modell wird angewendet. Die Ergebnisse können gemessen und visualisiert werden, um weitere Erkenntnisse zu gewinnen. Ggf. werden neue Daten gesammelt, die zur Weiterentwicklung der Modelle herangezogen werden.

Diese Schritte wurden auch bei MetaOpenData durchgeführt. Die Ergebnisse der Schritte 1 bis 3 sind in Kapitel 5.3 dargestellt, die Ergebnisse aus Schritt 4 werden in Kapitel 6.2 erläutert.

4.3 NLP

Um die Verarbeitung von natürlicher Sprache zu ermöglichen, geht es bei NLP darum, die Struktur sowie den Sinn eines Textes zu verstehen. Hierbei erweitert NLP die Textanalyse-Methoden, bei welchen zunächst die Analyse von Strukturen und Klassifikationen im Vordergrund steht. Ziel ist die Automatisierung von Aufgaben.

NLP (Natural Language Processing)	Text Mining or Text Analytics
Künstliche Sprache (Text-To-Speech)	Automatische Gruppierung (n grams)
Automatische Texterzeugung	Automatische Klassifikation (Bag of Words)
Automatische Übersetzung	Identifizierung von Patterns

Tabelle 6: Natural Language Processing

Folgende Techniken können u.a. auf einen Text angewandt werden:

- **Language Detection:** Um Texte zu bearbeiten muss in den meisten Fällen die Sprache des Textes identifiziert werden.

- **Segmentation:** Text in Abschnitte und Sätze segmentieren. Meist ist ein Satz durch ein Abschlusszeichen getrennt wie ein Punkt (.) oder Ausrufezeichen (!).
- **Tokenization:** Es werden Wörter, Zahlen oder andere Trennzeichen und Symbole identifiziert.
- **Stemming/ Lemmatization:** Hierbei wird der Wortstamm identifiziert. Beim Stemming werden Wortendungen entfernt, um damit weiter zu arbeiten. Im Englischen werden z.B. Wörter wie 'eating' zu 'eat' reduziert. Diese Methode muss für jede Sprache individuell angepasst werden.
- **Part of speech (POS) tagging:** Jedem Wort in einem Satz wird die entsprechende grammatische Bedeutung (Subjekt, Prädikat oder Verb) zugeordnet.
- **Parsing:** Basierend auf der POS-Zuordnung wird ein Syntax-Baum für einen Satz erstellt. Somit kann die syntaktische Korrektheit des Satzes überprüft werden.
- **Named Entity Recognition:** Dadurch können spezielle Objekte wie Personen und Orte sowie Zeiten im Text identifiziert werden.
- **Co-Reference resolution:** Hierbei werden Beziehungen von Wörtern satzübergreifend identifiziert.

Dies wird anhand der in Tabelle 7 folgenden Beispiele dargestellt.

Technik	Beispiel	Output
Sentence Segmentation	Mark met the president. He said:"Hi! What's up -Alex?"	<ul style="list-style-type: none"> • Sentence 1: Mark met the president. • Sentence 2: He said: "Hi! What's up – Alex?"
Tokenization	My phone tries to 'charging' from 'discharging' state.	<ul style="list-style-type: none"> • [My] [phone] [tries] [to] ['] [charging] ['][from] ['][discharging] ['] [state][.]
Stemming/ Lemmatization	Drinking, Drank, Drunk	<ul style="list-style-type: none"> • Drink
Part-of-Speech tagging	If you build it he will come.	<ul style="list-style-type: none"> • IN – prepositions and subordinating conjunctions. • PRP – Personal Pronoun • VBP – Verb Noun 3rd person singular present form. • PRP- Personal pronoun • MD – Modal Verbs • VB – Verb base form
Parsing	Mark and Joe went into a bar.	<ul style="list-style-type: none"> • (S(NP(NP Mark) and (NP(Joe))) • (VP(went (PP into (NP a bar))))
Named Entity Recognition	Let's meet Alice at 6 am in India.	<ul style="list-style-type: none"> • Person: Alice • Time: 6 am • Location: India
Coreference resolution	Mark went into the mall. He thought it was a shopping mall.	<ul style="list-style-type: none"> • [Mark] went into the mall. [He] thought it was a shopping mall. • [Mark] thought it was a shopping mall.

Tabelle 7: Beispiele zum NLP

4.4 Machine Learning

Zusätzlich zu den klassischen NLP-Techniken werden zur Verarbeitung von komplexen Texten auch Machine-Learning-Ansätze eingesetzt, um Modelle zu erzeugen (siehe Abbildung 9). Diese Modelle werden mit vorhandenen Daten trainiert und können dadurch komplexere Sachverhalte bearbeiten, die nur mit großem Aufwand regelbasiert zu lösen wären (Quelle: u.a. www.xenon-stack.com).

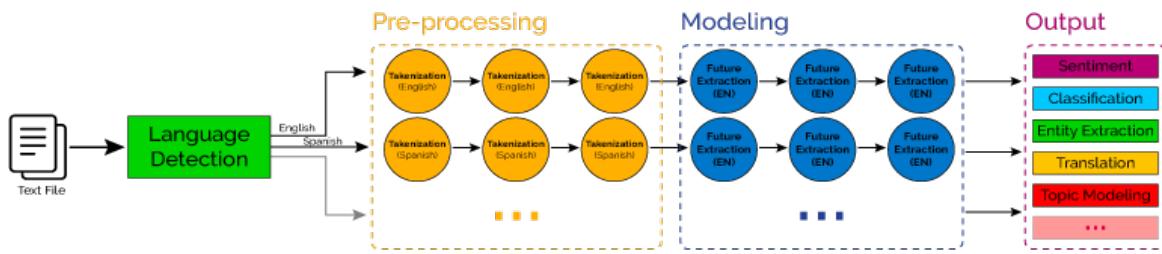


Abbildung 9: Klassische NLP Verfahren

Folgende ML – Modelle sind weit verbreitet:

1. Decision Tree:

Entscheidungsbäume (engl. Decision Trees) sind eine Methode zur automatischen Klassifikation von Datenobjekten und damit zur Lösung von Entscheidungsproblemen. Ein einfacher Baum beantwortet die Frage, ob ein Baum Früchte trägt anhand der Ausprägung von max. 3 Attributen (siehe Abbildung 10): Alter, Sorte, Boden. Der Algorithmus erkennt in den Ausgangsdaten, dass bestimmte Attribut-Werte in bestimmten Konstellationen irrelevant sind. Diese sind in der Abbildung mit „xxx“ gekennzeichnet.

Früchte	Alter	Sorte	Boden
nein	jung	xxx	xxx
ja	alt	veredelt	xxx
ja	alt	natürlich	reichhaltig
nein	alt	natürlich	mager

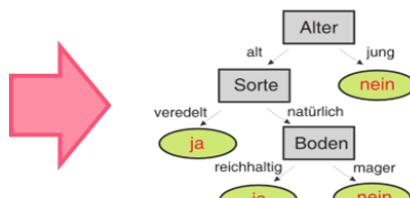


Abbildung 10: Entscheidungsbaum

2. Naive Bayes:

Der Naive Bayes - Klassifikator gehört zur Familie einfacher „wahrscheinlicher Klassifikatoren“, die auf der Anwendung des Satzes von Bayes mit starken (naiven) Unabhängigkeitsannahmen zwischen den Merkmalen basieren.

Die naive Grundannahme ist dabei, dass jedes Attribut nur vom Klassenattribut abhängt. Solange die Attribute nicht zu stark korrelieren, erzielen naive Bayes-Klassifikatoren bei praktischen Anwendungen häufig gute Ergebnisse, obwohl die Grundannahme selten die Realität abbildet.

3. Random Forest:

Ein Random Forest ist ein Klassifikationsverfahren, das aus mehreren nicht korrelierten Entscheidungsbäumen besteht. Alle Entscheidungsbäume werden auf eine bestimmte Art zufällig erzeugt und wachsen im Trainingsprozess. Für eine Klassifikation darf jeder Baum in diesem Wald eine Entscheidung treffen und die Klasse mit den meisten Stimmen entscheidet die endgültige Klassifikation.

Um die Modelle zu trainieren, werden die zur Verfügung stehenden Referenzdaten in Trainings- und Testdaten aufgeteilt. Dazu wird bspw. ein Verhältnis von 70% Trainingsdaten und 30% Testdaten verwendet. Mithilfe der Testdaten kann die Genauigkeit des Modells überprüft werden. Dabei können zu große Ungenauigkeiten festgestellt werden.

Bei diesen Ansätzen kann es auch zu einem sogenannten „overfitting“ (Deutsch: Überanpassung) kommen, wenn das Modell sich zu sehr an den Trainingsdaten orientiert und bei unbekannten Konstellationen keine sinnvollen Vorhersagen mehr machen kann.

4.5 Deep Learning

Im Gegensatz zu klassischen NLP-Techniken mit Machine Learning werden bei Deep-Learning-Ansätzen entsprechend künstliche neuronale Netze (KNN) erzeugt, die vom Grundprinzip her nicht sprachspezifisch implementiert werden (siehe Abbildung 11) müssen. Je nach Ausgangsdaten (Trainingsdaten) und Einsatzszenario muss jedoch das KNN verschiedene Eigenschaften besitzen, um einen sinnvollen Output zu erzeugen (Quelle: verschiedene Internetseiten, u.a. www.xeonstack.com).

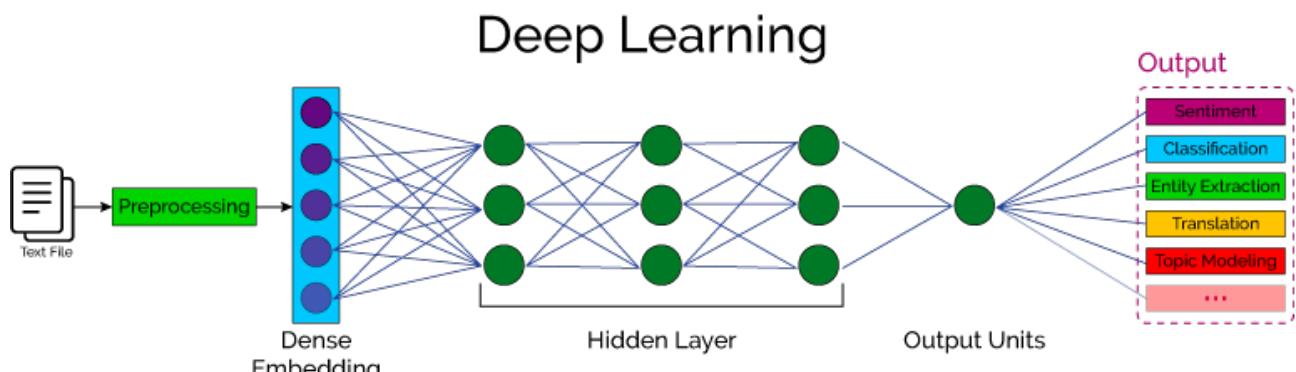


Abbildung 11: Schematische Darstellung von NLP mit Deep Learning

Ein KNN besteht aus einer Menge an Perzeptoren. Ein Perzeptron besteht aus zwei Komponenten: der sogenannten gewichteten Summe, also der Summe von gewichteten Eingabewerten, sowie der Aktivierungsfunktion (siehe Abbildung 12):

- **Gewichtete Summe:** Der Wert wird aus den Eingabewerten (x_1 bis x_n) gebildet, die jeweils mit ihren Gewichten (w_1 bis w_n) multipliziert werden. Diese Gewichte werden anfangs zufällig ausgewählt und durch den Trainingsvorgang basierend auf den Trainingsdaten sukzessive angepasst.

- Aktivierungsfunktion:** Die Aktivierungsfunktion bestimmt den Ausgabewert (o). Im einfachsten Fall überprüft eine Schwellwertfunktion die Summe. Sollte die Summe kleiner 0 sein wird auch 0 ausgegeben, ansonsten ist der Ausgabewert 1.

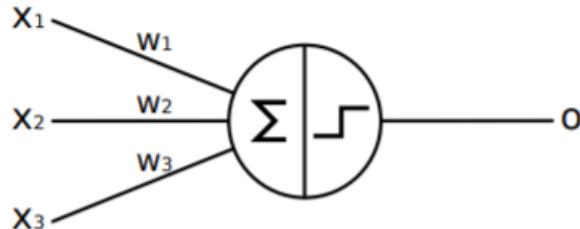


Abbildung 12: Ein Perzepron

Durch die Kombination von mehreren Perzeptren können komplexere Netze erzeugt werden, die auch nicht lineare Probleme lösen können. Das Multi-Layer Perzepron (kurz MLP) gehört zu den bekanntesten Ansätzen künstlicher neuronaler Netze. Es besteht aus einer Vielzahl von Perzeptren in mehreren Schichten, wobei jedes Perzepron einer Schicht mit jedem Perzepron der folgenden Schicht verbunden ist (siehe Abbildung 13). Dies wird auch ein „fully connected feedforward neuronal network“ genannt. Hierbei wird in Input Layer (deutsch Eingabeschicht), Hidden Layer (deutsch versteckte Schicht) sowie Output Layer (deutsch Ausgabeschicht) unterschieden. Jede Verbindung besitzt ein eigenes Gewicht ($w_{i,j}$), welches mit dem jeweiligen Eingabewert multipliziert wird.

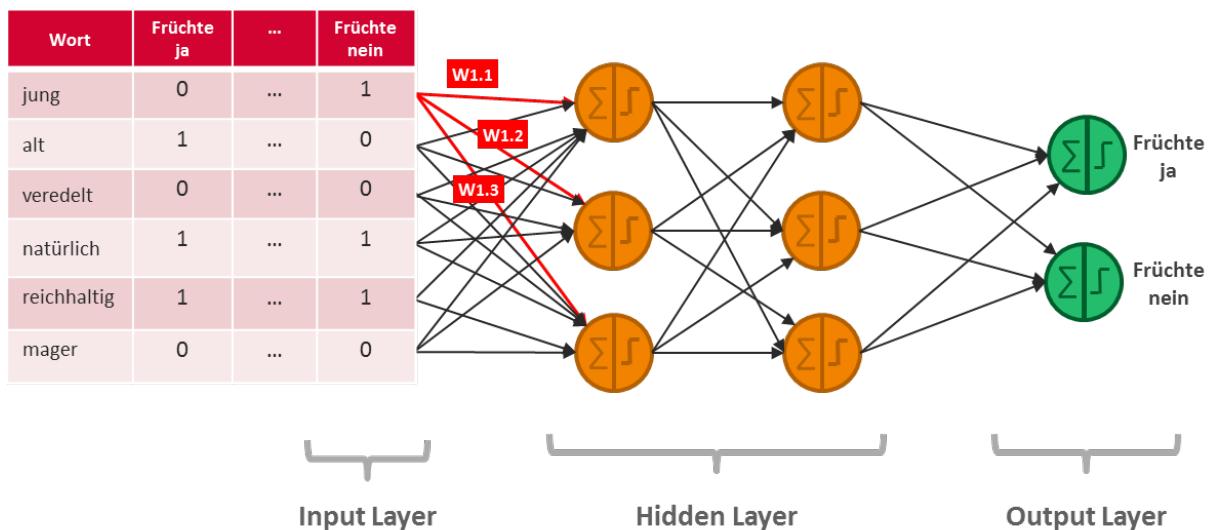


Abbildung 13: Multi Layered Perception (MLP)

In der Abbildung 13 ist das Beispiel, wann ein Baum Früchte trägt, dargestellt. Angenommen, es soll aus einer beliebigen Beschreibung eines Baumes in Textform erkannt werden, ob dieser Früchte trägt oder nicht. Um ein MLP zu trainieren müssen viele verschiedene bereits klassifizierte

Beschreibungen von Bäumen vorliegen, die entsprechende Begriffe enthalten bzw. nicht enthalten. Diese vorklassifizierten Texte werden eingelesen, wobei die Begriffe in einem Preprocessing-Schritt herausgefiltert und in einer einheitlichen Struktur zusammengeführt werden.

Im Trainingsprozess durchlaufen jeweils alle Werte einer Spalte das Netz. Dazu werden initial allen Gewichten ein zufälliger Wert ($-1 \leq w_{i,j} \leq 1$) zugewiesen. Sofern die Aktivierungsfunktion der Perzeptren im Output Layer nicht das erwartete Ergebnis widerspiegelt, werden alle Gewichte minimal korrigiert, bevor die Werte der nächsten Spalte das Netz durchlaufen. In dem Beispiel oben (Abbildung 13) wird für die rechte Spalte der Tabelle die errechnete Summe im Output Layer bei Früchte ja < 0 erwartet. Sollte der Wert > 0 sein, werden die Gewichte rückwirkend angepasst (engl. backpropagation).

Bei diesem sehr einfachen Beispiel ist schon zu erkennen, dass je nach Datenumfang und Netzgröße recht komplexe Berechnungen durchgeführt werden, die nur sehr schwer vorhersagbar sind. Um festzustellen, wie gut ein Netz trainiert ist, werden meist bei 30% der Ausgangsdaten keine Anpassungen der Gewichte durchgeführt. Der Vergleich von Vorhersagen mit realen Werten ermöglicht eine Aussage darüber, mit welcher Wahrscheinlichkeit das aktuell trainierte Netz mit unbekannten Daten eine korrekte Vorhersage treffen wird.

Es gibt eine Vielzahl an Möglichkeiten die Funktionsweise eines neuronalen Netzes zu beeinflussen, wie z.B. das Ändern der Aktivierungsfunktion, die Vorgehensweise bei der Anpassung der Gewichte oder das Ändern der Anzahl von Layern und der jeweils enthaltenen Neuronen.

Es kann aber auch die gesamte Struktur des Netzes verändert werden. Die bekanntesten Formen sind (siehe Abbildung 14):

- **Feedforward NN:** Die Hidden Layer sind immer gleich aufgebaut. Die Berechnung der folgenden Ebene basiert auf den Ergebnissen der vorherigen Ebene. Mit dieser Netzstruktur können gute Vorhersagemodelle basierend auf strukturierten Daten erstellt werden.
- **Convolutional NN:** Die Unterschiede zu einem Feedforward sind im Wesentlichen:
 - 2D- oder 3D-Anordnung der Neuronen
 - Geteilte Gewichte
 - Lokale Konnektivität der Perzeptren innerhalb eines Layers

Mit einem solchen Aufbau können Bilddaten effizienter verarbeitet werden.

- **Recurrent NN/LSTM:** In diesen Netzen gibt es verschiedene Arten von Rückkopplungen innerhalb eines oder mehrerer Layer. Solche Konstruktionen ermöglichen die Verarbeitung von Sequenzen, bei denen Zusammenhänge eine gewisse Rolle spielen. Beispiele dafür sind Handschrifterkennung, Spracherkennung und Maschinenübersetzung.

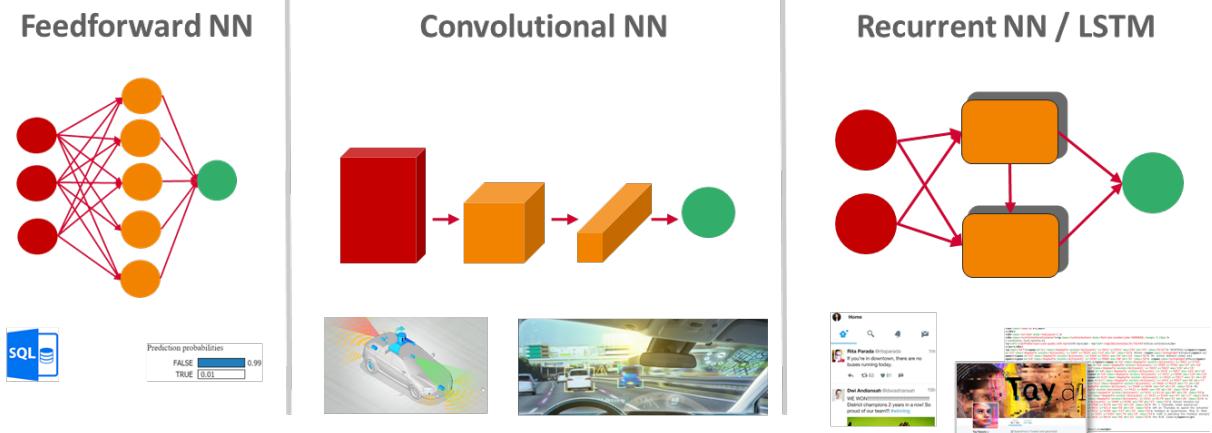


Abbildung 14: Arten von künstlichen neuronalen Netzen (KNN)

Zuletzt sei noch auf die verschiedenen Lernverfahren hingewiesen:

- **Überwachtes Lernen (supervised learning)**
Hier werden die Ausgabewerte mit dem erwarteten bzw. vorgegebenem Wert verglichen, woraufhin ggf. entsprechend die Gewichte angepasst werden (siehe Beispiel oben).
- **Unüberwachtes Lernen (unsupervised learning)**
Das neuronale Netz verändert sich entsprechend den Eingabemustern von selbst. Hierbei kommen komplexere Lernregeln zum Einsatz wie z.B. Adaptive Resonanztheorie oder die Hebb'sche Lernregel.
- **Bestärkendes Lernen (reinforced learning)**
Beim bestärkenden oder verstärkenden Lernen führt z.B. ein Agent einen Test durch. Auf Grund von einem verzögerten Feedback (positiv oder negativ) wird entsprechend das Netz angepasst.

5 AP IV: Entwicklung eines Proof of Concept

Im Rahmen der Machbarkeitsstudie wurde ein Proof of Concept durchgeführt. Dazu wurde ein Prototyp entwickelt, der auf der Basis der Erkenntnisse aus den vorangegangenen Arbeitspaketen konzipiert wurde.

Im Folgenden werden die Komponenten des Prototypen vorgestellt.

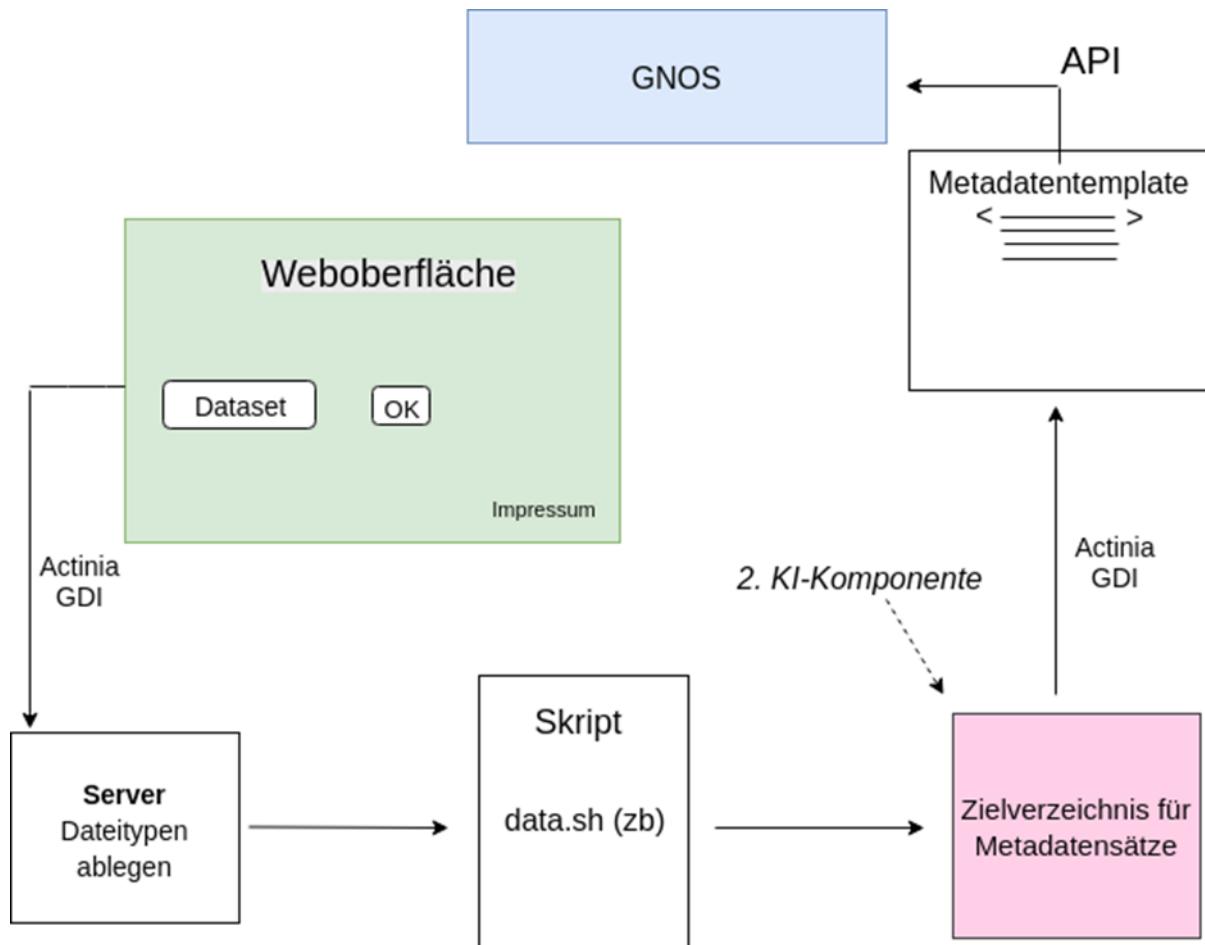


Abbildung 15: Abläufe zur automatisierten Metadatenerzeugung

Die Abbildung 15 zeigt die Abläufe im Prototypen, die zu Erzeugung von Metadatensätzen durchgeführt werden. Dieses Konzept kann ausgebaut als Vorlage für eine spätere Realisierung dienen.

Wie die vorangegangenen Kapitel gezeigt haben, können Metadaten grundsätzlich nach zwei unterschiedlichen Verfahren generiert werden:

- Durch Ableitung aus den Daten selbst, wie dies im unteren Teil der Abbildung zu erkennen ist, und
- über Deep-Learning-Methoden, wobei für die Beschaffung von benötigten „Trainingsdaten“ sogenannte Harvester verwendet werden. Diese beziehen automatisch Metadaten aus ex-

ternen Quellen. Dabei nutzen sie standardisierte CSW-Dienste (gemäß OGC-Standard „Catalogue Service for the Web“), dargestellt im rechten Bildbereich.

Im Prototyp wurden die beiden Verfahren in einem hybriden Ansatz kombiniert. Abbildung 16 gibt eine Übersicht über die Systemarchitektur und das Zusammenwirken der einzelnen Komponenten. Wichtig ist erneut der Hinweis, dass die Komponenten aufgrund von Verfügbarkeit (Open Source) ausgewählt wurden und dass diese – zumindest zum Teil – in einer späteren wirklichen Lösung ausgetauscht werden können bzw. die Architektur flexibler auch andere Komponenten einbinden kann.

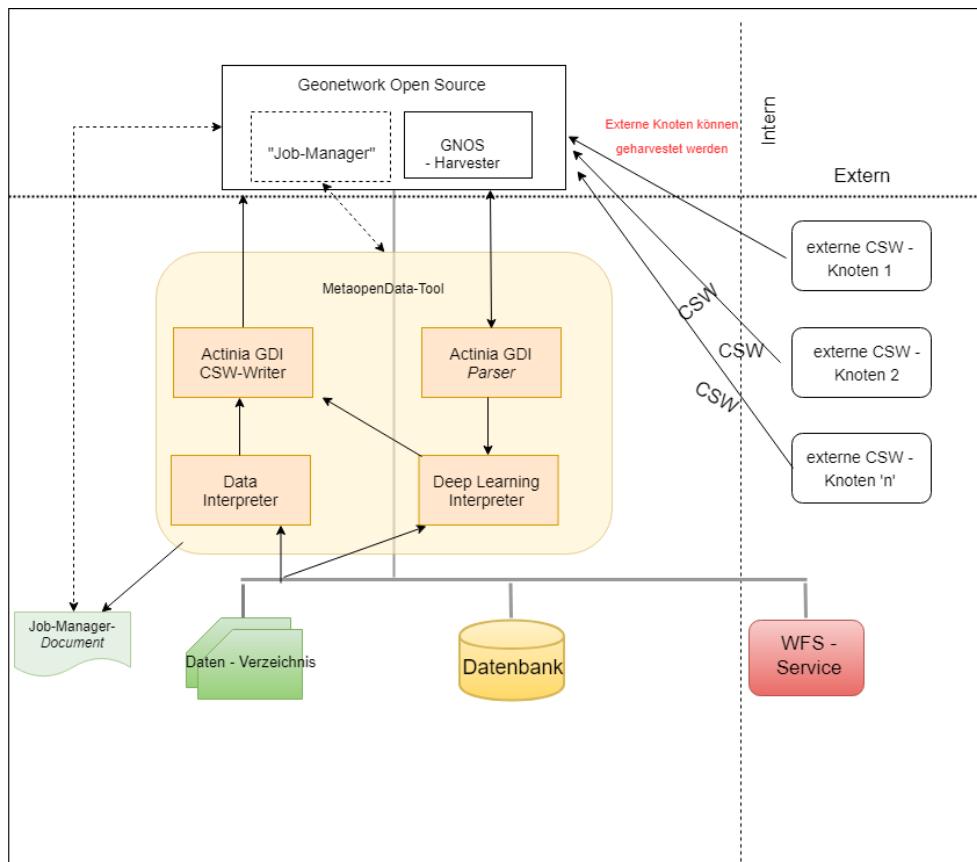


Abbildung 16: Architektur Prototyp

Insgesamt besteht das System aus vier Komponenten. Darüber hinaus besitzt das System Schnittstellen zu verschiedenen weiteren Systemen und Diensten, bspw. zu einem für das Projekt betriebenen CSW-Knoten auf der Basis von GeoNetwork Open Source.

Insgesamt wurde so eine moderne, Micro-Service-basierte Architektur entworfen, die auf bestehende Standards zurückgreift bzw. sich diese zunutze macht.

5.1 Data-Interpreter

Beim ersten der genannten Verfahren (der Ableitung der Metadaten aus den Daten selbst) werden bspw. die sog. Header-Dateien aus Rasterdaten (bspw. Sentinel Satellitendaten) ausgelesen, aber

auch die Daten selber betrachtet. Daraus können bereits relevante Informationen über Datentyp (bspw. Geotiffs), räumliche Auflösung (Pixelgröße), räumliche Abdeckung oder die Projektion der Daten extrahiert werden. Der sogenannte „Data-Interpreter“ extrahiert derartige Informationen aus den Daten, wobei der Zugriff daraus im Prototyp lediglich für hochgeladene Vektordaten umgesetzt wurde. Eine Ausweitung auf Daten und Dienste aus anderen Quellen und -formaten wie Datenbanken, Datenverzeichnissen oder über Geodatendienste wie Web Map Service (WMS), Web Feature Service (WFS) und Web Coverage Services (WCS) ist ohne weiteres möglich.

5.2 GNOS-Harvester

Für das zweite Verfahren, die Ableitung von Metadaten durch Deep Learning werden Trainingsdaten benötigt. Um Metadaten als Trainingsdaten zu bekommen, wird ein sogenanntes Harvesting verwendet, dies ist ein Verfahren, durch das Metadaten von öffentlichen Metadatenknoten „geerntet“ und zur Verwendung in einen eigenen Metadatenknoten kopiert werden. In unserer Architektur wurde diese Aufgabe von der Komponente Geonetwork Open Source (GNOS) bewerkstelligt und als Harvesting-Server das Geoportal Hamburg, dass bereits umfassende Metadaten zu verschiedenen Daten bereitstellt verwendet. In den Katalogen bzw. Knoten sind Metadaten zu gleichen Datentypen in ggf. unterschiedlichen Metadatenprofilen hinterlegt. Ein solcher Katalogdienst analysiert so die Inhalte von anderen Katalogen und kopiert diese Metadaten über Geodatendienste und Geodaten. Damit können Nutzer dieser Metadaten-Knoten herausfinden, welche Karten- (WMS) oder Daten (WFS)-Dienste für bestimmte Schlagwörter oder Regionen existieren und welche Fähigkeiten bzw. Eigenschaften diese besitzen. Ein CSW liefert Metadatensätze als Abfrageergebnisse (z.B. in XML), die der ActiniaGDI-Harvester ausliest und dem Deep-Learning-Interpreter als Eingangs-Trainingsdaten zur Verfügung stellt.

5.3 Deep-Learning-Interpreter

Der „Deep-Learning-Interpreter“ ist ein KI-Verfahren, das anhand bestehender (Meta-) Daten lernen kann, Objekte (Daten) zu identifizieren bzw. Daten automatisch mit geeigneten Metadaten zu beschreiben, also Metadaten selbstständig – ohne manuelle Eingabe – zu generieren. Dazu wird jedoch eine ausreichend große Menge Trainingsdaten benötigt. Durch das Abrufen mehrerer Kataloge sollen daher einerseits relevante Datenbestände aufgefunden und andererseits ausreichend umfassende Trainingsdaten für den Deep-Learning-Interpreter generiert werden.

Auf der Basis der Ergebnisse aus den vorangegangenen Arbeitspaketen wurde sich dafür entschieden, als Datenbasis INSPIRE-Daten zu nutzen. Hier wurden entsprechende Teile der GDI-DE betrachtet und letztlich entschieden, den Hamburger CSW-Knoten als Datenbasis für die KI-Verfahren zu nutzen, da hier die Quote von Daten-Metadaten-Kopplungen besonders hoch erschien.

Im Rahmen der Umsetzung des Deep-Learning-Interpreters wurden die Daten analysiert, um die KI-gestützte automatisierte Erzeugung einzelner Metadaten-Attribute für folgende INSPIRE- Attribute umzusetzen.

XPath	Beschreibung	Analyse
gmd:language/gmd:Language-Code	Sprache des Datensatzes	Wie kann eine Sprache erkannt werden?
gmd:topicCategory	Thema des Datensatzes	Welche der 19 TopicCategoryCode- Ausprägungen (Quelle: http://inspire.ec.europa.eu/metadata-codelist/TopicCategory:2) können vorhergesagt werden?
gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:thesaurusName/gmd:CI_Citation	Katalog für Schlagwörter	Wie kann der zu verwendende Katalog vorhergesagt werden?
gmd:descriptiveKeywords/gmd:MD_Keywords	Schlagwörter zum Datensatz	Welche Schlagwörter können vorhergesagt werden?

Tabelle 8: INSPIRE- Attribute

Hinweis: Bei der Sichtung der Daten wurde deutlich, dass für die Fragen rund um Schlagwörter keine ausreichende Datenbasis vorhanden war.

5.3.1 Daten sammeln

Zur Analyse und Erstellung der Modelle wurden Geodateien im .gml-Format sowie entsprechende INSPIRE Metainformationen im .xml-Format über WFS-GetFeature- bzw. CSW-Aufrufe vom Harvesting-Server heruntergeladen. Die Dokumentenstruktur ist wie folgt:

Name der Datei	Datei-Endung	Format	Anzahl
HH_WFS_Hoehenpunkte-1B253E9C-CF57-4727-AD23-FA809B9A8EB3-hoehenfestpunkte-2.0.0	.gml	XML	2.556
HH_WFS_Hoehenpunkte-1B253E9C-CF57-4727-AD23-FA809B9A8EB3-hoehenfestpunkte-2.0.0	.xml	XML	2.054

Tabelle 9: Dokumentenstruktur

5.3.2 Daten aufbereiten

Für eine Analyse der Daten müssen diese in einer einheitlichen Struktur vorliegen. Wenn die XML-Dateien einheitlich aufgebaut sind kann dazu die XML-Struktur in eine übergreifende Tabellenstruktur überführt werden, die dann mit den Werten der jeweiligen Dateien gefüllt werden (siehe Abbildung 17).

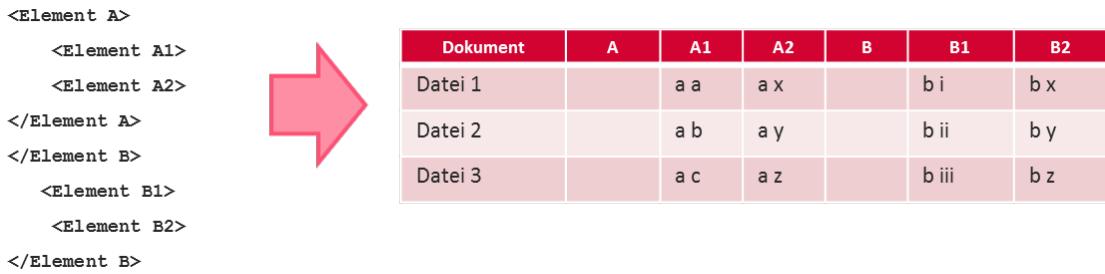


Abbildung 17: Erzeugung einer übergreifenden Datenstruktur

Da jedoch die Elemente der GML-Dateien nicht einheitlich sind und sehr stark variieren, können diese nicht in eine übergreifende Struktur überführt werden, die sich aus den XML-Elementen ableitet, wie in den zwei folgenden Beispielen dargestellt:

Nr	Ausschnitt der .gml-Datei
1	<wfs:member> <app:hoehenfestpunkte xmlns:app="http://www.deegree.org/app" gml:id="APP_HOEHENFESTPUNKTE_1"> </wfs:member> <wfs:member> <app:hoehenfestpunkte xmlns:app="http://www.deegree.org/app" gml:id="APP_HOEHENFESTPUNKTE_2"> </wfs:member>
2	<wfs:member> <app:swis_sensoren xmlns:app="http://www.deegree.org/app" gml:id="APP_SWIS_SENSOREN_1"> </wfs:member> <wfs:member> <app:swis_sensoren xmlns:app="http://www.deegree.org/app" gml:id="APP_SWIS_SENSOREN_2"> </wfs:member>

Tabelle 10: Elemente der GML-Datei

Während in der ersten .gml-Datei das XML-Element `<app:hoehenfestpunkte>` verwendet wird, kommt in der zweiten Datei stattdessen `<app:swis_sensoren>` vor. Daher genügt es nicht, die Values (deutsch: Werte) der XML-Elemente zu vergleichen. Es muss stattdessen die Verwendung der Tag-Bezeichnungen selbst zum Verstehen der Inhalte analysiert werden. Aus diesem Grund wurden die Inhalte der .gml-Dateien ausgelesen und in einer flachen Struktur zusammengefasst (siehe Abbildung 18).

	A	B	C	D	E	F
1	Datei	Dateityp	Text	Datum	Größe	MD_TopicCategoryCode
2	DE_HH_INSPIRE_WFS_Dienststellenstand.gml		tname Autobahnmeisterei Stiilhorn dstname dsttyp dsttyp strasse Altenfelder stras	2018-07-23 15:15:27	4578,0	utilitiesCommunication
3	DE_HH_INSPIRE_WFS_SVZ_Zaehlstellen.gml		REICHE GEOM srsName EPSG posList posList LineString geom zaehlstellenbereiche	2018-07-23 15:43:50	5304,0	environment
4	DE_HH_INSPIRE_WFS_SVZ_Zaehlstellen.gml		REICHE GEOM srsName EPSG posList posList LineString geom zaehlstellenbereiche	2018-07-23 15:43:50	5304,0	boundaries
5	DE_HH_INSPIRE_WFS_SWIS_Sensoren-5.gml		: xmlns http://deegree.SWIS SENSOREN kennung kennung bezeichnung bezeichnung s	2018-07-23 14:53:33	4777,0	structure
6	DE_HH_WFS_INSPIRE_A1_5_Verkehrsne.gml		:enherkunft europastrasse europastrasse geom Inlined geometry STRASSENNETZ IN	2018-07-23 14:45:22	7334,0	transportation
7	DE_HH_WFS_INSPIRE_A1_7_Verkehrsne.gml		:is http://deegree.STRASSENNETZ INSPIRE BESCHR strasse strasse geom Inlined geom	2018-07-23 14:45:22	6306,0	transportation
8	DE_HH_WFS_INSPIRE_A1_7_Verkehrsne.gml		:enherkunft laengenherkunft europastrasse europastrasse geom Inlined geometry S	2018-07-23 14:45:22	7196,0	transportation
9	DE_HH_WFS_INSPIRE_A1_7_Verkehrsne.gml		:is http://deegree.STRASSENNETZ INSPIRE BESCHR strasse strasse geom Inlined geom	2018-07-23 14:45:22	5645,0	transportation
10	DE_HH_WFS_INSPIRE_A1_7_Verkehrsne.gml		:unft laengenherkunft europastrasse europastrasse geom Inlined geometry STRASS	2018-07-23 14:45:23	7458,0	transportation
11	DE_HH_WFS_INSPIRE_A1_7_Verkehrsne.gml		:is http://deegree.STRASSENNETZ INSPIRE BESCHR strasse strasse geom Inlined geom	2018-07-23 14:45:23	6004,0	transportation
12	DE_HH_WFS_INSPIRE_A1_7_Verkehrsne.gml		:enherkunft europastrasse europastrasse geom Inlined geometry STRASSENNETZ IN	2018-07-23 14:45:23	8754,0	transportation
13	DE_HH_WFS_INSPIRE_A1_7_Verkehrsne.gml		:is http://deegree.STRASSENNETZ INSPIRE BESCHR strasse strasse geom Inlined geom	2018-07-23 14:45:23	13845,0	transportation
14	DE_HH_WFS_INSPIRE_A1_7_Verkehrsne.gml		:e abschnittslaenge laengenherkunft laengenherkunft europastrasse europastrasse	2018-07-23 14:45:24	6617,0	transportation
15	DE_HH_WFS_INSPIRE_A1_7_Verkehrsne.gml		:is http://deegree.STRASSENNETZ INSPIRE BESCHR strasse strasse geom Inlined geom	2018-07-23 14:45:24	4102,0	transportation
16	DE_HH_WFS_INSPIRE_A1_7_Verkehrsne.gml		:http://deegree.STRASSENNETZ INSPIRE NULLPUNKTE zusatz zusatz geom Inlined ge	2018-07-23 14:45:24	3822,0	transportation
17	DE_HH_WFS_INSPIRE_A3_6_Versorgung.gml		:OM srsName EPSG pointMember Point GEOMETRY cccfc eadef srsName EPSG Poi	2018-07-23 15:22:36	5739,0	utilitiesCommunication
18	HH_WFS_abfall_recycling-374EB872-F58.gml		:er Weissglas Depotcontainer Weissglas Depotcontainer Wertstoff Depotcontainer	2018-07-23 15:10:45	7032,0	utilitiesCommunication

Abbildung 18: Daten der .gml-Dateien mit TopicCategoryCode

Mit Blick auf die Analyse der INSPIRE TopicCategoryCode-Zusammenhänge wurden die Inhalte aus den .gml-Dateien ausgelesen, gefiltert und mit den TopicCategoryCode-Werten aus den .xml-Dateien verbunden. Somit liegen folgende Daten einheitlich vor:

1. Dateien verbunden. Somit liegen folgende Daten einheitlich vor:

- **Datei:** Dateiname der .gml-Datei
- **Dateityp:** In diesem Fall liegt der Focus auf den .gml-Dateien
- **Text:** Der Inhalt der .gml-Datei wurde eingelesen und wie folgt aufbereitet:
Tokenization: Wörter extrahiert. Sonderzeichen und Zahlen verworfen.
Filter Tokens by Length: Nur Wörter/Tokens mit mind. 3 und max. 25 Zeichen verwendet.
- **Datum:** Datum der Datei.
- **Größe:** Größe der .gml-Datei in Byte (Zeichen).
- **MD_TopicCategoryCode:** Code in Englisch zu welcher TopicCategory eine .gml-Datei zugewiesen wurde.

5.3.3 Daten Analysieren

Die Aufbereitung und Zusammenführung der Inhalte zeigt bereits den Umfang und die Qualität der vorliegenden Daten in Bezug auf die TopicCategoryCode-Werte:

- Nur 1.194 INSPIRE .xml-Dateien hatten einen TopicCategoryCode (von 2.054).
- Davon hatten 254 INSPIRE .xml-Dateien mehr als einen TopicCategoryCode.

Anzahl TopicCategoryCodes in INSPIRE .xml-Datei	Anzahl der .gml-Dateien	Summe der Codes
1	940	940
2	159	318
3	46	138
4	43	172
5	6	30
GESAMT	1.194	1.598

Tabelle 11: Dateien mit TopicCategoryCode

Um ggf. Zusammenhänge von TopicCategroyCode und Inhalten der .gml-Dateien zu erhalten, wurden die Vorkommnisse der einzelnen Begriffe gemessen. Je Token (Deutsch: Wort/Begriff) wurde geschaut:

- Anzahl der Dokumente, die das Wort enthalten
- Anzahl der Vorkommen des Wortes insgesamt
- Anzahl der Vorkommen in einer jeweiligen TopicCategory

Die Analyse der Vorkommen ergab jedoch keine markanten Alleinstellungsmerkmale von Begriffen. In Abbildung 19 ist ein Ausschnitt der Tabelle zu sehen, die ohne Erfolg auf entsprechende Cluster untersucht wurde.

word Category	in documents Number	total Number	utilitiesCommunication Number	environment Number	boundaries Number	structure Number
against	1160	1160	59	371	40	20
application	1160	1167	59	371	40	20
artifact	1160	1160	59	371	40	20
attribute	1162	1260	59	421	40	20
boundedby	438	1236	0	254	90	144
bxml	1160	1167	59	371	40	20
change	1160	2320	118	742	80	40
current	1160	1160	59	371	40	20
degree	894	6088	393	1595	185	98
encoding	1196	1196	59	383	40	32
envelope	438	1236	0	254	90	144

Abbildung 19: Aufkommen der Begriffe aus den .gml-Dateien samt TopicCategoryCode-Zuordnung

Die Darstellung der häufigsten Begriffe je TopicCategoryCode zeigt ebenfalls keine markanten Zusammenhänge (siehe Abbildung 20).

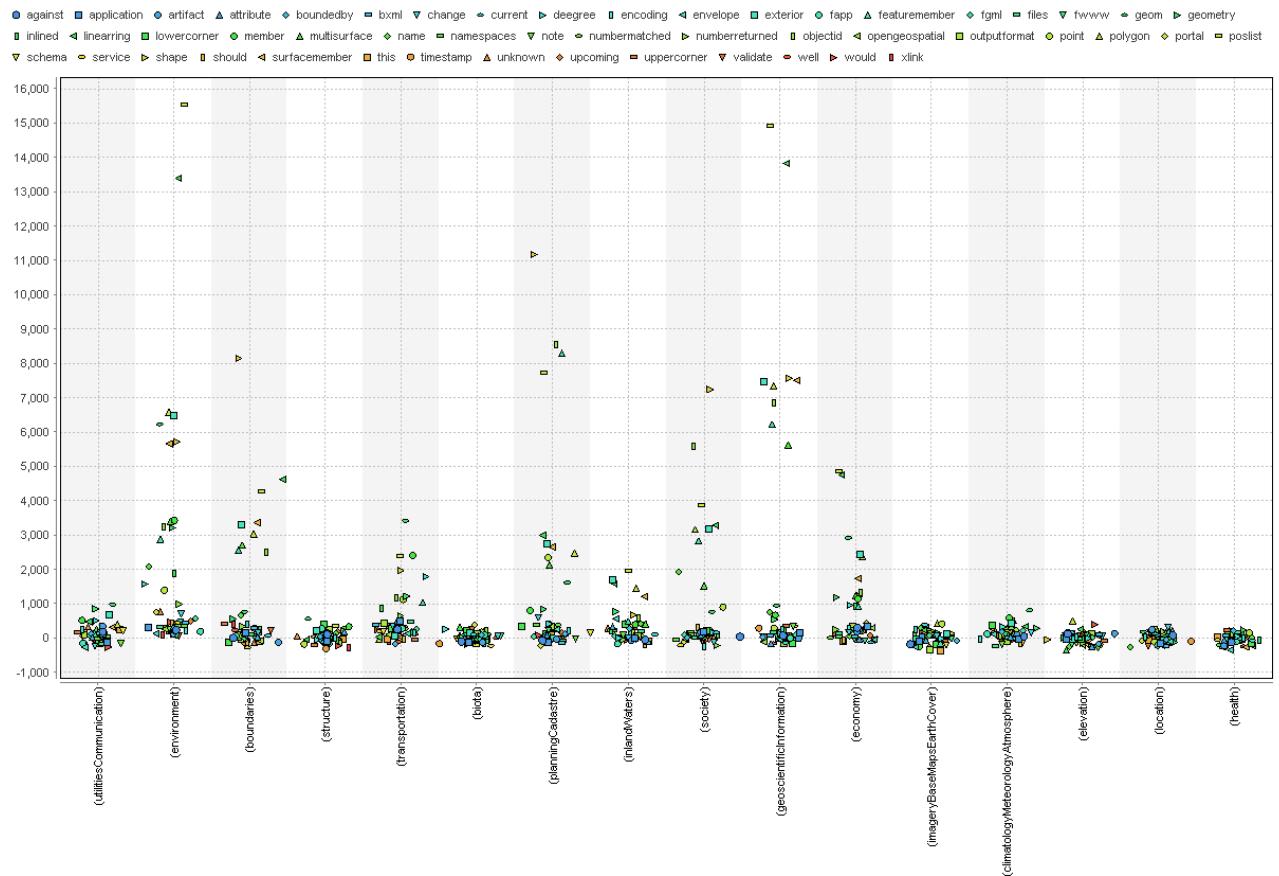


Abbildung 20: Keine Clusterung der Begriffe aus den .gml-Dateien erkennbar

Die folgende Darstellung (Abbildung 21) zeigt die TopicCategory und die Anzahl an .xml-Dateien, die in die Kategorie fallen.

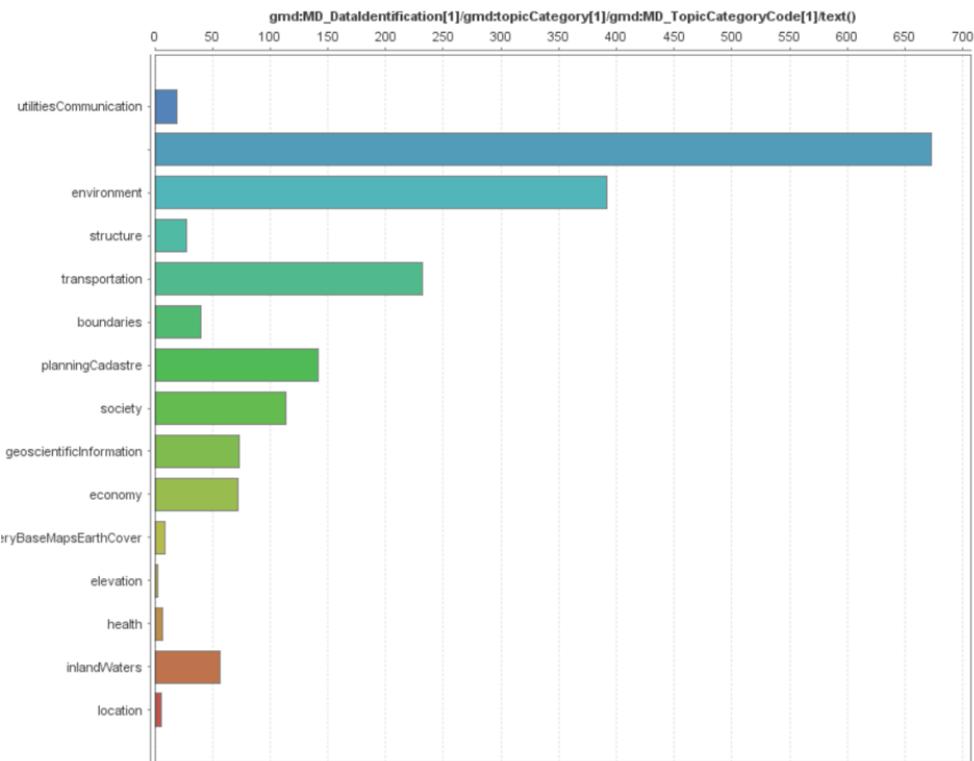


Abbildung 21: Verteilung der TopicCategoryCode- Werte

Insgesamt kommen 15 der verfügbaren 19 Kategorien (Quelle: <http://inspire.ec.europa.eu/metadata-code-list/TopicCategory:2>) vor. Da das Modell zum Trainieren eine gewisse Anzahl an Werten benötigt, wurde entschieden, Kategorien mit einem Anteil von unter 5% vom Gesamtkorpus zu eliminieren. Daraus wurden folgende Kategorien für die Erstellung eines Modells herangezogen:

- Environment
- Transportation
- PlanningCadastre
- Society
- Economy

5.3.4 Modell erstellen

5.3.4.1 Einsatz von NLP

Die Spracherkennung wurde mit Hilfe verschiedener Python-Bibliotheken getestet und umgesetzt.

Um die Sprache eines .gml-Dokuments zu analysieren, muss zunächst der Inhalt der jeweiligen Datei extrahiert werden. Dazu wird der komplette Text einer .gml-Datei eingelesen. Aus dem extrahierten Text werden alle numerischen Werte herausgefiltert, da Zahlen für eine Spracherkennung unerheblich sind.

Der Text wird dann in Tokens gespalten. Doppelte Tokens werden aussortiert, da die doppelte Be trachtung von Wörtern nicht relevant ist. Weiterhin werden Wörter aussortiert, die weniger als drei Zeichen haben, da sie meist keiner Sprache eindeutig zugeordnet werden können. Anschließend wird eine Spracherkennung über den nicht leeren Text mit mindestens drei Wörtern durchgeführt. Dabei hat sich nach einem Benchmark folgende Bibliothek durchgesetzt:

- **Python-Bibliothek:** TextBlob

TextBlob ist eine Python-Bibliothek zur Verarbeitung von Textdaten. Sie bietet eine einfache API für das Eintauchen in allgemeine NLP-Aufgaben, zum Beispiel für die Spracherkennung.

Die erkannte Sprache (z.B. „de“) wird noch über eine Mapping-Tabelle in die bestehende ISO-Norm überführt (z.B. „ger“) und ausgegeben.

5.3.4.2 Einsatz von DeepLearning

Um den Einsatz eines Deep-Learning-Modells zu evaluieren und mit klassischen Machine Learning-Ansätzen zu vergleichen, wurde im RapidMiner ein entsprechender Workflow erstellt. Abbildung 22 zeigt den Workflow im Überblick.

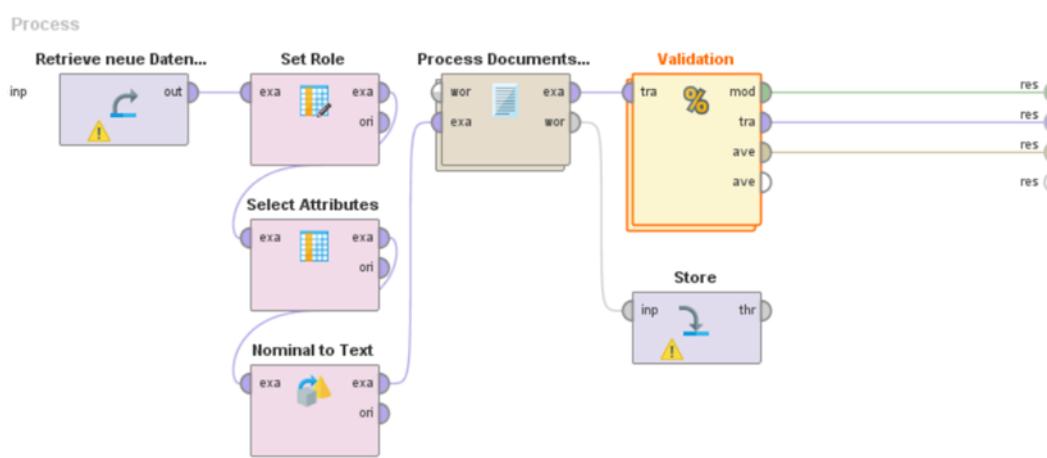
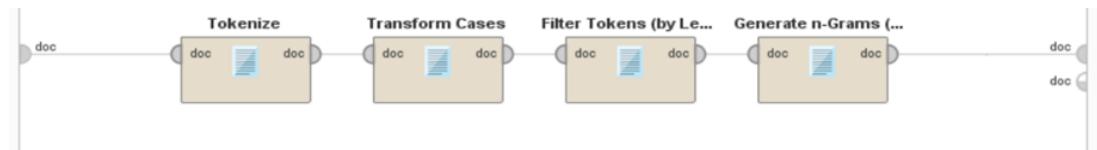


Abbildung 22: RapidMiner Workflow zum Vergleich von ML und Deep Learning

Die Komponenten des Workflows haben folgende Funktionalitäten:

- a. **Retrieve:** Die Datenabfrage kann entweder via .csv-Import oder über das Importieren der Daten in das Repository des RapidMiners erfolgen. In dem Fall wurden die Daten zuerst in das Repository geladen und dann abgefragt.
- b. **Set Role:** Als nächstes muss die Spalte Category als Label definiert werden. Das Label ist das Merkmal, welches zukünftig vorhegesagt werden soll.
- c. **Select Attributes:** Falls unnötige Spalten in dem Datensatz vorliegen, können diese durch „Select Attributes“ ausgeschlossen werden.
- d. **Nominal to Text:** Durch „Nominal to Text“ werden die Daten in Texte umgewandelt, damit diese für den nächsten Schritt bearbeitet werden können.
- e. **Process Document from Data:** Dieser Operator ist der entscheidende Schritt, um die Texte für die Klassifizierung vorzubereiten.



- I. Tokenize spaltet jedes Wort in Tokens.
- II. Transform Cases wandelt alle Buchstaben in Kleinbuchstaben um.
- III. Filter Tokens by Length: Wir haben uns dazu entschieden alle Wörter, die aus drei oder mehr als 25 Buchstaben bestehen, auszusortieren.
- IV. Durch n-Grams werden nicht nur einzelne Wörter betrachtet, sondern auch zwei miteinander in Beziehung stehenden Wörter. Das Resultat sind die primären Daten, aus denen das Modell lernt.
- f. **Store:** Mit „Store“ wird die generierte Wörterliste abgespeichert, die später beim Anwenden von neuen Daten genutzt wird.
- g. **Validation:** Diese Komponente ist dazu da, um das Modell zu trainieren und zu testen. Dabei wird eine Ratio von 70/30 angewandt. Das heißt, dass 70% der Daten zum Antrainieren (508 Datensätze) und 30% (218 Datensätze) zum Testen des Modells herangezogen werden.

Der Inhalt der Validation-Komponente ist in Abbildung 23 dargestellt. Im Trainings-Set wird mit Hilfe des Deep-Learning-Operators ein Vorhersagemodell erzeugt, welches für zukünftige Daten die entsprechende TopicCategory vorhersagt. Der Deep Learning besteht aus einer Klasse von Optimierungsmethoden künstlicher neuronaler Netze, die zwei hidden layers mit jeweils 50 Neuronen zwischen Input und Output haben.

Um die Ergebnisse des KNN besser einzuschätzen, wurden die Daten auch mit anderen ML-Modellen getestet. Die Ergebnisse sind in Kapitel 6.2 erläutert.

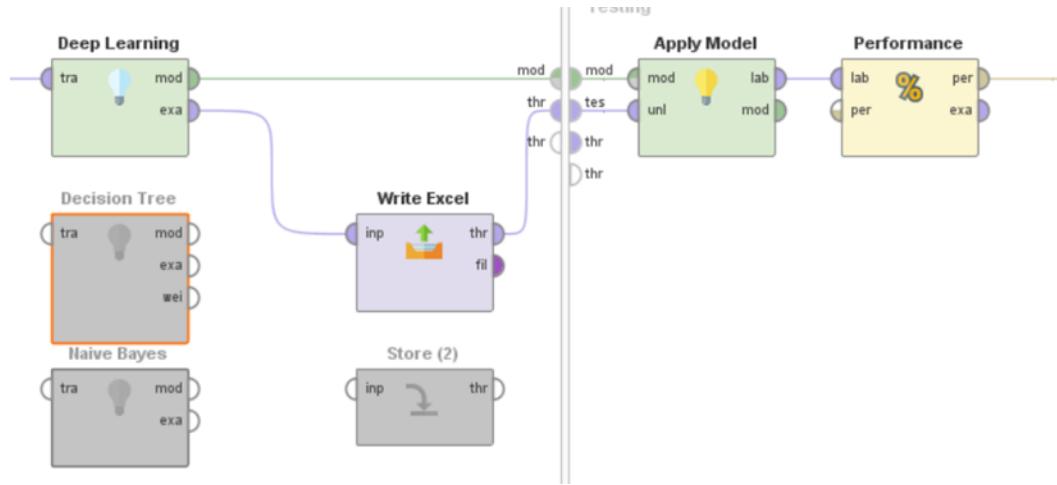


Abbildung 23: RapidMiner Workflow – Validation-Komponente

5.4 ActiniaGDI-CSW-Writer

Die Komponente „ActiniaGDI-CSW-Writer“ schreibt schlussendlich die durch die verschiedenen Methoden erfassten Metadaten in das entsprechende Metadatenprofil. Der CSW-Writer nutzt dazu die CSW-T-Schnittstelle von GNOS und schreibt die Metadaten transaktional in den Zielkatalog-Server.

Um die Metadaten strukturiert schreiben zu können, benötigt der Writer eine Profil-Vorlage, in einer späteren Ausbaustufe ist es möglich, dass hier auch unterschiedliche Metadaten-Profile (INSPIRE, ISO, DCAT, u.a.) aus einem Prozess heraus angesteuert und entsprechende Metadaten geschrieben werden.

6 AP V: Evaluierung der Ergebnisse und Herausstellung des Verwertungspotentials

6.1 Prototyp für die automatisierte Metadatenerzeugung

Der Prototyp des Projektes MetaOpenData ist unter <https://bmvimetadaten.mundialis.de/> abruf- und nutzbar.



Abbildung 24: Startseite des Prototypen MetaOpenData

Die Anwendung besteht aus einer Weboberfläche, über die eigene Geodaten hochgeladen werden können. Diese Oberfläche bietet die Möglichkeit, entweder eine lokale eigene Datei auszuwählen oder innerhalb einer Landkarte ein Polygon nach Wahl zu zeichnen.

Lädt man eine lokale Datei hoch (z.B. eine gezippte Shapedatei) wird, nach Betätigung der Schaltfläche „Submit“, der ausgewählte Datensatz mit Hilfe der Actinia-GDI-Komponente auf einem Server in einem eingerichteten Dateisystem abgelegt. Der Dateiname der ausgewählten Datei wird als Titel in ein Metadatentemplate (XML-Datei) übernommen.

Das Hochladen der Daten ist an dieser Stelle allerdings als Show-Case zu verstehen, eine professionell eingesetzte Lösung müsste an dieser Stelle lediglich den Zugriff auf die Daten haben.

Anschließend sendet die Actinia-GDI-Komponente das ersetzte Metadatentemplate an die CSW-T Schnittstelle von Geonetwork Open Source, wodurch ein Metadatensatz für den hochgeladenen Datensatz erzeugt wird. Auf der Weboberfläche des Prototyps spielt sich dieser Vorgang innerhalb von wenigen Sekunden ab.

Der Nutzer erhält nach dem Einfügen seines Datensatzes (mit „Submit“) eine Webadresse (engl. Uniform Resource Locator, URL), welche sich aus der Webadresse zum Geonetwork sowie der sogenannten „Universally Unique Identifier“ (UUID), also der eindeutigen ID des neuen Metadatensatzes, zusammensetzt. Ruft man diese URL in einem Browser auf, wird man auf den entsprechend angelegten Metadatensatz im Geonetwork weiterleitet. Dieser Metadatensatz wird im Hintergrund nun weiter mit Informationen angereichert. Das geschieht mit Hilfe des Data-Interpreters, der Informationen aus dem hochgeladenen Datensatz automatisch detektiert. An dieser Stelle ergibt sich außerdem die Möglichkeit, weitere Informationen, die aus der Deep-Learning-Komponente entstehen, ebenfalls in die XML-Datei zu schreiben und damit erneut die Komponente Geonet-

work anzufragen, wodurch der vorhandene Metadatensatz aktualisiert wird und immer weiter mit Informationen gefüllt werden kann

Der erstellte Prototyp und das bereits aufgekommene Interesse von Datenbereitstellern (z.B. Bundesstadt Bonn oder GovData) zeigen, dass die erarbeiteten Konzepte des Prototypen die Basis für weitere Entwicklungen bilden kann.

Grundsätzlich sehen wir alleine im verfolgten Ansatz der automatisierten, wenn auch unvollständigen, Erstellung von Metadatensätzen sowie insbesondere der Möglichkeit automatisiert Feature-Kataloge zu erstellen, ein hohes Potential, auch bestehende Metadatensätze zu verbessern und die Nutzung der Daten dahinter zu fördern.

Dabei muss das System nicht zwingend so eingesetzt werden, dass die Metadaten in einem eigenen System erfasst werden (eigener Metadaten-Knoten). Sofern das zu entwickelnde System die standardisierte OGC-CSW Schnittstelle unterstützt, können so auch externe Kataloge, wie beispielsweise Datenportale der Länder oder des Bundes angesteuert werden. Für dieses Szenario wäre lediglich die Authentifizierung und Autorisierung an den externen Portalen zu klären. Zudem müsste die Möglichkeit bestehen, externe Metadatenprofile automatisiert auszulesen (technisch kein Problem) und die zu erfassenden Metadaten-Tags, die das System füllen soll, entsprechend automatisiert zu konfigurieren.

Eine weitere denkbare Entwicklung wären Plugins für verschiedene Metadatenmanagement-Systeme bzw. Datenportale wie beispielsweise Comprehensive Knowledge Archive Network (CKAN) oder GeoNetwork Open Source. Diese Plugins können ebenso die Basis für eine weitere Verlinkung zu den nachfolgend beschriebenen KI-Ansätzen sein.

6.2 Potenziale zum Einsatz KI-basierter Verfahren bei der Metadatenerzeugung

Im Folgenden wird dargestellt, inwiefern sich KI-Verfahren eignen, die Metadatenerzeugung zu automatisieren. Im Rahmen der Machbarkeitsstudie wurden vorhandene Daten aufbereitet und zum Trainieren verschiedener Vorhersagemodelle genutzt (siehe Kapitel 5.3.4.2). Ein Teil der Daten (30%) wurde zur Validierung herangezogen, um die Güte der Modelle zu messen.

Um die nachfolgenden Auswertungen nachvollziehen zu können, werden einleitend die wesentlichen Begrifflichkeiten wie Trefferquote, Genauigkeit und Korrektheit kurz erläutert:

- Die Trefferquote (engl. recall) eines realen Wertes ist das Maß, wie oft dieser Wert korrekt vorhergesagt wurde. Also wie gut ein realer Wert vorhergesagt werden kann.
- Die Genauigkeit (engl. precision) eines vorhergesagten Wertes ist das Maß, wie oft dieser Wert mit dem tatsächlichen Wert übereinstimmt (positiver Vorhersagewert). Also wie wahrscheinlich ist es, dass die Vorhersage dieses Wertes auch der Realität entspricht.
- Die Korrektheit des Modells (engl. Accuracy) gibt an wie oft die Vorhersagen insgesamt richtig lagen. Also wie wahrscheinlich ist es, dass der vorhergesagte Wert – egal welcher – der Realität entspricht. Wenn z.B. von 218 Vorhersagen 164 korrekt waren, ergibt das eine Genauigkeit von 75,23%. Dieser Prozentwert ist kein Durchschnittswert der anderen Prozentwerte (recall, precision).

Die Korrektheit gibt dabei auch an, mit welcher Wahrscheinlichkeit im Produktivbetrieb die korrekten TopicCategoryCode- Werte vorhergesagt werden würden.

Die Ergebnisse der einzelnen Algorithmen werden im Folgenden beschrieben.

6.2.1 Decision Tree

Entscheidungsbäume sind eine Methode zur automatischen Klassifikation von Datenobjekten und damit zur Lösung von Entscheidungsproblemen. Die grafische Darstellung als Baumdiagramm veranschaulicht hierarchisch aufeinanderfolgende Entscheidungen.

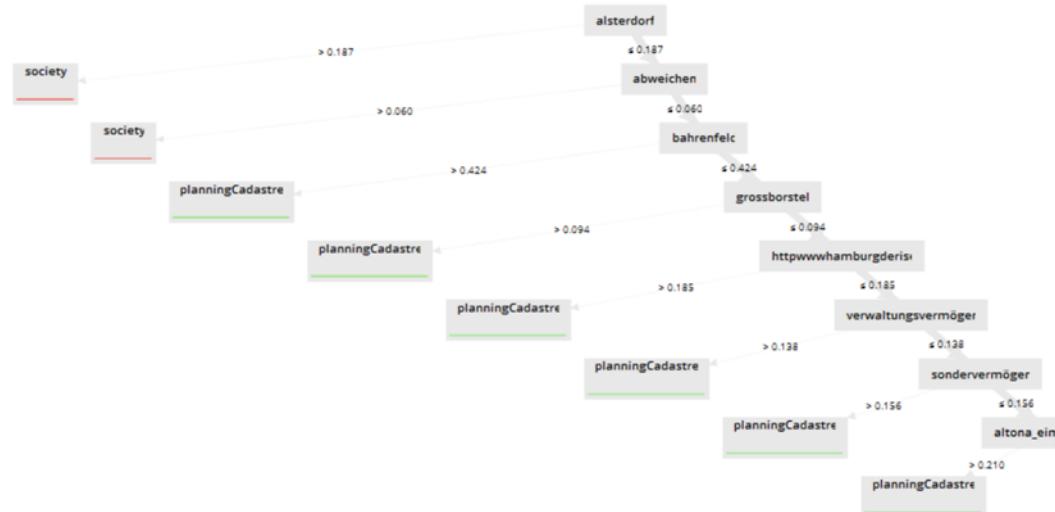


Abbildung 25: Darstellung des berechneten Decision Tree

Mit einer Genauigkeit von 61,93% ist der Decision Tree auf fünf Kategorien relativ schlecht. In den einzelnen Spalten ist deutlich zu erkennen, nur zwei Kategorien werden korrekt betrachtet. Das gute Abschneiden in diesen Kategorien hängt damit zusammen, dass der Algorithmus die kleineren Kategorien, die weniger Unterscheidungsmerkmale haben, kaum berücksichtigt und nur Kategorien vorhersagt, die eine hohe Anzahl von Dateien haben.

accuracy: 61.93%							
	true environment	true transportation	true planning	Cada...	true economy	true society	class precision
pred. environment	104	56	8	16	0		56.52%
pred. transportation	0	0	0	0	0		0.00%
pred. planning	0	0	15	0	0		100.00%
pred. economy	1	0	0	6	0		85.71%
pred. society	0	1	1	0	10		83.33%
class recall	99.05%	0.00%	62.50%	27.27%	100.00%		

Abbildung 26: Testergebnisse Decision Tree

6.2.2 Naive Bayes

Der Naive Bayes-Klassifikator gehört zur Familie einfacher "wahrscheinlicher Klassifikatoren", die auf der Anwendung des Satzes von Bayes mit starken (naiven) Unabhängigkeitsannahmen zwischen den Merkmalen basieren.

Die Genauigkeit des Naive Bayes Algorithmus ist mit 75,23% neben dem Deep Learning am besten. Die einzelnen Werte innerhalb der Klassengenauigkeit und –präzision sind allerdings nicht so ausgewogen, wie bei dem Deep Learning Algorithmus, was heißt, dass bei umfangreicheren sowie komplexeren Daten die Korrektheit vermutlich schlechter wird. Solange die Komplexität jedoch auf fünf Kategorien begrenzt bleibt, ist der Naive Bayes eine Alternative zum Deep Learning.

accuracy: 75.23%						
	true environment	true transportation	true planningCada...	true economy	true society	class precision
pred. environment	65	0	2	1	0	95.59%
pred. transportation	17	49	1	1	0	72.06%
pred. planningCada...	5	1	20	0	0	76.92%
pred. economy	18	7	0	20	0	44.44%
pred. society	0	0	1	0	10	90.91%
class recall	61.90%	85.96%	83.33%	90.91%	100.00%	

Abbildung 27: Testergebnisse Naive Bayes

6.2.3 Random Forest

Der Random Forest ist ein Klassifizierungsverfahren bei der zur Entscheidung eine Vielzahl von Entscheidungsbäumen erstellt wird und die Klasse ausgegeben wird, die mit den meisten Stimmen der einzelnen Entscheidungsbäume übereinstimmt. Das Ergebnis ist mit unter 50% sehr schlecht und nicht für den Anwendungsfall nutzbar.

accuracy: 49.54%						
	true environment	true transportation	true planningCada...	true economy	true society	class precision
pred. environment	105	56	24	22	8	48.84%
pred. transportation	0	1	0	0	0	100.00%
pred. planningCada...	0	0	0	0	0	0.00%
pred. economy	0	0	0	0	0	0.00%
pred. society	0	0	0	0	2	100.00%
class recall	100.00%	1.75%	0.00%	0.00%	20.00%	

Abbildung 28: Testergebnisse Random Forest

6.2.4 Deep Learning

Der Deep Learning Algorithmus ist mit zwei Hidden Layer und jeweils 50 Neuronen der aussagekräftigste Algorithmus für das Zuordnen von .gml-Dateien zu den Kategorien (TopicCategoryCode). Bis auf die Kategorie „economy“, die fälschlicherweise zu oft der Kategorie „environment“ zugeordnet wurde, ist der Algorithmus am vielversprechendsten, weil man diesen ebenfalls noch per

Fine-Tuning verbessern kann (z.B. andere Aktivierungsfunktion, weitere Hidden Layer, mehr/weniger Neuronen pro Layer).

Die inkorrekt Zuordnungen entstehen hauptsächlich bei den .gml-Dateien, die aus Geodaten bestehen. Da die Geodaten nur aus Ziffern bestehen, werden diese aus der Analyse entfernt, sodass ein fast leerer Textinhalt zurückbleibt. Dieser Textinhalt wird von dem Modell trotzdem angelernt, wenn die entsprechende .xml-Datei eine TopicCategory enthält. Bei der Zuordnung wird also eine Datei ohne interpretierbaren Geodaten einer bestimmten Kategorie zugeordnet, obwohl aufgrund der Geodaten ggf. eine andere Kategorie in Betracht käme. Das Modell klassifiziert also die nicht interpretierbaren Geodaten innerhalb der .gml-Dateien in die am häufigsten vorkommende Kategorie „environment“, weil anhand der Testdateien die Kategorie „environment“ die Kategorie mit den diversifizierten Inhalt war.

	true environment	true transportation	true planningCada...	true economy	true society	class precision
pred. environment	87	19	5	15	0	69.05%
pred. transportation	14	37	0	0	0	72.55%
pred. planningCada...	1	0	17	0	0	94.44%
pred. economy	3	1	2	7	0	53.85%
pred. society	0	0	0	0	10	100.00%
class recall	82.86%	64.91%	70.83%	31.82%	100.00%	

Abbildung 29: Testergebnisse Deep Learning

6.2.5 Zusammenfassung der Projektergebnisse im Bereich KI-Verfahren

Die besten Ergebnisse werden mit dem Deep-Learning-Algorithmus sowie dem Naive-Bayes-Ansatz erreicht (siehe Abbildung 30). Hierbei hat jedoch der Deep-Learning-Ansatz noch mehr Potential, welches durch einen größeren Umfang an Daten und weiteres Training weiter ausgeschöpft werden könnte.

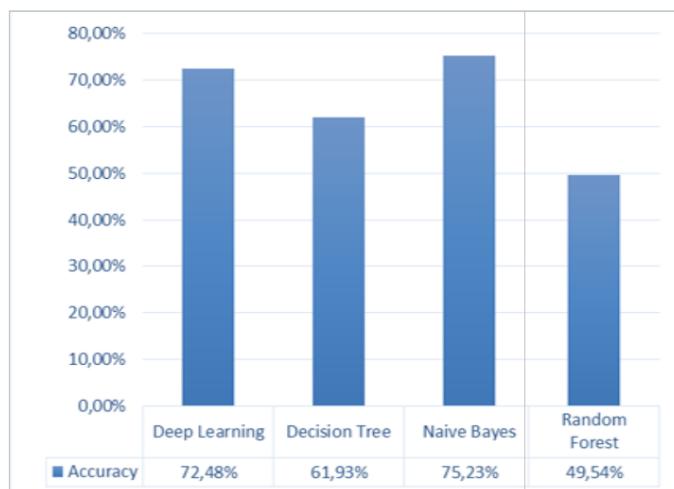


Abbildung 30: Testergebnisse aller KI-Verfahren

Die Schwierigkeit besteht darin, die Inhalte der .gml-Dateien in der notwendigen Form aufzubereiten und in einer geeigneten übergreifenden Struktur zusammenzufassen. In dieser ersten Evaluierung wurden noch Sonderzeichen und Zahlen und damit auch Geodaten ignoriert. Dadurch gab es eine Reihe an „vermeintlich“ fast leeren Dateien, die bestimmten Kategorien wie „economy“ zugeordnet waren. Aus Sicht des Modells ist diese Zuordnung jedoch nicht trainierbar und damit auch nicht vorhersagbar.

Wenn die betroffenen Dokumente aussortiert werden, wird die Korrektheit der Modelle bei den Testläufen sehr wahrscheinlich auf über 80% bzw. sogar 90% steigen. Allerdings spiegelt diese Situation noch weniger die Realität wider, in der schätzungsweise ca. 15% der Dateien hauptsächlich aus Geodaten bestehen. Für die Praxis müssten diese Fälle sowie die übrigen der 19 Kategorien ebenfalls zum Training des Modells verwendet werden, andernfalls ist eine automatische Zuordnung in der Praxis nicht einsetzbar.

Grundsätzlich ist anzumerken: Je mehr Daten zur Verfügung stehen, desto genauer können die Modelle trainiert werden. Hinsichtlich der Untersuchung des KI-Einsatzes zur automatisierten Metadatenerzeugung im Kontext MetaOpenData wird folgendes Fazit gezogen:

- Zur Erstellung eines Modells werden weit mehr Daten in besserer Qualität benötigt.
Für alle 19 Kategorien (TopicCategoryCode) sollten jeweils mindestens 500 bis 1.000 Datensätze vorliegen. In Kapitel 5.3 sind die Schritte zur Datenaufbereitung sowie Datenanalyse beschrieben, um dies zu überprüfen. Dann könnten darüber hinaus auch andere KNN-Arten wie die rekursiven neuronale Netze (RNN) betrachtet werden.
- Eine Zuordnung zu Kategorien sollte auch Geodaten berücksichtigen.
Da einige .gml-Dateien auch in Zukunft ggf. (fast) nur Geodaten enthalten, sollte ein zweites Modell erstellt werden, welches mit Geodaten umgehen kann. Analog zur ortsabhängigen Vorhersagemodellierung (engl. geospatial predictive modeling, Quelle: https://en.wikipedia.org/wiki/Geospatial_predictive_modeling) könnte ein Modell zur Kategorisierung von Geodaten entwickelt oder ein bestehender Ansatz wiederverwendet werden.
- Eine Zuordnung zu mehreren Kategorien muss zusätzlich berücksichtigt werden.
Die vorliegenden Modelle sagen je Testdatensatz immer nur eine Kategorie vorher. Um für alle Kategorien jeweils einen Vorhersagewert zu erhalten muss u.a. die Aktivierungsfunktion angepasst werden. Die Evaluierung dieses Ansatzes macht jedoch erst Sinn, wenn Punkt 1 und 2 erfüllt sind.

Die Ergebnisse der Machbarkeitsstudie zeigen, dass es grundsätzlich möglich ist, KI-gestützt Metadatenattribute aus den eigentlichen Datensätzen abzuleiten. Im Rahmen der Machbarkeitsstudie stand jedoch keine Datenmenge zur Verfügung, um ein produktionsnahes Modell zur Metadatenerzeugung zu entwickeln.

6.2.6 Ausblick KI-Verfahren in der Metadatenerzeugung

Die Metadaten der Geodatensätze sind ein wichtiger Bestandteil der Indexierung. Abhängig vom Suchverhalten der Nutzer, könnten weitere wichtige Attribute identifiziert werden, die dabei helfen

in der Fülle der zur Verfügung stehenden Geoinformationen die brauchbaren Datensätze zu finden.

Hierbei birgt das Verständnis der Geokoordinaten samt Informationen aus den .gml-Dateien großes Potential. Zum einen kann ein gewisses Verständnis über die Struktur der Ortsvektoren untereinander aufgebaut werden. Zum anderen können die Zusammenhänge zu den bekannten Metainformationen als auch Informationen aus weiteren Quellen wie z.B. Google Maps verwendet werden, um ein Modell zu erstellen. Mit einem umfangreichen Wissen wären eventuell weitere Aussagen möglich, als nur die Zuordnung der Kategorien (topicCategoryCode).

Bei den künftigen Einsatzszenarien von KI-Ansätzen sollten folgende Eigenschaften erfüllt sein:

- Die Herleitung bzw. Vorhersage von Daten kann nicht durch „einfache“ Regeln/Workflows wie z.B. Geodatenbereich (min./max.) abgebildet und automatisiert werden.
- Die Auswahl oder Vorhersage erfordert keinen hohen kreativen Prozess (wie z.B. Beschreibung), sondern kann aus vorhandenen Daten abgeleitet werden.
- Die zur Vorhersage notwendigen Daten liegen in allen möglichen eintretenden Konstellationen ggf. mehrfach vor. D.h. die Daten sind in ausreichender Menge und Qualität vorhanden, um ein Modell zu trainieren und zu testen.

Sofern diese Bedingungen erfüllt sind, birgt der Einsatz entsprechender KI-Ansätze sehr hohes Potential.

6.2.7 Fazit

Generell wurde durch die Studie deutlich, dass es möglich ist, den Prozess der Metadatenerfassung zu automatisieren und damit deutlich zu vereinfachen. Dabei trägt zum Einen die Möglichkeit bei, Metadatensätze durch reine Betrachtung der Daten anzulegen. Hier fehlt allerdings noch eine abschließende Betrachtung von Geodaten-Diensten, wobei wir zuversichtlich sind, dass sich die angewandte Technologie auch auf Dienste ausweiten lässt.

Das größere Potential der automatisierten Erfassung birgt im betrachteten Ansatz mit Sicherheit aber der KI-Ansatz.

Aus unserer Sicht ist es erstrebenswert auf Basis unserer Erkenntnisse die Entwicklung einer laufähigen und anpassbaren Software-Lösung anzustreben. Aus unserer Sicht muss diese Lösung Open Source sein, um das große Potential dieser angestrebte Lösung so vielen Nutzern wie möglich zugänglich zu machen.

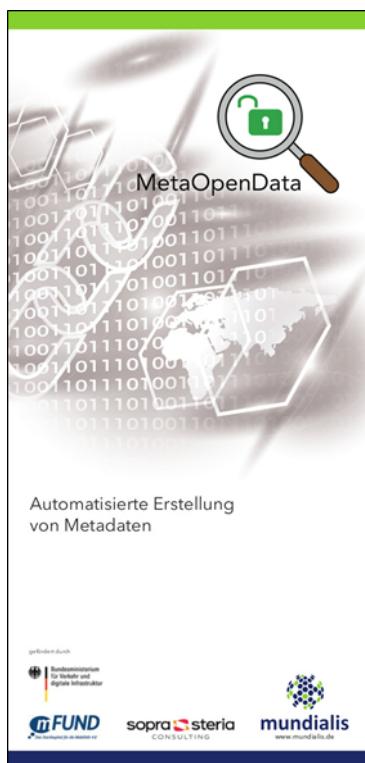
6.3 Ergebnispräsentationen und Folgeaktivitäten

6.3.1 Vorträge

Im Rahmen der Machbarkeitsstudie wurden auf verschiedenen Veranstaltungen Vorträge rund um das Projekt eingebrochen. Diese stießen jeweils auf großes Interesse. Im Einzelnen wurden Projektergebnisse bei den folgenden Veranstaltungen vorgestellt:

- FOSSGIS 2018, Bonn: MetaOpenData – Manuelle Metadatenerzeugung war gestern (Sebastian Goerke), Aufzeichnung: <https://media.ccc.de/v/2018-5276-openmetadata>
- Bratwurst, Bier & GIS 2018, Bonn: Open Data sind toll! Doch wie finde ich, was ich brauche? (Sebastian Goerke)
- mFUND Konferenz 2018, Berlin: (Till Adams)
- GeoIT Round Table NRW 05/2019, Münster : (Sebastian Goerke)

6.4 Projektflyer



Für Werbezwecke und zur Projektvorstellung wurde der hier abgebildete Flyer erstellt. Dieser stellt in kurzen Worten das Projekt MetaOpenData und dessen Ziele vor. Hierfür wurden unter anderem auch die Ergebnisse aus der bereits oben beschriebenen Umfrage ausgewertet und eingearbeitet. Auch eine Skizze der vollständigen Projektarchitektur wird gezielt beschrieben, um auch Interessenten über die technische Umsetzung zu informieren. Die Partner (Sopra Steria Consulting und mundialis), Kontaktdaten, ein Hinweis auf den mFUND sowie die Projekt-Homepage befinden sich auf der letzten Seite des Flyers.

Bonn, den 28.06.2019

gez. Till Adams

- Geschäftsführer mundialis GmbH & Co. KG -