Bundesministerium
für Verkehr und
digitale Infrastruktur

Forschungsprojekt

Machbarkeitsstudie zum Aufbau einer KI-basierten Wissensplattform im Umweltbereich

Schlussbericht: 31.03.2020

BMVI-Auftragsforschung: FE-Nr. 97.404/2019

Erstellt durch:

Condat AG

Es wird darauf hingewiesen, dass die unter dem Namen des Verfassers bzw. der Verfasser veröffentlichten Berichte nicht in jedem Fall die Ansicht des Herausgebers wiedergeben. Die Verantwortung für den Inhalt liegt daher ausschließlich beim Autor.

Die Verwendung dieses Forschungsberichts oder Auszüge hieraus durch Dritte sind nur mit Quellenangabe zulässig.

BMVI Machbarkeitsstudie 2019

Aufbau einer Wissensplattform im Umweltbereich zur Planungsbeschleunigung von Infrastrukturmaß-nahmen

Abschlussbericht
Teil A: Systembeschreibung

Jan Thomsen, Sacha Prelle, Thomas Langer, Ulrike Freitag (Condat AG), Jan Nehring, Philippe Thomas, Sven Schmeier (DFKI)

Stand: 31.03.2020

Auftraggeber
Bundesministerium für Verkehr und digitale Infrastruktur (BMVI)
Referat G10
Invalidenstr. 44
10115 Berlin

condat®

Inhaltsverzeichnis

1. Einleitung 7
2. Anforderungen und Ziele 8
2.1 Anforderungen aus der Leistungsbeschreibung 8
2.2 Anforderungen aus dem 1. Workshop 9
2.2.1 Tierarten 9
2.2.2 Quellen 10
2.2.3 Dokumentarten 11
2.2.4 Suchbegriffe 11
2.2.5 Sprachen 12
2.2.6 Relevanzkriterien 12
2.2.7 Gültigkeitsdauer von Informationen 12
2.2.8 Suchfunktionen 12
2.2.9 Vokabular 13
2.3 Anforderungen aus dem 2. Workshop 14
2.3.1 Zentrales Suchfeld 14
2.3.2 Suchergebnisse 14
2.3.3 Explorative Suche/Clustering 15
2.3.4 Anzeige von Metadaten 15
2.3.5 Sortierung der Suchergebnisse 15
2.3.6 Sanduhrprinzip 15
2.3.7 Navigation innerhalb von Dokumenten 16
2.3.8 Vorhabenbezogene Sicht 16
3. Überblick über das Gesamtsystem 17

3.1 Datenspeicherung 18
3.2 Integrations- und Kommunikationsschicht (Message Broker) 19
3.3 Plattformdienste (Microservices) 19
3.4 Schnittstelle zu Daten und Diensten 19
3.5 Applikationsserver / Benutzeroberfläche 19
4. Der Vorverarbeitungs-Workflow 21
4.1 Finden relevanter Dokumente 22
4.1.1 Import bekannter Quellen 22
4.1.2 Regelmäßige automatische Internet-Suchanfragen 22
4.1.3 Manuelle Suchanfragen 23
4.1.4 Semantische Suchraumerweiterung 23
4.1.5 Relevanzbewertung 24
4.1.6 Metadatenanreicherung 24
4.2 Import in die Wissensplattform 25
4.3 Download und Textextraktion 26
4.4 Segmentierung 26
4.5 Textanalyse 26
4.6 Named Entity Recognition 27
4.7 Semantische Indexierung 27
5. KI-basierte Dienste für die Benutzeroberfläche 29
5.1 Explorative Suche durch *Clustering* 29
5.2 Semantische Suche durch *Semantic Fingerprinting* 29
6. Benutzeroberfläche 31
7. Komponenten und Dienste 37

7.1 Importer (Crawling / Scraping) 37
7.1.1 Prozessbeschreibung 37
7.1.2 Analyse des Quell-Portals 37
7.1.3 Implementierung eines quell-spezifischen Scrapers 38
7.2 Internet-Suchkomponente 39
7.3 Relevanzfilter 40
7.4 Regionalerkennung 41
7.4.1 Technisches Verfahren 42
7.4.2 Realisierung 42
7.5 Dokumenttyperkennung 43
7.5.1 Erläuterung der Funktion 43
7.5.2 Technisches Verfahren 44
7.5.3 Realisierung 45
7.6 Explorative Suche/Clustering 48
7.6.1 Erläuterung der Funktion des Clusterings 48
7.6.2 Technisches Verfahren 48
7.6.3 Machbarkeit 49
7.6.4 Ausblick 49
7.7 Textextraktion 50
7.8 Segmentierung 52
7.8.1 Textextraktion mittels Apache Tika 52
7.8.2 Speicherung und Indizierung von Dokumenten in Elasticsearch 53
7.9 Semantische Indexierung 54
7.9.1 Grundlage: die BMVI-Ontologie 54
7.9.2 Textanalyse 55

7.9.3 Semantische Analyse 56

7.10 Semantische Suche 58

7.10.1 Semantische Suchraumerweiterung 59

7.10.2 Ähnlichkeitssuche 60

8. Zusammenfassung und Fazit 62

Anhang 65

A. Technische Komponenten des Systems 65

B. Quellensammlung 66

C. Vokabularsammlung 71

D. Referenzen 84

Abbildungsverzeichnis

Abbildung 1: Gesamtüberblick Systemarchitektur 18

Abbildung 2: Workflow der Wissensplattform Umwelt 21

Abbildung 3: Benutzeroberfläche 31

Abbildung 4: Explorative Sicht der Suchergebnisse 33

Abbildung 5: Sortierung und Filterung der Suchergebnisse 33

Abbildung 6: Konfigurationsparameter 34

Abbildung 7: Detailansicht 35

Abbildung 8: Konfusionsmatrix der Bundeslanderkennung 43

Abbildung 9: Konfusionsmatrix zur Dokumenttyp-Erkennung 45

Abbildung 10: Ausschnitt aus der BMVI-Ontologie 55

Abkürzungsverzeichnis

API	Application Programming Interface
BAV	Bundesanstalt für Verwaltungsdienstleistungen
BMVI	Bundesministerium für Verkehr und digitale Infrastruktur
CEF	Continuous ecological functionality
DFKI	Deutsches Forschungsinstitut für Künstliche Intelligenz
EBA	Eisenbahn-Bundesamt
FFH	Flora-Fauna-Habitat
GDWS	Generaldirektion Wasserstraßen und Schifffahrt
HTML	Hypertext Markup Language
JSON	JavaScript Object Notation
OWL	Web Ontology Language
RDF	Resource Description Framework
REST	Representational State Transfer
SME	Smart Media Engine
URL	Uniform Resource Locator
WNA	Wasserstraßen-Neubauamt

1. Einleitung

Das vorliegende Dokument enthält den Teil A des Abschlussberichts der durch das Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) in Auftrag gegebenen *Machbarkeitsstudie 2019 - Aufbau einer Wissensplattform im Umweltbereich zur Planungsbeschleunigung von Infrastrukturmaßnahmen*. Teil A beschreibt das im Rahmen der Studie entwickelte System. Es wurde unter Zugrundelegung der Anforderungen der Leistungsbeschreibung sowie der besonderen Nutzeranforderungen, die im Rahmen zweier Workshops ermittelt wurden, durch den Auftragnehmer Condat AG unter Mitwirkung des Deutschen Forschungsinstitutes für Künstliche Intelligenz (DFKI) erstellt. Das System besteht im Wesentlichen aus zwei Bestandteilen: Der Wissensplattform einerseits, die im Rahmen der Machbarkeitsstudie prototypisch realisiert wurde und deren Zweck u.a. darin besteht, die Eignung verschiedener Verfahren der Künstlichen Intelligenz (KI) bzgl. des automatisierten Aufbaus einer solchen Wissensplattform für den Umweltbereich zu prüfen. Andererseits der Benutzeroberfläche, die exemplarisch zeigt, wie man seitens der Nutzenden auf die Informationen in dieser Wissensplattform zugreifen kann; die Benutzeroberfläche wird im Folgenden auch als „Demonstrator“ bezeichnet.

Teil A des Abschlussberichts ist wie folgt aufgebaut: Im folgenden Kapitel 2 werden die Ziele und Anforderungen an das System zusammengefasst. In Kapitel 3 wird ein Überblick über das Gesamtsystem gegeben. In Kapitel 4 wird der technische Workflow beschrieben, dessen Ziel der automatisierte Aufbau der Wissensplattform ist. In Kapitel 5 werden spezielle KI-basierte Verfahren, die entwickelt wurden, um spezielle Funktionen in der Benutzeroberfläche anbieten zu können, beschrieben. In Kapitel 6 wird die Benutzeroberfläche vorgestellt. In Kapitel 7 werden die einzelnen Komponenten des Systems detailliert erläutert. Teil A des Berichts endet in Kapitel 8 mit einem Fazit.

2. Anforderungen und Ziele

Die Anforderungen, welche im Rahmen der Machbarkeitsstudie an Funktionalität und Nutzung einer Wissensplattform im Umweltbereich gestellt wurden, ergaben sich aus den folgenden drei Quellen:

- Leistungsbeschreibung vom 17.05.2019 entsprechend den Vergabeunterlagen im Rahmen der Ausschreibung des Auftrags
- Ergebnisse des ersten Workshops vom 09.09.2019
- Ergebnisse des zweiten Workshops vom 12.11.2019

2.1 Anforderungen aus der Leistungsbeschreibung

Das übergeordnete Ziel der Machbarkeitsstudie war, herauszufinden, ob mit Mitteln der Künstlichen Intelligenz (KI) bei der Umsetzung von Verkehrsvorhaben eine Planungsbeschleunigung durch Herstellung größtmöglicher Rechtssicherheit im Hinblick auf die Beachtung von umweltrechtlichen Anforderungen im Rahmen des Planungs- und Genehmigungsverfahrens erzielt werden kann. Die Rechtssicherheit sollte dabei durch größtmögliche Vollständigkeit der der Planung und Genehmigung zugrunde zulegenden rechtlichen und fachlich/wissenschaftlichen Informationen erzielt werden. Die hierbei maßgeblichen Informationen beziehen sich dabei insbesondere auf den Umgang mit geschützten Tierarten im Rahmen der Umweltverträglichkeitsprüfung eines Vorhabens.

Aus diesem übergeordneten Ziel ergaben sich gemäß Leistungsbeschreibung die folgenden funktionalen Anforderungen an die zu erstellende Wissensplattform:

- Die Machbarkeitsstudie soll klären, inwieweit sich der Suchaufwand bzgl. des Umgangs mit gefährdeten Tierarten im Rahmen der Planung und Genehmigung von Verkehrsvorhaben durch die Anwendung von Methoden der Künstlichen Intelligenz (KI) minimieren lässt.
- Die Wissensplattform soll dabei die erforderlichen Informationen selbständig finden, identifizieren, verknüpfen und dem Anwender zur Verfügung stellen.
- Dies soll durch KI gewährleistet werden, d.h. ein manueller Aufbau der Wissensplattform sollte nicht erforderlich sein.
- Mit Hilfe der Wissensplattform sollen alle relevanten Dokumente aus planerischer, fachlicher und rechtlicher Sicht gefunden und dem Nutzer zur Verfügung gestellt werden.
- Für die Machbarkeitsstudie wird der Umfang, der durch die Plattform zu re-

cherchierenden Aspekte auf zehn geschützte Tierarten, sowie Rechercheanfragen aus den Bereichen Wasserstraße und Schienenwege begrenzt.

2.2 Anforderungen aus dem 1. Workshop

Der erste Workshop fand am 09.09.2019 statt. An diesem Workshop haben Vertreterinnen und Vertreter des BMVI und der Bundesanstalt für Verwaltungsdienstleistungen (BAV) für den Aufraggeber, des Eisenbahn-Bundesamtes (EBA) und des Wasserstraßen-Neubauamtes Berlin (WNA Berlin) für die Planungsbehörden, der Generaldirektion Wasserstraßen und Schifffahrt (GDWS) für die Genehmigungsbehörden, sowie der Condat AG und des Deutschen Forschungsinstitutes für Künstliche Intelligenz (DFKI) für den Auftragnehmer teilgenommen. Ziel des Workshops war die Präzisierung der Anforderungen an Funktionalität und Inhalt einer Wissensplattform im Umweltbereich. Durch Gespräche und Diskussion mit den potentiellen Nutzenden aus den Planungs- und Genehmigungsbehörden wurde ihr konkreter Recherchebedarf identifiziert. Im Folgenden werden die Ergebnisse des Workshops aufgeführt und erläutert. [1]

2.2.1 Tierarten

Folgende Tierarten wurden für die Entwicklung der Wissensplattform im Rahmen der Machbarkeitsstudie gemeinsam mit den Nutzern festgelegt:

- Baumfalke
- Biber
- Eisvogel
- Eremit
- Fledermaus
- Haselmaus
- Kreuzkröte
- Nachtkerzenschwärmer
- Rapfen
- Zauneidechse

[1] In diesem Abschnitt werden nur die Ergebnisse des Workshops wiedergegeben, die als Anforderungen an die Realisierung des Demonstrators relevant sind. Für den vollständigen Bericht vgl. AP1 Dokumentation Workshop zur Ermittlung der Nutzeranforderungen am 09.09.2019“ vom 16.10.2019.

2.2.2 Quellen

Die potentiellen Anwendenden benannten die nachfolgend aufgeführten digitalen Quellen, die sie im Rahmen ihrer üblichen Online-Recherche verwenden und deren Dokumente in einer Wissensplattform aus Sicht der Nutzenden enthalten sein sollten:

- juris.de[2]
- Portale der Länder
- artenschutz.naturschutzinformationen.nrw.de
- Geo-Portale der Länder[3]
- Leitfäden für Artenschutz
- Arbeitshilfen der Länder
- Artenschutzdatenbank BfN/ WISIA
- Homepage Vorhabenträger
- Homepage Genehmigungsbehörden
- Homepage Anhörungsbehörden
- UVP Portal Länder/ Bund
- Amtsblätter Länder

Im Nachgang des Workshops wurden aus diesen allgemeinen Kategorien von Quellen dann konkrete Quellen (d.h. Internet-Adressen) abgeleitet, evaluiert und diese als Grundlage für einen Import bekannter Quellen in die Wissensplattform herangezogen. Zur Auflistung dieser konkreten Quellen s. Anhang B Quellensammlung.

Die grundsätzliche Aufnahme dieser Quellen in eine Wissensplattform und deren regelmäßige Verarbeitung wurde für sinnvoll erachtet. Gleichzeitig wurde klargestellt, dass im Rahmen der Machbarkeitsstudie bei der Entwicklung des Demonstrators ausschließlich kostenfrei digital verfügbare Quellen herangezogen werden können.

Nichtdigitale Quellen wurden für die Berücksichtigung im Rahmen der Entwicklung des Demonstrators ausgeschlossen.

---

[2] juris.de ist kostenpflichtig und konnte deshalb nicht berücksichtigt werden. Eine Kontaktaufnahme mit open-jur.de wurde nicht beantwortet. Als Quelle für Gerichtsurteile etc. wurde dann openlegaldata.io herangezogen, ein Open Data Archiv bei dem alle Gerichtsurteile, Beschlüsse, etc. veröffentlicht werden, und dies auch mit dem Ziel einer europaweiten Abdeckung.

[3] Die Geo-Portale der Länder werden hier der Vollständigkeit halber mit erwähnt, sie waren aber explizit vom Umfang der zu berücksichtigenden Quellen im Rahmen der Machbarkeitsstudie ausgeschlossen.

2.2.3 Dokumentarten

Folgende Dokumentarten wurden durch die im Workshop anwesenden potentiellen Nutzenden beispielhaft als für eine Recherche relevant identifiziert:

- Gutachten, Handlungsempfehlungen, Best-Practice-Beispiele, Aufsätze in Fachzeitschriften sowie Standard- und Regelwerke,
- relevante Leitfäden und Fachkonventionen,
- Dokumente zur Methodik der Bestandserfassung und -bewertung,
- Dokumente zu Auswirkungen der Verkehrsprojekte auf bestimmte Arten,
- Dokumente zur Wirksamkeit und zu Einsatzmöglichkeiten von Vermeidungs- und Kompensationsmaßnahmen,
- Dokumente zur Rechtsprechung und deren Auswertung und
- Forschungsberichte sowie Dokumente zu Monitoring-Ergebnissen.

2.2.4 Suchbegriffe

Folgende Suchbegriffe bzw. Suchkategorien, die im Zuge einer Recherche im Internet bzgl. des Umgangs mit einer geschützten Tierart relevant sind, wurden von den potentiellen Nutzenden identifiziert:

- Wissenschaftlicher Name der Art
- Gattungsname
- Synonyme
- CEF-Maßnahmen
- Name Gewässer/ Fluss
- Name Gemeinde/ Region
- Bezeichnung Schutzgebiet
- FFH-Gebietsnummer
- Ausgleichmaßnahme
- Ersatzmaßnahme
- Ökologische Baubegleitung
- Flächenpool
- Kompensationsmaßnahme
- FCS-Maßnahme

2.2.5 Sprachen

Es wurde klargestellt, dass in der Version des Demonstrators als Hauptsprache ausschließlich deutsch verwendbar sein wird. Lediglich bei den Tierarten und Gattungen sollten auch die lateinischen Fachnamen suchbar sein.

2.2.6 Relevanzkriterien

Für die Bewertung der Relevanz von Quellen bzw. Dokumenten benannten die potentiellen Anwendenden die folgenden Kriterien nach Wichtigkeit in dieser Reihenfolge:

- Veröffentlichungsdatum
- Regionaler Bezug der Quelle
- Ersteller Quelle / Dokument

2.2.7 Gültigkeitsdauer von Informationen

Es wurde deutlich, dass es für einzelne Dokumentarten unterschiedliche Gültigkeitsdauern geben kann. So seien nach Aussage der Nutzenden z.B. Gutachten für Kartierungen nur fünf Jahre gültig. Hieraus leitete sich der Wunsch der Nutzenden ab, die Gültigkeitsdauer den Dokumenten manuell zuweisen zu können. Diese Anforderung wurde im Demonstrator nicht umgesetzt, da es sich um eine funktionale Anforderung handelt, deren grundsätzliche Machbarkeit außer Frage steht. Diese Funktionalität wird als wünschenswerte Anforderung für die Produktivversion mit aufgenommen. Eine automatische Zuordnung direkt beim Import ist nach derzeitigem Wissensstand nicht möglich

2.2.8 Suchfunktionen

In der Diskussion mit den potentiellen Nutzenden ergaben sich die folgenden funktionalen Anforderungen an die Suchfunktionen in der Benutzeroberfläche:

Feld	Beispiel / Anmerkungen	Pflichtangabe
Tierart	Aus der Liste der zehn Tierarten	ja
Region	Bundesländer, Schweiz, Österreich	nein
Quelle	z.B. eba.de	nein
Dokumentenart	Fachlich, Planerisch, Rechtlich	nein
Stichwortfeld		nein
Datumsbereich	Publikationsdatum von - bis	nein

Suchhistorie		nein
Suchergebnisliste		ja
Einzelansicht Suchergebnis		ja
Nutzerrückkopplung	Möglichkeit eines Feedbacks durch Nutzende, ob das Suchergebnis relevant ist[4]	nein

2.2.9 Vokabular

Im Nachgang zum ersten Workshop wurde gemeinsam mit den potentiellen Nutzenden eine Vokabularsammlung erstellt, in welcher wichtige Begriffe im Rahmen einer fachbezogenen Recherche aufgenommen wurden. Diese Vokabularsammlung verfolgte den Zweck, den Inhalt der Quellen[5] zu beschreiben. Sie ist zentraler Bestandteil der Wissensplattform, da sie jenes fachliche Wissen enthält, das aus einer einfachen Internet-Suchmaschine und Dokumentensammlung eine tatsächliche *Wissensplattform* macht. Dies wird u.a. dazu benutzt, über ein spezielles KI-Verfahren, die sog. *semantische Indexierung* relevante von irrelevanten Dokumenten zu unterscheiden und bei einer Suchanfrage die relevantesten Dokumente finden zu können.

Seitens Condat AG wurde diese Vokabularsammlung in eine digital verarbeitbare Form überführt, eine sog. *Ontologie*:

> *Ontologien in der Informatik sind meist sprachlich gefasste und formal geordnete Darstellungen einer Menge von Begrifflichkeiten und der zwischen ihnen bestehenden Beziehungen in einem bestimmten Gegenstandsbereich. Sie werden dazu genutzt, „Wissen" in digitalisierter und formaler Form zwischen Anwendungsprogrammen und Diensten auszutauschen. Wissen umfasst dabei sowohl Allgemeinwissen als auch Wissen über sehr spezielle Themengebiete und Vorgänge.* [6]

[4] Die Möglichkeit des Nutzer-Feedbacks wurde im finalen Demonstrator nicht umgesetzt, zur Diskussion hierzu s. Abschnitt 6.

[5] Zur Liste der Quellen siehe Anhang B Quellensammlung.

[6] Wikipedia, Ontologie (Informatik), https://de.wikipedia.org/wiki/Ontologie_%28Informatik%29

Das Verfahren der semantischen Indexierung wird in Abschnitt 7.9 ausführlich erläutert.

2.3 Anforderungen aus dem 2. Workshop

Der zweite Workshop fand am 12.11.2019 statt. An diesem Workshop haben Vertreterinnen und Vertreter des BMVI und der Bundesanstalt für Verwaltungsdienstleistungen (BAV) für den Aufraggeber, des Eisenbahn-Bundesamtes (EBA) und des Wasserstraßen-Neubauamtes Berlin (WNA Berlin) für die Planungsbehörden, der Generaldirektion Wasserstraßen und Schifffahrt (GDWS) und ebenfalls des EBA für die Genehmigungsbehörden, sowie der Condat AG und des Deutschen Forschungsinstitutes für Künstliche Intelligenz (DFKI) für den Auftragnehmer teilgenommen. Der Workshop verfolgte das Ziel, anhand der Präsentation des aktuellen Entwicklungsstandes des Demonstrators durch die Condat AG die finalen Anforderungen an die Benutzeroberfläche und etwaiges Verbesserungspotential in der Funktionalität der Plattform zu identifizieren.[7]

Als Ergebnis des Workshops wurden gemeinsam mit den potentiellen Nutzenden die folgenden Festlegungen für den Funktionsumfang des finalen Demonstrators getroffen.

2.3.1 Zentrales Suchfeld

Es soll einen Sucheinstieg über ein zentrales Suchfeld geben, über das i.d.R. sehr allgemeine Suchbegriffe eingegeben werden.

2.3.2 Suchergebnisse

Im Workshop wurde demonstriert, dass als Suchergebnisse sowohl ganze Dokumente als auch Teile von Dokumenten (sog. Fragmente) angezeigt werden können. Von den Teilnehmenden des Workshops wurde erkannt, dass die Möglichkeit der direkten Suche in Fragmenten nicht sinnvoll ist, d.h. als Suchergebnisse nur ganz Dokumente bzw. Internet-Quellen aufgelistet werden sollten. Die Verwendung von Fragmenten

[7] In diesem Abschnitt werden nur die Ergebnisse des Workshops wiedergegeben, die Anforderungen an den zu realisierenden Funktionsumfang des finalen Demonstrators betreffen. Für den vollständigen Bericht zum Workshop inkl. der Darstellung des Status Quo zu diesem Zeitpunkt vgl. „AP2/AP3 Bericht - Workshop zur Präsentation und Diskussion des Zwischenstands 12.11.2019“ vom 19.12.2019.

sei aber, so das Feedback, bei der Navigation innerhalb von gefundenen Dokumenten eine sinnvolle Zusatzfunktion.

2.3.3 Explorative Suche/Clustering

Im Workshop wurde als besonderes Feature die *Explorative Suche* demonstriert. Diese gruppiert die im Internet gefundenen Suchergebnisse automatisch unter Verwendung von KI-Verfahren des maschinellen Lernens in verschiedene Unterthemen („Clustering“). Die explorative Suche wird angeboten, um eine Vorauswahl treffen zu können und die Suchergebnisse auf einer ersten Stufe zu verfeinern, indem v.a. viele irrelevante Suchergebnisse gleich ausgeblendet werden können. Das themenzentrierte Clustering basiert auf einer Kombination der explorativen Suche und der semantischen Indexierung auf Basis des Vokabulars. Dieses Clustering wurde von den potentiellen Nutzenden als sehr hilfreich empfunden und sollte deshalb in die finale Version des Demonstrators aufgenommen werden.

2.3.4 Anzeige von Metadaten

Die relevanten Suchergebnisse sollen möglichst bereits im Rahmen der Ergebnisliste mit den wesentlichen Metadaten angezeigt werden (Datum, Autoren/Organisation, Lokaler Bezug, Dokumenttyp, ggf. Vorhaben).

2.3.5 Sortierung der Suchergebnisse

Die Sortierung der Suchergebnisse soll durch Relevanzkriterien nach Quelle, Datum und Dokumenttyp erfolgen; die Sortierung soll manuell geändert werden können.[8]

2.3.6 Sanduhrprinzip

Über Eingabe weiterer Suchkriterien soll es möglich sein, die Suchergebnisse immer weiter einzuschränken, bis nur noch wenige Ergebnisse übrig bleiben; sollten hierbei keine die Fragestellung beantwortenden Dokumente gefunden werden, wird die Su-

[8] Im finalen Demonstrator (s. Kapitel 6) wurde keine Sortierung nach Dokumenttyp realisiert, da stattdessen für die verschiedenen Dokumenttypen eigene Ergebnislisten angezeigt werden. Innerhalb dieser Ergebnislisten wiederum wurde dann eine Sortiermöglichkeit nach Relevanz, Titel und Datum angeboten. Eine Sortiermöglichkeit nach Quelle wurde nicht mit aufgenommen, da sie einerseits technisch sehr einfach umsetzbar ist, und damit die Machbarkeit außer Frage steht, andererseits angesichts der Tatsache, dass einzelne Quellen tausende von Dokumenten enthalten (wie z.B. eba.bund.de) inhaltlich nicht sinnvoll erscheint. Für eine ggf. zu realisierende Erweiterung sollte diese Frage dann noch einmal bewertet werden.

che durch Wegnahme von Suchbegriffen und ggf. Hinzufügen anderer Suchbegriffe wieder erweitert (Sanduhrprinzip / Zoom-In-Zoom-Out).

2.3.7 Navigation innerhalb von Dokumenten

Viele relevante Dokumente haben 100 und mehr Seiten. Um das Auffinden der relevanten Textstellen innerhalb eines Dokuments zu erleichtern, sollte eine zweigeteilte Detailsicht angeboten werden, bei der auf der linken Seite die im Suchkontext relevanten Textstellen angezeigt werden und auf der rechten Seite das komplette Dokument. Durch eine Phrasensuche soll erreicht werden, dass bei Eingabe mehrerer Begriffe in Absätze gesprungen werden kann, die diese Begriffe enthalten, sofern dies technisch bei PDF-Dokumenten mit vertretbarem Aufwand realisierbar ist.

2.3.8 Vorhabenbezogene Sicht

Als wünschenswert bezeichneten die potentiellen Nutzenden die Abbildung einer „Vorhabenssicht“ im Suchergebnis, d.h. eine qualifizierte und möglichst in richtiger zeitlicher Abfolge sortierte Zusammenstellung aller mit einem bestimmten Vorhaben verknüpften Dokumente, sodass auch der Status der vorgeschlagenen Maßnahmen schnell ersichtlich wird (z.B. genehmigt/ nicht genehmigt).

3. Überblick über das Gesamtsystem

In diesem und den nächsten Kapiteln dieses Abschlussberichts Teil A wird das aus den zuvor beschriebenen Anforderungen realisierte System beschrieben. Hierbei wird zunächst einen Gesamtüberblick über das System (Kapitel 3) gegeben. Anschließend wird der Workflow von der Suche im Internet zur Vorverarbeitung der in die Wissensplattform importierten Dokumente (Kapitel 4) beschrieben. In Kapitel 5 werden spezielle KI-basierte Verfahren beschrieben, die in der Benutzeroberfläche bei Suche und Navigation zur Anwendung kommen. In Kapitel 6 wird die Benutzeroberfläche beschrieben. In Kapitel 7 werden dann die Systemkomponenten, Dienste und Verfahren aus technischer Sicht detaillierter beschrieben.

Die Wissensplattform Umwelt ist eine KI-basierte Plattform, die so konzipiert ist, dass beliebige Dienste über einfache Schnittstellen in die Plattform integriert werden und über eine zugrundeliegende Kommunikationsinfrastruktur (ein sog. *Message Broker*) miteinander verknüpft werden können. Die einzelnen Komponenten sind als einfach wart- und austauschbare Microservices[9] (diese entsprechen in Abbildung 1 den hellblauen Kästchen) realisiert.

[9] Unter Microservices wird ein Ansatz der Entwicklung von Software-Systemen verstanden, bei denen einzelne Funktionen in kleine, selbständig funktionsfähige Einheiten aufgeteilt werden, die dann z.B. auch separat zur Steigerung der Performance auf anderen Servern laufen können. Diese einzelnen Einheiten kommunizieren dann z.B. wie in der Wissensplattform über eine Software wie RabbitMQ miteinander.

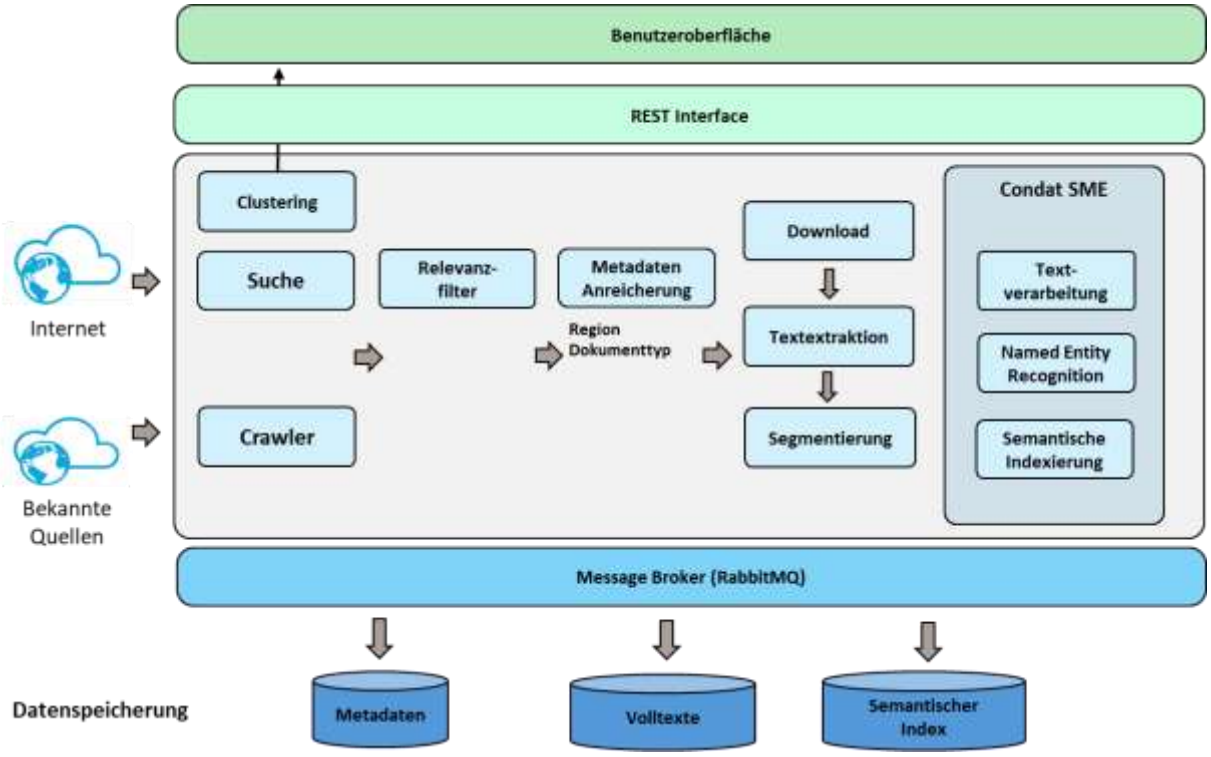

Abbildung 1: Gesamtüberblick Systemarchitektur

Abbildung 1 gibt einen Überblick über das Gesamtsystem. Dieses besteht aus fünf konzeptionellen Schichten:

3.1 Datenspeicherung

Es gibt drei wesentliche Arten von permanent gespeicherten Daten: 1) Metadaten zu Quellen, Suchergebnissen und Dokumenten, 2) Volltexte zu heruntergeladenen Dokumenten und 3) semantischen Index. Metadaten werden in einer speziellen Datenbank abgespeichert (hierbei kommt die Datenbank MongoDB[10] zum Einsatz), die Volltexte und erzeugten Fragmente in einem speziellen Volltextindex (hierfür wird die Suchmaschine Elasticsearch verwendet)[11] und der semantische Index in der Condat Smart Media Engine (SME), die intern wiederum ebenfalls Elasticsearch nutzt.

[10] MongoDB ist eine dokumentenorientierte Datenbank, die speziell für die Speicherung flexibler, d.h. nicht gleichförmiger und vorher festgelegter Datenstrukturen geeignet ist; z.B. kann hier einfach zu einzelnen gespeicherten Dokumenten weitere Metadaten hinzugefügt werden, ohne dass vorher ein Datenschema definiert und aktualisiert werden muss. Zu MongoDB s. www.mongodb.com

[11] Elasticsearch ist eine hocheffiziente Suchmaschine, die zur Indexierung von Volltexten, auch im Big Data Bereich, verwendet wird, s. www.elastic.co

3.2 Integrations- und Kommunikationsschicht (Message Broker)

Für die Integrations- und Kommunikationsschicht wird ein Message Broker eingesetzt. Über diesen werden die einzelnen Microservices miteinander gekoppelt, sodass hierüber der technische Workflow (s. nächster Abschnitt) in der Plattform abgewickelt wird.

Technisch kommt hierbei RabbitMQ[12] zum Einsatz.

3.3 Plattformdienste (Microservices)

Die Plattformdienste stellen den eigentlichen funktionalen Kern der Wissensplattform dar. Dies sind zum einen die voneinander unabhängigen Microservices, die jeweils einzelne Funktionen zur Verfügung stellen, wie z.B. die regionale Zuordnung von Texten, Dokumentarterkennung oder Textextraktion. Die speziellen Dienste zur Textverarbeitung und semantischen Indizierung sind in der Condat SME gebündelt. Technisch handelt es sich bei den einzelnen Microservices um unterschiedliche Komponenten, die z.T. als Java-, z.T. als Python-Komponenten realisiert sind.

3.4 Schnittstelle zu Daten und Diensten

Um die in der Wissensplattform gespeicherten und verarbeiteten Daten und Dienste nach außen zur Verfügung zu stellen, insbesondere zur Darstellung in der Benutzeroberfläche, wurde eine einheitliche Schnittstelle implementiert. Prinzipiell könnten über diese Schnittstelle z.B. auch externe Systeme angebunden werden.

Technisch handelt es sich hierbei um eine *REST-API-Schnittstelle*, die mittels eines eigenen Webservers realisiert wurde.

3.5 Applikationsserver / Benutzeroberfläche

Für die Realisierung der Benutzeroberfläche kommt zum einen ein Applikationsserver zum Einsatz, der auf Basis von node.js[13] entwickelt wurde und der die Kommuni-

[12] RabbitMQ ist eine sog. Message Broker Software, über die verschiedene Software-Komponenten per Nachrichtenaustausch miteinander kommunizieren können, s. https://www.rabbitmq.com/

[13] Node.js ist ein bei Webanwendungen weitverbreiteter Applikationsserver, der auf sog. *Server Side JavaScript* basiert, d.h. auf die Komponenten die mittels der Programmiersprache JavaScript realisiert sind, aber auf dem

kation mit dem REST-API abwickelt. Die Realisierung der eigentlichen Benutzeroberfläche erfolgte mittels des Web Framework Angular.[14]

Alle hier erwähnten und im Rahmen der Machbarkeitsstudie entwickelten oder angepassten Komponenten und Dienste werden im Folgenden genauer beschrieben.

Server ausgeführt werden und nicht wie bei JavaScript sonst meistens der Fall im Browser. Node.js zeichnet sich durch eine besonders hohe Performanz aus; s. https://nodejs.org/

[14] Angular ist eines der populärsten Frameworks für die Web-Applikations-Entwicklung und basiert auf der Programmiersprache *TypeScript*, die aus JavaScript entwickelt wurde:; s. https://angular.io/

4. Der Vorverarbeitungs-Workflow

Um die komplexen KI-basierten Funktionen auf der Benutzeroberfläche anbieten zu können, müssen von der Suche bis zur Verarbeitung der einzelnen Dokumente bzw. Texte eine Reihe von einzelnen vorbereitenden Schritten durchgeführt werden. Dies ist der Vorverarbeitungs-Workflow, der in diesem Abschnitt genauer beschrieben wird.

Den Kern der Wissensplattform stellen die einzelnen singulären Dienste dar, die sowohl einfache Basisdienste sein können (wie z.B. Download von Dokumenten), als auch komplexere, aber auf konventionellen Algorithmen basierende Dienste (z.B. Segmentierung), beziehungsweise genuine KI-basierte Dienste, wie z.B. Dokumentarterkennung oder *Named Entity Recognition*. Diese Dienste bauen teilweise aufeinander auf, teilweise können sie auch unabhängig voneinander aufgerufen werden. Sie werden in einem KI-basierten Workflow zu einer Verarbeitungskette kombiniert. Dieser Workflow wird im Folgenden beschrieben (Abbildung 2).

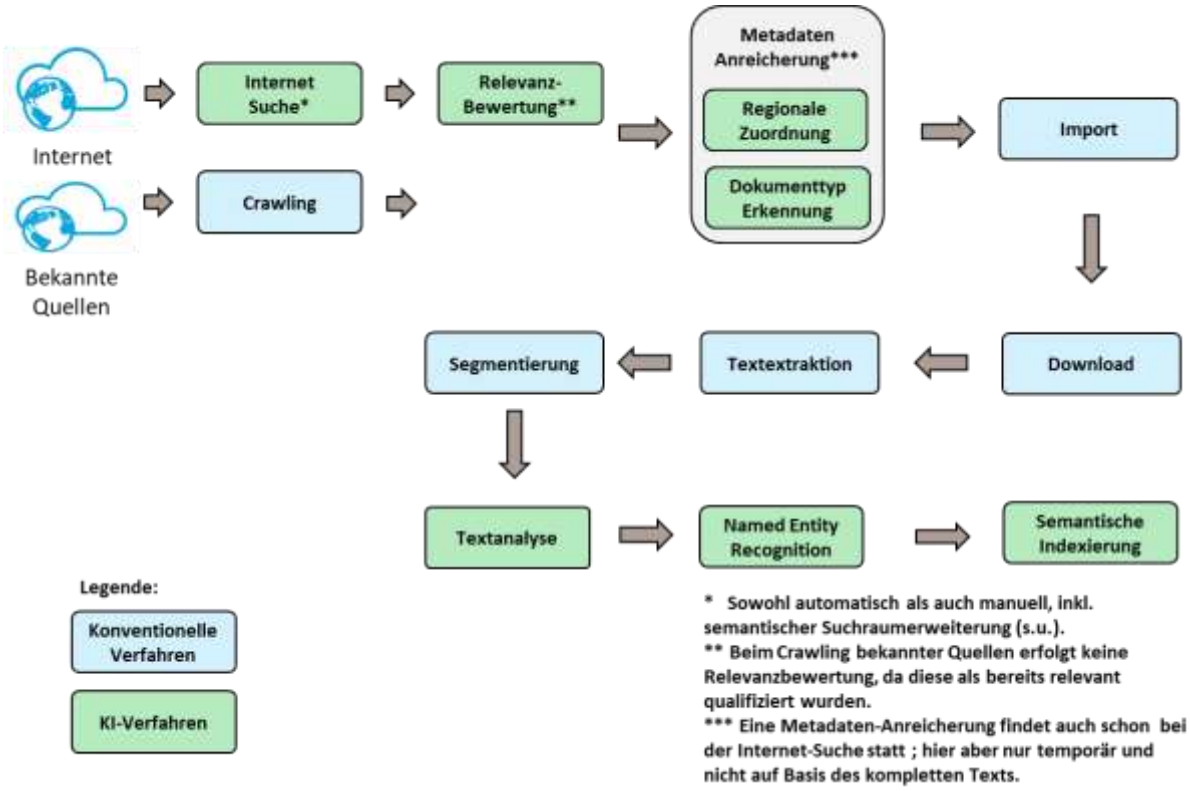

Abbildung 2: Workflow der Wissensplattform Umwelt

4.1 Finden relevanter Dokumente

Die Suche nach relevanten Dokumenten stellt den Beginn der gesamten Verarbeitungskette dar. Sie erfolgt durch drei unterschiedliche Kanäle:

1. Import bekannter Quellen
2. Regelmäßige automatische Internet-Suchanfragen
3. Manuelle Suchanfragen durch Anwender

Zu den einzelnen Kanälen:

4.1.1 Import bekannter Quellen

Die im ersten Workshop identifizierten Quellen (s. Anhang B) wurden durch das System initial komplett in die Plattform importiert. Wenn es sich um eine normale Website handelt (wie in den meisten Fällen), ist hierzu eine Komponente erforderlich, die den Links der Quelle folgt (*Crawling*) und die Inhalte extrahiert (*Scraping*) bzw., sofern es sich um PDF-Dokumente handelt, der Downloader-Komponente übergibt. Das Scraping und der Download musste für jede Quelle separat konfiguriert und implementiert werden.

Für den Fall, dass es sich um datenbankbasierte Quellen handelt, die eine eigene Abfrageschnittstelle anbieten, wird hierfür jeweils ein eigener Konnektor implementiert, d.h. eine Komponente, die die Daten der jeweiligen Quelle importiert. Exemplarisch wurde dies für die Datenquelle *openlegaldata.io* realisiert, die eine Datenbank mit allen Gerichtsurteilen und Beschlüssen als Open Data Quelle zur Verfügung stellt.

Zur detaillierten Beschreibung der Import-Komponente s. Abschnitt 7.1.

4.1.2 Regelmäßige automatische Internet-Suchanfragen

Sehr viele relevante Dokumente werden bereits über das Crawlen bekannter Quellen gefunden. Über die Ausführung regelmäßiger automatischer Internet-Suchanfragen soll darüber hinaus sichergestellt werden, dass relevante Dokumente auch aus bisher unbekannten Quellen bzw. von Websites gefunden werden, die nicht als Ganzes relevant sind, wie z.B. die Web-Auftritte von Gemeinden oder Planungsbüros, die nur einzelne relevante Dokumente enthalten. Zu diesem Zweck wurde eine Reihe von

Suchanfragen auf Grundlage der im ersten Workshop identifizierten Suchbegriffe (s. 2.2.4) definiert. Diese automatischen Internet-Suchanfragen werden jeweils nächtlich ausgeführt.

Technische Details zur Realisierung der Suchkomponente sind in Abschnitt 7.2 beschrieben.

4.1.3 Manuelle Suchanfragen

Die Benutzeroberfläche (s. Kapitel 6) bietet die Möglichkeit, manuell Internet-Suchanfragen zu stellen. Dies dient zum einen dazu, Ergebnisse zusätzlich über Abfragen zu finden, die weder durch das Crawlen bekannter Sites gefunden werden, noch durch die automatischen Abfragen. Zum anderen soll damit den Benutzenden auch das Vertrauen gegeben werden, dass die Plattform eine Suchfunktionalität bietet, die der reinen Internet-Suche auf jeden Fall mindestens äquivalent ist.

Die Ergebnisse der manuellen Suchanfragen werden über die reine Darstellung analog einer Suche mit Bing noch um folgende Funktionalitäten erweitert:

- Dokumentkategorie (Fachinformation, Rechtliches, Planung)
- Regionale Zuordnung
- Relevanzbewertung
- Automatisches Clustering der Ergebnisse

Ergebnisse, die über den Relevanzfilter als relevant eingestuft werden und noch nicht in der Plattform gespeichert sind, werden gespeichert und in die weitere Verarbeitungskette mit aufgenommen. Ebenso werden die Suchabfragen der Benutzenden (anonym) gespeichert, sofern sie mehr Suchbegriffe enthalten als die Tierart und diese Suchbegriffe in der Ontologie enthalten sind.[15] Durch diese beiden Maßnahmen wird bewirkt, dass die Plattform durch die Benutzung „lernt“ und sowohl Suchen als auch Ergebnisse kontinuierlich und automatisch verbessert werden. Dieses Lernen ist in diesem Fall aber keine KI-Technologie, sondern lediglich ein kontinuierliches Abspeichern neuer Suchabfragen, die dann auch wieder für die automatischen Suchen zusätzlich verwendet werden.

4.1.4 Semantische Suchraumerweiterung

[15] Für den Produktivbetrieb müssten dann Suchbegriffe, die nicht in der Ontologie enthalten sind, separat gesammelt werden und es müsste administrativ entschieden werden, ob diese Suchbegriffe in die Ontologie mit aufgenommen werden sollen.

Des Weiteren werden manuelle Suchanfragen durch die semantische Suchraumerweiterung ergänzt, um noch bessere Suchergebnisse zu erhalten. Diese beruht auf dem gemeinsam mit den potentiellen Anwendern definierten Vokabular. Auf diese Weise wird z.B. der Suchbegriff „Fledermaus“ durch den lateinischen Begriff (*Microchiroptera*) und die Unterarten der Fledermaus ergänzt.

4.1.5 Relevanzbewertung

Für Dokumente, die über die automatische oder die manuelle Suche gefunden werden, wird eine Relevanzbewertung vorgenommen, bevor diese zur weiteren Verarbeitung in die Plattform importiert werden. Dokumente, die aus bekannten und als relevant identifizierten Quellen, bzw. relevanten Unterkategorien dieser Quellen stammen, werden ohne gesonderte Bewertung direkt in die Plattform importiert.

Zur detaillierten Beschreibung des Relevanzfilters s. Abschnitt 7.3.

4.1.6 Metadatenanreicherung

Für alle Suchergebnisse wird eine Anreicherung der Standardmetadaten um spezielle Metadaten durchgeführt, die für die Wissensplattform von besonderer Bedeutung sind. Dies sind regionale Zuordnung und Dokumenttyperkennung.

Die Suchergebnisse aus einer Bing-Suchabfrage enthalten z.B. standardmäßig die folgenden Metadaten:

Name	**Beispiel**
Name	Natura 2000 | Startseite | MLUL - mlul.brandenburg.de
URL	https://mlul.brandenburg.de/mlul/de/umwelt/natur/natura-2000
Snippet[16]	Die Umsetzung von Natura 2000 gehört in Brandenburg zu den wichtigsten Fachaufgaben der Naturschutzbehörde. 27 Vogelschutzgebiete und 595 FFH-Gebiete umfassen die Natura 2000-Gebietskulisse in Brandenburg. Dies entspricht einem Anteil von rund 26 Prozent an der Landesfläche.
dateLastCrawled	2020-01-12 12:37:59
Language	de

Durch die Metadatenanreicherung werden diese Daten wie folgt angereichert:

[16] Die Snippets werden von Bing erzeugt und dienen (wie man das analog auch von Google kennt) der Anzeige kurzer Textstellen aus den Suchergebnissen, die die Suchbegriffe enthalten.

Region	BB für Brandenburg
Dokumenttyp	Fachinformation

Für beide Metadatentypen wurden spezifische KI-basierte Verfahren entwickelt bzw. angepasst.

Region

Kern der regionalen Zuordnung ist die Erkennung des Bundeslandes aus beliebigen Texten, bei denen es sich um URLs handeln kann (wie eberswalde.de), Text-Snippets wie in der Tabelle angegeben oder auch ganzen Texten, die aus PDF-Dokumenten extrahiert wurden.

Dieses KI-basierte Verfahren wurde noch um ein heuristisches Verfahren erweitert, das URLs mit Domains, die mit *.ch*, *.at*, etc. enden, das entsprechende Land ableitet.

Die technischen Details der Realisierung sind in Abschnitt 7.4 beschrieben.

Dokumenttyp

Im Kontext der Wissensplattform sind drei Dokumenttypen als besonders relevant klassifiziert worden: fachliche Dokumente, planungsbezogene Dokumente und rechtliche Dokumente (z.B. Gerichtsurteile). Die Dokumenttyperkennung dient der entsprechenden Klassifikation. Hierbei handelt es sich um ein genuines KI-Verfahren, das auf Training und der Definition von Trainings-Sets beruht. Für eine detaillierte Beschreibung s. Abschnitt 7.5.

4.2 Import in die Wissensplattform

Alle Dokumente bzw. Suchergebnisse, die über die oben angegebenen drei Suchverfahren gefunden und ggf. durch den Relevanzfilter als relevant eingestuft wurden, werden dann permanent in die Wissensplattform aufgenommen und weiterverarbeitet.

Der erste Schritt hierzu ist die Speicherung der Metadaten der Suchergebnisse inklusive der erweiterten Metadaten Region und Dokumenttyp in der Metadatendatenbank (vgl. Abschnitt 3.1).

Technisch kommt hierbei MongoDB zum Einsatz.

Der Import stellt den Beginn der eigentlichen Verarbeitungskette der Wissensplatt-

form dar.

4.3 Download und Textextraktion

Der Import der Quellen umfasst im Wesentlichen deren Metadaten, hierüber werden die Quellen dem System bekannt gemacht und als bekannte und relevante Quellen in der Wissensplattform gespeichert. Wenn es sich hierbei um PDF-Dokumente handelt – und dies ist in sehr vielen Fällen der Fall – werden diese zusätzlich physisch heruntergeladen, um sie weiter analysieren zu können. Dies bietet auch den Vorteil, dass diese Dokumente auch dann noch zur Verfügung stehen, wenn deren Veröffentlichungszeitraum im Internet überschritten ist, da viele vorhabensbezogene Dokumente von den Behörden nur für die Dauer der Planung bzw. Umsetzung veröffentlicht werden.

Das Verfahren der Textextraktion und verwendete Werkzeuge werden in Abschnitt 7.7 genauer beschrieben.

4.4 Segmentierung

Auf Basis der Extraktion der kompletten Texte eines Dokuments wird dieses im nächsten Schritt in einzelne Segmente unterteilt. Dies dient dazu, in der Detailansicht die im Suchkontext relevanten Abschnitte anzeigen und filtern zu können. Dies ist insbesondere bei Dokumenten hilfreich, die 100 und mehr Seiten lang sind, was bei vielen, insbesondere planungsbezogenen Dokumenten der Fall ist.

Die Segmentierungskomponente ist in Abschnitt 7.8 detaillierter beschrieben.

4.5 Textanalyse

Im Anschluss an die Textextraktion und die Segmentierung erfolgt eine Analyse der Texte. Hierbei erfolgt die grammatikalische Analyse der Texte, wobei neben Satzbestandteilen wie Subjekt, Prädikat und Objekt insbesondere auch die sogenannten *Nominalphrasen*, d.h. diejenigen Satzbestandteile, die die in der Vokabularsammlung aufgeführten Begriffe ob als einzelner Begriff („Zauneidechse“), oder als größerer Satzbestandteil („die in diesem Gebiet selten vorkommende Zauneidechse“)) erkannt werden sollen. Hier kommen Techniken der Verarbeitung natürlicher Sprache (engl. Natural Language Processing, NLP) zum Einsatz. Dies wird im Abschnitt 7.9.2 näher

erläutert.

4.6 Named Entity Recognition

Auf der Grundlage der erfolgten Textanalyse kann dann die sog. *Named Entity Recognition* erfolgen:

> *Named-entity recognition (NER) oder Eigennamenerkennung ist eine Aufgabe in der Informationsextraktion und bezeichnet die automatische Identifikation und Klassifikation von Eigennamen. Ein Eigenname ist eine Folge von Wörtern, die eine real existierende Entität beschreibt, wie z. B. ein Firmenname.*[17]

Die zu erkennenden Eigennamen sind im Fall der Wissensplattform Umwelt die in der Vokabularsammlung aufgeführten Begriffe, wie z.B. „Zauneidechse“, „Bundesnaturschutzgesetz“ oder „FFH-Gebiet“; hierbei soll insbes. erkannt werden, dass es sich bei verschiedenen Ausdrücken um dieselbe Entität handelt (z.B. Bundesnaturschutzgesetz entspricht „BNatschG“).

4.7 Semantische Indexierung

Auf der Grundlage der erfolgten Named Entity Recognition kann dann die KI-Kernfunktionalität der Plattform realisiert werden, die *Semantische Indexierung*, und zwar sowohl der kompletten Dokumente als auch der jeweils einzeln daraus extrahierten Fragmente. Bei der semantischen Indexierung werden relevante Begriffe (im Kontext der Ontologie als *Konzepte* bezeichnet) erkannt. Dies erfolgt nicht nur wie in einer normalen Volltextsuche in der direkten Weise ihres textuellen Vorkommens, sondern vielmehr auch als Begriff, der z.B. in verschiedenen Schreibweisen, Singular/Plural, anderen Sprachenauftauchen kann und damit in Beziehung zu anderen Begriffen aus dem gleichen Kontext gesetzt werden kann. So kann über die semantische Indexierung beispielsweise erkannt werden, dass ein Text von Zauneidechsen handelt, auch wenn hierin nur der Terminus „L. agilis“ auftaucht.

Grundlage der semantischen Indexierung ist die sog. *Ontologie*, d.h. ein nach formalen Kriterien aufgebautes Vokabular. Wie in Abschnitt 2.2.9 beschrieben, wurde für die Machbarkeitsstudie zusammen mit den potentiellen Anwendern ein spezielles fachbezogenes Vokabular erstellt, da dies in diesem Kontext wesentlich genauere Ergebnisse ermöglicht als eine Generierung des Vokabulars beispielsweise auf

[17] Wikipedia, https://de.wikipedia.org/wiki/Named-entity_recognition

Grundlage von Wikipedia und auch sukzessive erweitert und angepasst werden kann.

Ergebnis der semantischen Indexierung ist ein sog. *Semantischer Fingerabdruck (Semantic Fingerprint).* Hierunter ist die Gesamtheit aller Begriffe aus der Vokabularsammlung (bzw. genauer deren technischem Äquivalent, den Konzepten) zu verstehen, die explizit oder implizit dem entsprechenden Gesamttext oder Fragment inkl. einer Gewichtung zugeordnet werden können.

Die semantische Indexierung ist auch Grundlage für den Aufbau eines sog. *Knowledge Graph*, über den die Informationen innerhalb der Plattform miteinander verknüpft werden können. Der *Knowledge Graph* ermöglicht es, sehr strukturiert alle Informationen zu einem bestimmten Thema darzustellen und durch den dadurch entstehenden Wissensraum zu navigieren. Z.B. können dadurch alle Dokumente zu einem bestimmten Vorhaben (vorhabensbezogene Darstellung) miteinander verknüpft werden, oder bestimmte relevante Eigenschaften strukturiert gesammelt werden, wie Fluchtdistanzen oder Schutzstatus bei geschützten Tierarten.

Für die semantische Indexierung kommt bei der Wissensplattform die Condat Smart Media Engine (SME) zum Einsatz. Für eine detaillierte Beschreibung des Prozesses und der Komponente s. Abschnitt 7.9.

5. KI-basierte Dienste für die Benutzeroberfläche

Der in Abschnitt 4 beschriebene Workflow ist die Voraussetzung für die Realisierung der speziellen KI-basierten Dienste für die Suche, Navigation und Darstellung innerhalb der Benutzeroberfläche, um die besonderen Anforderungen der Machbarkeitsstudie zu erfüllen. Hierbei handelt es sich insbesondere um die beiden Verfahren *Explorative Suche durch Clustering* und *Semantische Suche durch Semantic Fingerprinting*.

5.1 Explorative Suche durch *Clustering*

Das Clustering der Suchergebnisse ist ein sogenanntes ad-hoc-Verfahren, das zur Gruppierung der Suchergebnisse der manuellen Internetsuche in der Benutzeroberfläche verwendet wird. Hierbei werden aus der Menge der Suchergebnisse Gruppen (engl. Cluster) gebildet, die thematisch zusammenhängende Suchergebnisse gruppieren. Es dient dazu, schnell durch eine große Anzahl von Suchergebnissen zu navigieren und durch eine Vorklassifikation relevante von irrelevanten Ergebnissen unterscheiden zu können („*Explorative Suche*“). Da die Clustering-Ergebnisse immer abhängig von den jeweiligen aktuellen Ergebnissen sind und damit immer nur temporär gültig sind, ist es nicht sinnvoll, diese Cluster für die weitere Verarbeitung abzuspeichern.

Für eine detaillierte Beschreibung der Clustering-Komponente vgl. Abschnitt 7.6)

5.2 Semantische Suche durch *Semantic Fingerprinting*

Der große Vorteil der semantischen Fingerabdrücke liegt darin, dass sie von der Struktur und der Sprache der zugrundeliegenden Dokumente unabhängig sind. Sie ermöglichen es dadurch, völlig unterschiedliche Dokumente miteinander zu vergleichen, z.B. deutsch- und englischsprachige[18] und bei diesen hierdurch festzustellen, dass sie thematisch eine hohe Übereinstimmung haben, wenn beide von CEF-Maßnahmen bei Zauneidechsen handeln. Dies geschieht durch eines der Kernmodule der Condat Smart Media Engine, der sog. *Recommender*-Komponente. Durch die

[18] Nicht-deutschsprachige Dokumente waren im Rahmen dieser Machbarkeitsstudie ausgeschlossen. Insofern soll dieses Beispiel hier v.a. der Illustration der Möglichkeiten dienen. Für den inhaltlichen Vergleich zweier deutschsprachiger Dokumente gilt dies entsprechend.

semantische Ähnlichkeitssuche lassen sich zu einem Dokument somit thematisch ähnliche Dokumente finden.

Die *semantische Suche* geht von einem oder mehreren Suchstichwörtern aus. Diese werden ähnlich wie die Dokumente selbst textuell und semantisch analysiert; anschließend wird ein virtueller Fingerabdruck erzeugt. Dieser gestattet die Suche nach relevanten Dokumenten durch den Recommender, es werden also die zu den Stichwörtern thematisch passenden Dokumente gefunden.

Alle semantischen Suchprozesse verfügen über die Möglichkeit der *semantischen Suchraumerweiterung*. Grundlage dieser semantischen Verfahren ist die aus der gemeinsam erstellten Vokabularsammlung erzeugte Ontologie, die die Begriffe („Konzepte“), deren Synonyme (z.B. Zauneidechse – L. agilis), Hierarchiebeziehungen (Zauneidechse – Reptilien) und Beziehungen (Zauneidechse – lebt in Habitat) enthält. Die Konzepte aus dem Suchdokument oder die Suchstichwörter können so in mehreren Schritten um weitere Suchkonzepte ergänzt werden. Ein Beispiel wäre die Erweiterung der Suche von einer Tierart auf nah verwandte Tierarten oder von einer konkreten Maßnahme auf ähnliche Maßnahmen bzw. allgemeinere Maßnahmen (sog. *Fuzzy-Suche*) Diese Suchraumerweiterung kann für spezielle Teilbereiche der Wissensbasis einzeln aus- oder abgewählt werden, bzw. in ihrer Stärke konfiguriert werden.

Für die semantischen Suchen kommt bei der Wissensplattform die Condat Smart Media Engine (SME) zum Einsatz.

6. Benutzeroberfläche

Die Anforderungen an die Benutzeroberfläche wurden im ersten Workshop initial definiert (vgl. 2.2.) und dann anhand eines ersten Prototyps, der exemplarisch verschiedene Such- und Navigationsverfahren demonstrierte, gemeinsam mit potentiellen Anwendern in einem zweiten Workshop (vgl. 2.3.) präzisiert. Hieraus wurde dann die finale Version der Benutzeroberfläche entwickelt, die im Folgenden beschrieben wird.

Abbildung 3 zeigt die Startansicht der Benutzeroberfläche nach Eingabe eines Suchbegriffs. Im Folgenden werden die einzelnen Funktionen jeweils gesondert erläutert.

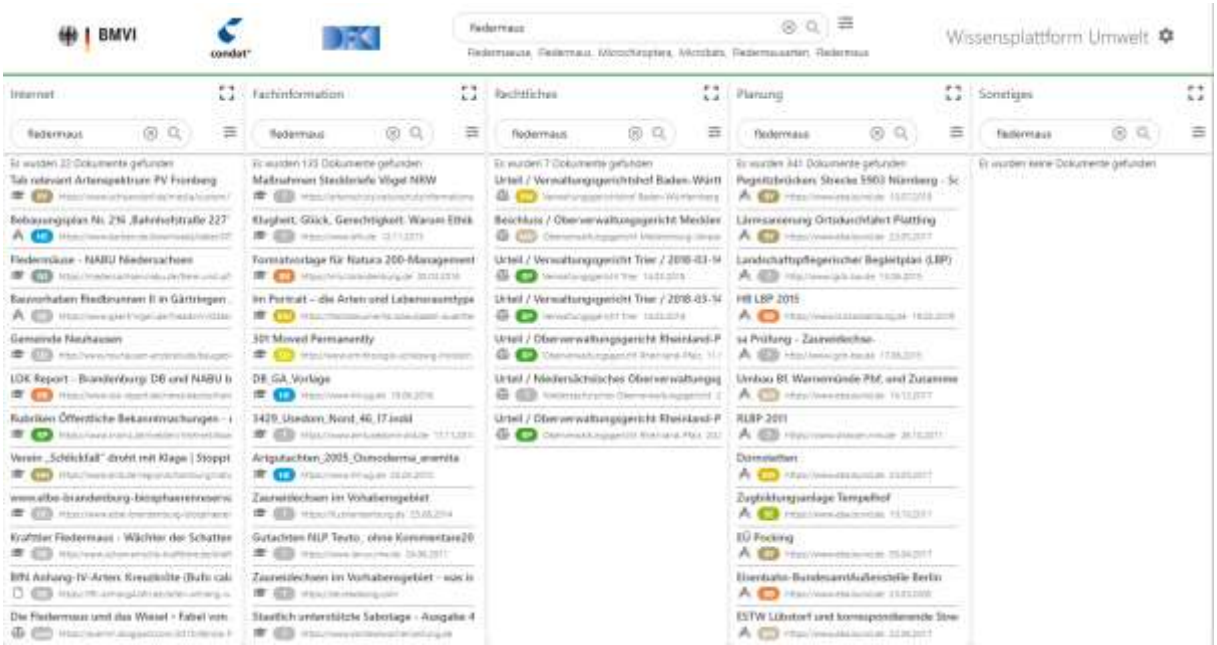

Abbildung 3: Benutzeroberfläche

Zentrales Suchfeld

Die Benutzeroberfläche bietet für bestimmte Kategorien (Fachinformationen, Rechtliches, Planung, etc.) spezialisierte Suchen im Internet sowie in der Datenbank, in der die identifizierten Dokumente abgelegt wurden. Es können einzelne Dokumentarten gesucht werden. Über das zentrale Suchfeld im oberen Teil wird die Suche in allen Kategorien/spezialisierten Suchen aufgerufen, und deren Ergebnisse werden sukzessive angezeigt.

Unterhalb des zentralen Suchfeldes wird der semantische Fingerabdruck des Suchbegriffes angezeigt. Dieser wird an die Internetsuche übergeben, um eine erweiterte

Suchanfrage mit allen Suchbegriffen stellen zu können.

Anzeige Suchergebnisse

Die Suchergebnisse werden in fünf separaten Spalten angezeigt, welche jeweils eine Kategorie abbilden:

Internet
Suchergebnisse über die Bing-Suche; hierbei lässt sich umschalten zwischen der flachen Anzeige der Suchergebnisse und der explorativen Sicht mit der nach Themen geclusterten Sicht.

Fachinformation
Dokumente, die der Kategorie „Fachinformation“ zugeordnet wurden.

Rechtliches
Dokumente, die der Kategorie „Rechtliches“ zugeordnet wurden. Hierbei handelt es sich sowohl um Dokumente, die direkt aus openlegaldata.io importiert wurden, als auch um Dokumente, die über die allgemeine Internet-Suche gefunden wurden.

Planung
Dokumente, die der Kategorie „Planung“ zugeordnet wurden.

Sonstiges
Dokumente, die keiner der drei Kategorien zugeordnet werden konnten.

Diese Aufteilung hat demonstrativen Charakter. Prinzipiell lassen sich durch den semantischen Fingerabdruck beliebige Kategorien erstellen und in Verbindung mit persönlichen Profilen auch personalisieren.

Z.B. wäre es mit dieser Technik möglich, eine individuelle Suchspalte „Planungsbezogene Dokumente zu Vorhaben im Bereich Schiene der Regionen Berlin, Brandenburg und Mecklenburg-Vorpommern mit Veröffentlichungsdatum ab 2016“ zu konfigurieren, oder „Fachartikel zu CEF-Maßnahmen bei der Zauneidechse“. Eine solche Detaillierung wäre aber nur in Verbindung mit einer Personalisierung sinnvoll, da dann jeder Nutzende eigene Suchprofile anlegen und anpassen kann.

Bei allen Suchergebnissen werden folgende Informationen angezeigt:

- Name (Titel);
- Dokumenttyp (Fachinformation, Rechtlich, Planung, Sonstiges), angezeigt durch ein Icon;
- Regionale Zuordnung, angezeigt durch kleine Labels mit Kürzeln der Bundesländer bzw. Länder;
- Domain, aus der das Dokument stammt.

Explorative Suche / Clustering

In der ersten Suchspalte werden unter „Internet“ die Suchergebnisse aus dem Internet angezeigt. Diese können mittels einer Schaltfläche „Ergebnis gruppieren“ in thematische Gruppen unterteilt werden. Hierbei wird jeweils die Anzahl der Ergebnisse mitangezeigt.

Abbildung 4: Explorative Sicht der Suchergebnisse

Sortierung und Filterung der Suchergebnisse

Die Suchergebnisse lassen sich jeweils nach Relevanz, Titel oder Datum sortieren und nach Bundesländern bzw. Ländern filtern.

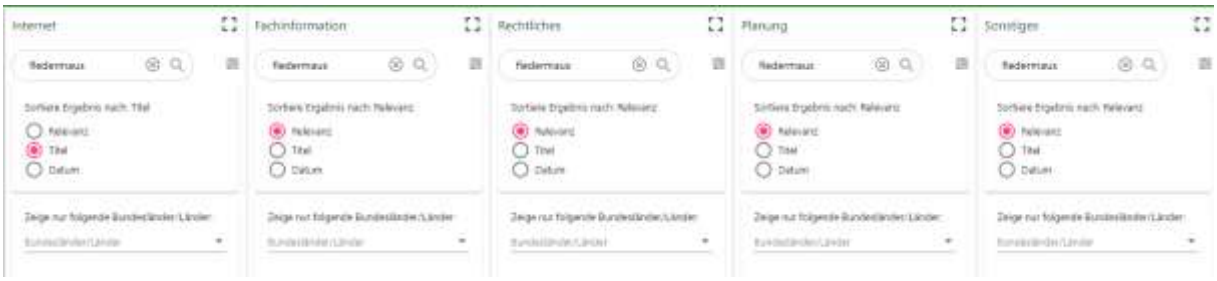

Abbildung 5: Sortierung und Filterung der Suchergebnisse

Konfigurationseinstellungen

Über die Benutzeroberfläche lassen sich auch verschiedene Eigenschaften bzgl. der Sucheinstellungen sowie der Anzeige der Suchergebnisse festlegen (Abbildung 6). Zum einen lässt sich einstellen, wie eng oder weit die automatische Suchraumerweiterung greifen soll. Hierüber lässt sich das von den Anwendern gewünschte „Sanduhrprinzip“ steuern, d.h. ob nur ganz exakte Suchtreffer gewünscht sind, oder auch ähnliche bzw. übergreifende Treffer. So lässt sich z.B. anhand des Parameters „Tierart“ einstellen, ob bei der Suche nach „Zauneidechse“, nur exakt nach Dokumenten gesucht werden soll, die „Zauneidechse“ (inkl. semantischer Synonyme wie „L. agilis“ und „sand lizard“) beinhalten, oder auch verwandte Tierarten (alle in der Vokabularsammlung enthaltenen Reptilien) oder alle Tierarten.

Abbildung 6: Konfigurationsparameter

Zum anderen lassen sich die anzuzeigenden Suchkarten aus- bzw. abwählen, wodurch die verbleibenden Suchkarten grösser und damit informativer werden.

Detailansicht

Wenn man auf ein einzelnes Suchergebnis klickt, öffnet sich die Detailansicht.

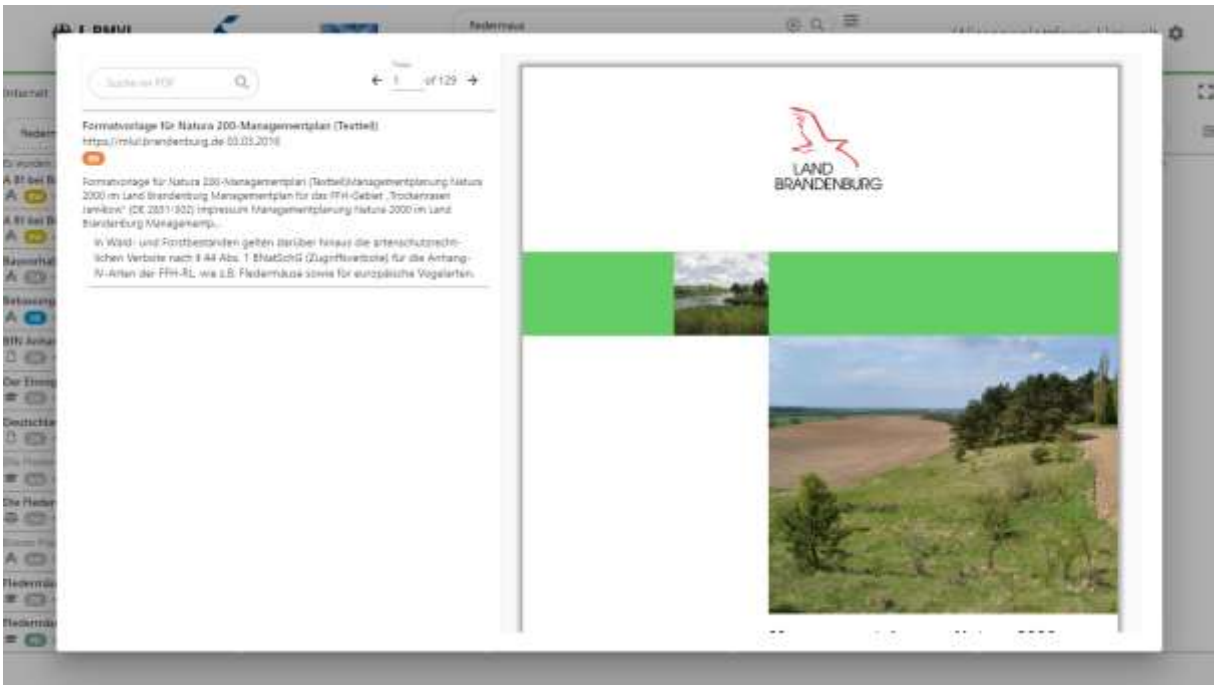

Abbildung 7: Detailansicht

Die Detailansicht ist zweigeteilt: Auf der rechten Seite befindet sich das eigentliche Dokument bzw. die HTML-Seite, auf der linken Seite werden die Fragmente angezeigt, die durch den Suchbegriff als relevant identifiziert wurden. Hiermit soll ein schnelles Navigieren zu den jeweiligen Stellen ermöglicht werden, ohne dass der Anwender das komplette Dokument Seite für Seite durchblättern muss. Auch ermöglicht die Anzeige des Fragments eine schnelle Übersicht über den Kontext, in welchem das Suchwort im Dokument auftaucht. Über das Suchfeld lassen sich darüber hinaus auch Schlagworte im Dokument suchen.

Allgemeine Anmerkungen zur Benutzeroberfläche

Neben den in den Workshops identifizierten Funktionalitäten soll die realisierte Benutzeroberfläche insbesondere Konfigurations- und Einstellungsmöglichkeiten demonstrieren. Der derzeitige Demonstrator ist nicht personalisiert. Für die Benutzeroberfläche einer eventuellen späteren Weiterentwicklung der Plattform, wäre eine Personalisierung empfehlenswert, über die spezielle definierte Suchkarten oder Do-

kumentlisten angelegt und gepflegt werden könnten.

Von den in den Workshops identifizierten Funktionalitäten wurden im Rahmen der Machbarkeitsstudie nicht realisiert: a) die Nutzerrückkopplung (Feedback-Mechanismus), b) die vorhabenbezogene Sicht. Hierzu die folgenden Erläuterungen:

Nutzerrückkopplung (Feedback-Mechanismus)

Die Nutzerrückkopplung sollte der Bewertung der Relevanz von Suchergebnissen durch die Anwender dienen. Diese wurde im zweiten Workshop von den Anwendern nicht mehr als wichtige Funktion identifiziert. Unabhängig davon ist diese Funktionalität aber sinnvoll nur in Verbindung mit einer Personalisierung zu realisieren, da Relevanzbewertungen meistens vom Kontext der Suche bzw. dem jeweils persönlichen Kenntnisstand und Bedarf abhängig sind. Ein generelles, unpersonalisiertes Feedback macht nur Sinn bei Suchergebnissen, die in keinem Kontext relevant sind (wie z.B. die Anzeige der Operette „Die Fledermaus“ bei Suchen nach Fledermaus). Ansonsten müsste noch genauer untersucht werden, ob eine einfache ja/nein-Antwort möglich sein sollte, oder eine graduelle z.B. von „irrelevant“, über „relevant“ bis „sehr relevant“ und dies jeweils in Abhängigkeit von Suchanfrage und Nutzerin bzw. Nutzer.

Vorhabenbezogene Sicht

Die vorhabenbezogene Sicht war ein Wunsch, der im zweiten Workshop geäußert wurde (vgl. 2.3.8). Nach Vorstellung der Nutzenden sollte diese Sicht alle Dokumente und Informationen zu einem Vorhaben zusammenfassen und in einer zeitlichen Reihenfolge darstellen, sodass Planungsdokumente, Gutachten, Klagen, Gerichtsurteile (Genehmigung oder Ablehnung) zu einem Vorhaben auf einen Blick sichtbar sind. Die insoweit angedachte Darstellung ist äußerst sinnvoll, jedoch sehr aufwändig zu realisieren. Voraussetzung für die Realisierung einer solchen Funktion ist die Erstellung eines Knowledge Graph, der Dokumente und Informationen miteinander verknüpft und spezielle Abfragemöglichkeiten bereitstellt. In der im Rahmen der Machbarkeitsstudie erstellten Wissensplattform wurde mit der Verwendung der Condat SME zwar die Grundvoraussetzung hierfür geschaffen, die jedoch für die Umsetzung der vorhabenbezogenen Sicht noch einer entsprechenden Erweiterung bedürfen würde. Es wird jedoch empfohlen, diese Funktion in die Anforderungen an das Produktivsystem mit aufzunehmen.

7. Komponenten und Dienste

7.1 Importer (Crawling / Scraping)

7.1.1 Prozessbeschreibung

Die gezielte Einbindung von Dokumenten aus einer bekannten Quelle erfordert drei grundlegende Prozessschritte:

1. Analyse des Quell-Portals (vgl. 7.1.2);

2. Implementierung eines quell-spezifischen Scrapers (vgl. 7.1.3) und

3. Aufbereitung und Indizierung gefundener Dokumente (vgl. 7.1.4).

7.1.2 Analyse des Quell-Portals

Die Portale der verschiedenen Dokumentquellen sind sehr heterogen. Jede Quelle hat ihre eigene Struktur, spezifische Such- und Zugriffsmöglichkeiten auf bereitgestellte Dokumente und teilweise eine spezielle API, die Dokumente nicht nur als PDF oder HTML, sondern auch als JSON bereitstellt.

Im ersten Schritt war deshalb eine detaillierte Analyse der Eigenschaften und Besonderheiten der zu integrierenden Dokumentquelle vorzunehmen.

Viele Quellen bieten eigene Suchmasken zum Suchen nach Stichwörtern und fachlichen Kriterien sowie zum Anzeigen und Durchblättern von Trefferlisten. Diese Trefferlisten erwiesen sich als der beste Ansatz zur automatischen Dokumentextraktion.

Andere Quellen bieten vordefinierte statische Dokument- oder Veröffentlichungslisten. Oft sind die Dokumentverweise hier direkt aus den Listen ermittelbar, manchmal ist aber auch jeweils noch ein separater Navigationsschritt auf eine zugehörige Detailmaske erforderlich.

Aus den Dokumentlisten oder Detailmasken in den Portalen lassen sich oft auch zusätzliche Metadaten auslesen, wie z.B. Autoren, Titel, Beschreibung oder Veröffentlichungsdatum – in jeweils sehr spezifischer Strukturierung und Zusammensetzung.

Die Analyse eines Quell-Portals erfolgte im Rahmen der Machbarkeitsstudie manuell, wobei gezielt untersucht wurde, an welchen Stellen und auf welche Art und Weise in dem Portal relevante Dokumente veröffentlicht und z.B. als PDF zum Download an-

geboten werden. Erste Ideen einer Automatisierung dieses Prozesses mit Hilfe generischer Web-Crawler wurden recht schnell verworfen, da der Aufwand für die quell-spezifische Konfiguration und Steuerung des Crawlers sowie die Bewertung der Crawl-Ergebnisse den Nutzen der einmaligen gezielten manuellen Suche überstiegen hätte.

7.1.3 Implementierung eines quell-spezifischen Scrapers

Generischer vs. spezifischer Scraper

Aufgrund der Heterogenität der Quell-Portale erschien die Entwicklung eines im Voraus geplanten, für möglichst viele oder alle Quellen funktionierenden generischen und damit sehr komplexen Dokument-Extraktors (Scrapers) aussichtslos. Jede neu zu integrierende Quelle wies Besonderheiten auf, die bis dahin noch nicht berücksichtigt werden konnten.

Ein erfolgversprechenderer Weg ist dagegen die sukzessive Entwicklung einzelner quell-spezifischer Scraper, wobei in die Entwicklung jedes Scrapers mindestens die bisherigen Erfahrungen all seiner Vorgänger einfließen und sich nach und nach auch einzelne universell einsetzbare Teilkomponenten herauskristallisieren werden. Dadurch verringert sich der jeweilige Entwicklungsaufwand für die Einbindung neuer Quellen mit jeder weiteren Einbindung. Dies wurde im Rahmen der Machbarkeitsstudie auch so realisiert. Für eine eventuelle Weiterentwicklung bedeutet dies, dass die Anbindung weiterer Datenquellen zwar weiterhin weitgehend individuell erfolgen müsste, aber der jeweilige Aufwand deutlich geringer wäre.

Funktionsweise des Scrapers

Wenn nach der Analyse eines Quell-Portals die Einstiegspunkte für die Extraktion veröffentlichter Dokumente bekannt sind, kann man hierfür einen spezifischen Scraper entwickeln. Dies wurde für die im Anhang (vgl. Anhang B) aufgeführten Quellen durchgeführt.

Werden statische Dokumentlisten zu verschiedenen Themen oder nach anderen Kriterien wie Zeit oder Ort angeboten, so beschränkt sich das „Aufsammeln“ der Dokumentverweise und eventueller Metadaten auf das Auslesen der jeweiligen HTML-Strukturen dieser Listen. Wenn die Dokumentverweise nicht direkt angeboten werden, so ist in einigen Fällen pro Dokument noch ein weiterer Request auf eine Detailseite notwendig.

Wenn die Dokumente in dynamischen Trefferlisten über Suchformulare angeboten werden, so muss der Scraper automatisiert die Nutzung des Suchformulars simulie-

ren, bei großen Treffermengen ein automatisches sog. Paging (Durchblättern durch die Suchergebnisse) durchführen und die Trefferlisten dann ähnlich den statischen Listen auswerten. Um über die Suchformulare fachlich relevante Dokumente zu erhalten, wird vom Scraper sukzessive nach einer Liste typischer Schlüsselwörter gesucht. Duplikate bei den Treffern werden dabei herausgefiltert.

Zur Analyse der HTML-Seiten wird das Java-basierte Werkzeug „JSoup“[19] verwendet, das eine Navigation im HTML-Baum, eine gezielte Selektion von Elementen mittels CSS-Selektoren und eine Text-Extraktion bietet. JSoup stößt allerdings an seine Grenzen, wenn HTML-Seiten dynamische Bestandteile enthalten, deren Inhalt erst während der Anzeige im Browser erzeugt wird. Hier musste für das Scraping zusätzlich eine Selenium-Bibliothek[20] verwendet werden, die einen Browser automatisiert fernsteuert und dadurch auch die dynamischen Inhalte von HTML-Seiten zugreifbar macht.

Als Ergebnis des Scrapings werden die Dokumentverweise und eventuell zusätzlich ermittelte Metadaten, wie Autor, Titel oder Veröffentlichungsdatum in der Metadaten-Datenbank gespeichert. Weiterhin werden administrative Daten, wie der Zeitpunkt des Scrapings und das Quell-Portal gespeichert, die für den Prozess der späteren Aktualisierung der Datenbasis und Wiederholung des Scrapings erforderlich sind.

7.2 Internet-Suchkomponente

Während der in Abschnitt 7.1 beschriebene Importer der Anbindung bekannter und als relevant identifizierter Quellen dient, dient die Internet-Suchkomponente der Suche im Internet nach unbekannten Quellen, sowohl durch die manuelle Suche über die Benutzeroberfläche, als auch automatisch regelmäßig durchgeführte Suchen. Hierfür wurde eine Internet-Suchkomponente entwickelt und integriert. Technisch wurde hierbei im Rahmen der Machbarkeitsstudie das *Search API* von Microsoft Bing[21] eingebunden. Prinzipiell kann aber z.B. auch die Google Search API hierfür

[19] JSoup ist eine Open-Source-Bibliothek zur Extraktion von Text aus Websites basierend auf der Programmiersprache Java; https://jsoup.org/

[20] Selenium ist ein Framework für automatisierte Softwaretests von Webanwendungen, https://www.selenium.dev/. In diesem Zusammenhang wurde es aber nicht für Tests verwendet, sondern um Inhalte von Websites zu scrapen, die interaktive Benutzereingaben erforderten, wie z.B. einen Suchbegriff in ein Suchformular einzugeben und einen Button zu drücken.

[21] Das Bing API ist Bestandteil der von Microsoft angebotenen Cloud-Plattform Azure Cloud Services; https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/

genutzt werden. Diese unterscheiden sich im Wesentlichen durch die Lizenzbedingungen.

Suchanfragen an die Search APIs von Microsoft und Google triggern keine eigene Suche, sondern fragen den von Microsoft bzw. Google bereits gepflegten Index ab, d.h. sie entsprechen einer manuellen Suchabfrage über das Internet. Das bedeutet insbesondere auch, dass die Aktualität dem letzten jeweiligen Indexierungslauf entspricht.

Alternativ wäre es bei einer Weiterentwicklung des Systems auch möglich, eine eigene Suchmaschine zu entwickeln und einzusetzen, z.B. auf Basis von Apache Nutch[22], wobei dies sehr aufwändig wäre und zu bezweifeln ist, dass hierdurch ein besseres Ergebnis zu erzielen wäre als mithilfe von Bing oder Google Search. Diesen Weg sollte man nur dann weiterverfolgen, wenn inhaltliche oder regulatorische Gründe gegen eine Verwendung verfügbarer Suchdienste sprechen.

Abgespeichert werden in der Wissensplattform nur diejenigen Suchergebnisse, die nicht bereits über das Crawlen bereits bekannter Sites gefunden wurden.

7.3 Relevanzfilter

Die Bewertung der Relevanz eines Dokuments erfolgt auf Basis der gemeinsam mit dem BMVI erstellten Ontologie (Vokabularsammlung) (s. 7.9.1). Als Kriterium gilt im Wesentlichen die Anzahl der enthaltenen Begriffe aus der Ontologie im Verhältnis zur Länge des Textes. Die Relevanzbewertung findet im Demonstrator nur im Zuge der manuellen Internetsuche statt, da davon ausgegangen werden kann, dass alle Dokumente, die durch das Crawlen bekannter Quellen und durch die regelmäßige automatische Suche gefunden werden, per se relevant sind. Diese Annahme wäre im Zuge eines längeren Benutzertests bzw. des laufenden Betriebes noch zu überprüfen. Da zum Zeitpunkt der Relevanzbewertung nur der kurze Textausschnitt („Snippet“) der Internet-(Bing-)-Suche zur Verfügung steht und nicht der komplette Text, erfolgt die Bewertung nur anhand dieses Textausschnitts.

Die Relevanzbewertung erfolgt im Demonstrator immer allgemein, d.h. nicht nutzerbezogen oder kontextbezogen.

[22] Nutch ist ein Java-Framework für Internet-Suchmaschinen; https://nutch.apache.org/

Die hier zum Einsatz kommende Relevanzberechnung ist sehr einfach und linear, da sie lediglich die Anzahl der auftauchenden Begriffe bezogen auf die BMVI-Vokabularsammlung in Betracht zieht; sie stellt damit nur eine erste Näherung dar. Zum einen setzt dieses Verfahren voraus, dass die relevanten Begriffe möglichst vollständig in der Ontologie enthalten sind, da ansonsten relevante Dokumente ggf. nicht als solche erkannt werden. Das bedeutet, dass für den späteren Produktivbetrieb eine Initialisierungsphase einzuplanen ist. In dieser wäre manuell in Zusammenarbeit mit den Nutzenden zu überprüfen, welche Dokumente aufgrund fehlender Begriffe nicht als relevant erkannt werden, und welche Begriffe in die Ontologie aufgenommen werden müssen. Dies sollte in regelmäßigen Abständen auch während des Betriebs erfolgen.

Zum anderen könnte eine qualitative Verbesserung dadurch erzielt werden, dass die Begriffe nicht gleich gewertet werden, sondern ein Begriff wie „CEF-Maßnahme“ z.B. höher und ein allgemeiner Begriff wie „Gerichtsurteil“ niedriger. Dies kann aber nur in Zusammenarbeit mit den späteren Nutzerinnen und Nutzern geschehen.

Eine weitere mögliche Verbesserung des Relevanzfilters bestünde darin, durch die Pflege von Negativlisten (sog. Blacklists) bestimmte Quellen von vorneherein auszuschließen (z.B. alle Ergebnisse aus Youtube), oder auch eine Negativklassifizierung gegen Begriffe aus Wikipedia: Während die im Rahmen des Demonstrators angewandte Positivklassifizierung feststellt, wie relevant ein Dokument ist, könnte die Negativklassifizierung die Irrelevanz berechnen. Wenn z.B. als Suchbegriff „Fledermaus“ eingegeben wird, stellt die Positivklassifizierung fest, wie viele weitere Begriffe aus der Ontologie (s. 7.9.1) das Dokument explizit oder implizit enthält. Sofern hieraus kein weiterer Begriff auftaucht, ist das Dokument höchstwahrscheinlich irrelevant. Die Negativklassifizierung hingegen würde z.B. feststellen, dass das Dokument noch die Begriffe „Johann Strauß“ und „Operette“ o.ä. enthält, und damit das Dokument in die Kategorie „Unterhaltung“ einordnen. Da man alle diese Negativbegriffe auf keinen Fall selbst pflegen würde, müsste diese Negativklassifizierung aber anhand einer vorhandenen Ontologie erfolgen, vorzugsweise Wikipedia bzw. Wikidata.

7.4 Regionalerkennung

Hinweis: im Folgenden werden die Begriffe „Regionalerkennung“ und „Bundeslanderkennung“ verwendet. Die Bundeslanderkennung ist hierbei ein KI-basiertes Verfahren, das aus Ortsnamen in Texten abzuleiten versucht, welchem Bundesland der Text zugeordnet werden kann. Die Regionalerkennung ist das allgemeinere Verfahren, Bundesländer über die Bundeslanderkennung abzuleiten, darüber hinaus aber noch über einfache URL-Adressenanalyse Länder wie Schweiz, Österreich, etc.

7.4.1 Technisches Verfahren

Aus Wikipedia wurde eine Liste aller deutschen Städte und ihrer Bundesländer heruntergeladen. In den Texten wird nach diesen Bundesländern standardmäßig gesucht und dann eine Verteilung über die unterschiedlichen Bundesländer berechnet. Wenn mehr als 70% der erkannten Ortsnamen aus einem Bundesland kommen, so wird der Text diesem Bundesland zugeordnet. Andernfalls wird der Text keinem Bundesland zugeordnet.

Für die Evaluation wurden von der Homepage des Eisenbahn-Bundesamts 16 Testdokumente heruntergeladen, die jeweils einem Bundesland zugeordnet wurden. Darüber hinaus wurden 4vier Dokumente ohne Bezug zu einem Bundesland heruntergeladen.

7.4.2 Realisierung

Die Bundeslanderkennung zeigt i.d.R. hervorragende Ergebnisse. Von den 20 Testdokumenten wurden 19 korrekt zugeordnet. Micro Precision, Micro Recall und Micro F2 Score[23] liegen bei jeweils 0,95. Die folgende Abbildung zeigt die dazugehörige Konfusionsmatrix.

[23] Micro Precision und Micro Recall sind Metriken, die angeben, ob ein Algorithmus die klassifizierten Elemente tatsächlich richtig klassifiziert (sog. „True Positives“) und wie viele Elemente er als richtig klassifiziert („Recall“). F2 beschreibt das Verhältnis zwischen beiden.

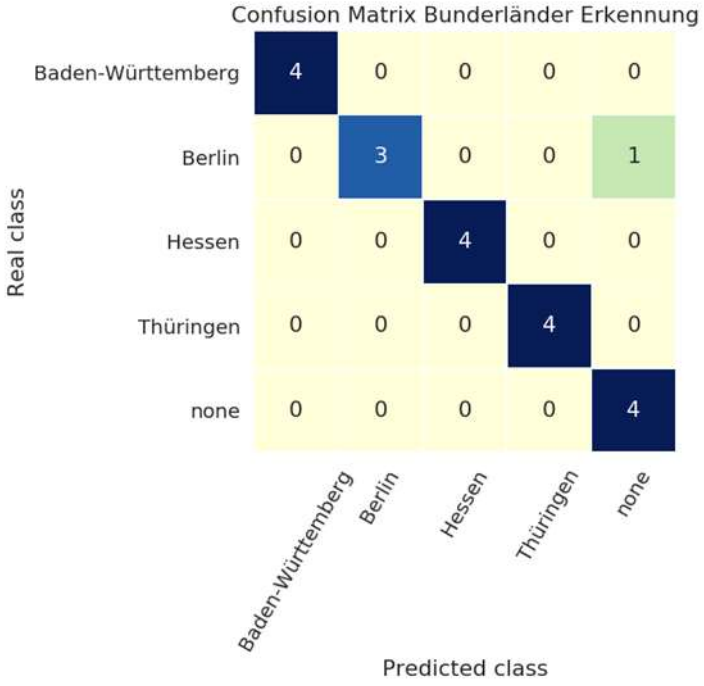

Abbildung 8: Konfusionsmatrix der Bundeslanderkennung

Wenn es einen Ortsnamen allerdings mehrfach gibt (z.B. Frankfurt), und der Ortsname dann ohne Ergänzung (Frankfurt (Main), Frankfurt (Oder)) mehrdeutig ist, und der Text aber im Übrigen auch keine sonstigen Ortsnamen enthält, ergibt die automatische Bundeslanderkennung kein Resultat.

Technisch ist das Verfahren leicht umzusetzen und wäre auch auf weitere Länder wie z.B. Österreich oder Schweiz übertragbar. Das Verfahren kann z.B. auch auf Seen, Flüsse, Berge, FFH-Gebiete und Weiteres erweitert werden. Da im Rahmen der Machbarkeitsstudie für außerdeutsche Dokumente aber nur die generelle Zuordnung zum Land erforderlich ist, wurde diese hier durch die einfache Analyse der URL realisiert, d.h. z.B. werden URLs mit .ch-Endung als CH klassifiziert.

Des Weiteren wird zur Regionalzuordnung neben einem Text (wie z.B. dem Snippet aus dem Bing-Resultat) auch die URL herangezogen, da in einem solchen kurzen Text oft kein Ortsname auftaucht, die zugehörige Gemeinde aber aus der URL ableitbar ist (z.B. *eberswalde.de*).

7.5 Dokumenttyperkennung

7.5.1 Erläuterung der Funktion

Bei den Anforderungsworkshops (vgl. 2.2. und 2.3.) wurden drei Kategorien von Do-

kumenten als besonders wichtig dargestellt:

- Fachinformation,
- Planung
- Rechtliches

Mithilfe intelligenter Technologien werden Dokumente automatisch mit einer dieser drei Kategorien, oder mit der zusätzlichen Kategorie „Sonstiges", markiert.

Dies erleichtert die Recherchearbeit auf mehrere Arten:

- Zum einen kann diese *Kategorisierung* in der Liste der Ergebnisse der Suchfunktion *visualisiert* werden. Damit kann sich der oder die Suchende einen schnelleren Überblick über den Inhalt der Dokumente verschaffen, ohne diese jeweils zuerst öffnen zu müssen. Wird z.B. nach wissenschaftlichen Fachinformationen gesucht, so müssen die Planungsdokumente nicht angeklickt werden.
- Zum anderen kann die *Kategorisierung als Filter* genutzt werden. Es kann gezielt z.B. lediglich die Kategorie "Rechtliche Dokumente" durchsucht werden. Dadurch werden zahlreiche, für die konkrete Suche irrelevante Suchergebnisse gar nicht erst angezeigt und hierdurch die Recherchearbeit weiter beschleunigt.

7.5.2 Technisches Verfahren

Für die Dokumenttyperkennung wurde ein datengestütztes Verfahren eingesetzt. Zunächst wurden zu allen vier Kategorien (*Rechtliches*, *Planung, Fachinformation* und *Sonstiges*) jeweils zehn Dokumente gesammelt. Die Dokumente stammten vom Eisenbahn-Bundesamt (*Rechtliches*, *Planung)*, ResearchGate (*Fachinformation*) und dem weiteren, unspezifischen Internet (*Sonstiges*). Die Dokumente wurden in zwei Hälften geteilt, eine Hälfte für Training und eine für die Evaluation. Dann wurde der Machine Learning Algorithmus *Multinomial Naive Bayes*[24] auf diesen Daten trainiert und evaluiert. *Multinomial Naive Bayes* erstellte Tabellen mit Wörtern, die einen Hinweis auf die Klassen ergeben. So war z.B. das Wort "Gerichtsurteil" ein starker Hinweis darauf, dass es sich um ein rechtliches Dokument handelt. Der Algorithmus konnte anhand dieser Tabellen für ein einzelnes Dokument Wahrscheinlichkeits-

[24] Für weitergehende Informationen zum Multinomial Naive Bayes Verfahren vgl. z.B. https://en.wikipedia.org/wiki/Naive_Bayes_classifier

verteilungen über die vier Kategorien errechnen. Wenn der Algorithmus mit hoher Wahrscheinlichkeit (> 75%) eine Kategorie vorschlug, so wurde das Dokument dieser Kategorie zugeordnet. Ansonsten wurde es der Kategorie „Andere“ zugeordnet.

7.5.3 Realisierung

Die Dokumenttyperkennung war das Schwierigste der drei KI-basierten Verfahren zu Regionalzuordnung, Clustering und Dokumenttyperkennung. Datengetriebene Verfahren benötigen große Mengen von Daten, die in dem Bereich Umweltschutz noch nicht existieren. Die hier berechnete Evaluation des technischen Verfahrens zeigt, dass 18 von 20 Testdokumenten korrekt erkannt wurden. Auch wenn diese Evaluation einen ersten Eindruck der Leistungsfähigkeit des Systems bietet, wird künftig eine größere Trainingsbasis erforderlich sein, um die Ergebnisse aus den Laborbedingungen auf die echte Welt zu übertragen.

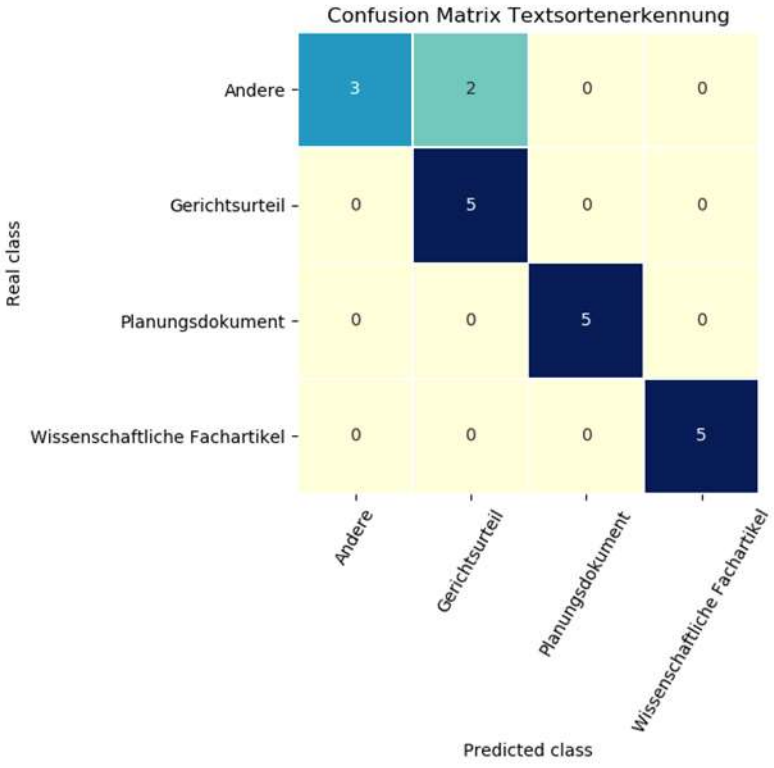

Abbildung 9: Konfusionsmatrix zur Dokumenttyp-Erkennung

Dokumenttypen können auch über die Datenquelle erkannt werden. So haben Dokumente von Juris z.B. sehr sicher einen rechtlichen Bezug, Dokumente aus einer wissenschaftlichen Suchmaschine sind sehr wahrscheinlich wissenschaftliche Fach-

artikel. Wenn bei der späteren Umsetzung dieses Projekts entsprechende Datenquellen angebunden und große Mengen von Dokumenten gesammelt wurden, kann die Textsortenerkennung mit einer wesentlich größeren Menge von Dokumenten trainiert werden.

Folgende Schwierigkeiten ergaben sich in der Aufgabe der Textsortenerkennung:

1. Eine klare Definition der Textsorten war schwierig. Wissenschaftliche Fachartikel zeichnen sich vor allem dadurch aus, dass sie einen wissenschaftlichen Peer Review Prozess durchlaufen haben, d.h. ein wissenschaftlicher Text, der in einer wissenschaftlichen Zeitschrift oder in den Proceedings einer wissenschaftlichen Konferenz veröffentlicht wird, wird zuvor von anderen Wissenschaftlern begutachtet, die den Text nach wissenschaftliche Relevanz und anderen Kriterien beurteilen und ggf. die Veröffentlichung verhindern. Dieser Prozess ist dem Text selbst jedoch nur schwer anzusehen. Sie werden in wissenschaftlichen Zeitschriften oder auf entsprechenden Konferenzen veröffentlicht, jedoch gibt es hier zahlreiche Grenzfälle, die unter einem pseudowissenschaftlichen Deckmantel Falschinformationen verbreiten.
2. Die Definition eines rechtlichen Dokuments war ebenfalls schwierig. Reicht die Nennung eines einzelnen Gesetzes aus, um ein Dokument als rechtliches Dokument einzustufen? Hier muss noch eine Definition ausgearbeitet werden, die einerseits die Anforderungen der Anwendenden und andererseits die verfügbaren Trainingsdaten miteinbezieht.
3. Während der Arbeit wurde deutlich, dass die meisten Dokumente mehreren Kategorien zugeordnet werden. So haben Planungsdokumente oft einen starken rechtlichen Bezug. Wissenschaftliche Fachartikel beziehen sich sowohl auf Planungsverfahren als auch auf rechtliche Fragestellungen. Somit sollte eruiert werden, ob es für Anwendende hilfreich ist, wenn Dokumente mit mehreren Kategorien markiert werden.[25]
4. Oft folgen Planungsdokumente aus dem Umweltbereich einer wissenschaftlichen Form, z.B. mit Quellenangaben und Literaturverzeichnis und sind somit nur schwer von wissenschaftlichen Fachartikeln zu unterscheiden.

[25] Wie in der Abschlusspräsentation am 12.02.2020 deutlich wurde, wird eine mehrfache Zuordnung von Dokumenten zu Kategorien von den potentiellen Nutzern explizit gewünscht und sollte deshalb auch für die eventuelle Weiterentwicklung als Anforderung mit aufgenommen werden.

7.6 Explorative Suche/Clustering[26]

7.6.1 Erläuterung der Funktion des Clusterings

Eine Internetsuche nach z.B. „Fledermaus“ liefert zahlreiche Dokumente zu den unterschiedlichsten Themen. Neben Informationen zu CEF-Maßnahmen für Fledermäuse werden allgemeine Informationen zu den Tieren gezeigt. Dazu gibt es beispielsweise auch Webseiten über die Operette von Johann Strauß und den Comic-Helden Batman. Anwendende im Kontext dieses Projekts interessieren sich zumeist nur für CEF-Maßnahmen, nicht aber für die Operette. Mit Hilfe der *explorativen Suche* werden automatisch Themen aus den Suchergebnissen gebildet, die einzelnen Suchergebnisse werden diesen Themen zugeordnet. Benutzende können sich schnell einen Überblick über die einzelnen Themen verschaffen und einzelne Themen aus- oder abwählen.

7.6.2 Technisches Verfahren

Die Themenerkennung beruht auf sog. Clustering-Algorithmen.[27] Diese sind auf Basis des Open Source Tools *Carrot2*[28] implementiert worden.

Carrot2 erkannte allerdings zahlreiche semantisch ähnliche Cluster, die lediglich auf syntaktischer Ebene unterschiedlich sind, z.B. “CEF Maßnahme Fledermaus”, “CEF-Maßnahmen Fledermaus” und “CEF Maßnahmen Fledermaus”. Alle drei Cluster können und sollten allerdings zu einem Cluster zusammengeführt werden. Um dies zu realisieren, wurde ein eigenständiges linguistisches Verfahren implementiert, das *semantisch ähnliche Clustertitel erkennen* und damit die Qualität des Clusterings verbessern kann. Das Verfahren besteht aus den folgenden Arbeitsschritten:

1. In der *Tokenisierung* werden die Clustertitel in einzelne Wörter aufgeteilt.

[26] Die Begriffe *explorative Suche* und *Clustering* werden im Kontext dieses Abschlussberichts austauschbar verwendet. Genauer betrachtet bedeutet die *explorative Suche* die Suchfunktion aus Nutzersicht, während *Clustering* das technische, KI-basierte Verfahren meint, mit dessen Hilfe die Funktionalität der explorativen Suche realisiert wird.

[27] *Unter Clusteranalysen (Clustering-Algorithmen, gelegentlich auch: Ballungsanalyse) versteht man Verfahren zur Entdeckung von Ähnlichkeitsstrukturen in (meist relativ großen) Datenbeständen. Die so gefundenen Gruppen von „ähnlichen“ Objekten werden als Cluster bezeichnet, die Gruppenzuordnung als Clustering.*
Quelle: https://de.wikipedia.org/wiki/Clusteranalyse

[28] Carrot2 ist eine Open Source Clustering Engine basierend auf der Programmiersprache Java; vgl. https://project.carrot2.org/

2. Durch *Stopword Removal* werden häufige, bedeutungsschwache Wörter aus den Titeln entfernt, so dass beispielsweise die Gleichheit von Clustern wie “Fledermaus der” und “Fledermaus” erkannt werden kann. Ein Cluster mit einem automatisch erzeugten, aber offensichtlich unsinnigen Titel wie “Die Keine” werden automatisch “Other Topics” zugeordnet.
3. Durch *Stemming* werden alle Wörter auf ihre Normalformen reduziert, so dass z.B. die Gleichheit von “Maßnahme” und “Maßnahmen” erkannt werden kann.
4. Durch diese syntaktische *Normalisierung* können semantisch ähnliche Cluster erkannt und zusammengeführt werden. Wenn aus zwei Clustern einer entsteht, so bleibt der kürzere Clustertitel erhalten.

7.6.3 Machbarkeit

Die grundsätzliche Machbarkeit und Nützlichkeit des Clustering-Verfahrens im Kontext der Wissensplattform Umwelt wurde demonstriert. Die Qualität der Ergebnisse hängt jedoch stark von der Anzahl der zugrundeliegenden Dokumente bzw. Suchergebnisse ab, d.h. je mehr Dokumente in verschiedene Gruppen aufgeteilt werden, umso besser ist die Qualität der generierten Cluster und auch der Cluster-Titel. Sinnvoll wird das Clustering erst ab einer Anzahl von 400 Dokumenten. Gleichzeitig steigen aber auch der Rechenaufwand und damit auch die Zeit, bis Nutzenden die Ergebnisse angezeigt werden können. Da KI-basierte Verfahren generell sehr rechenintensiv sind, müsste bei einer eventuellen Weiterentwicklung dafür Sorge getragen werden, dass hierfür ausreichende, dedizierte Rechenkapazität zur Verfügung steht.

7.6.4 Ausblick

Die Verbesserung des Clusterings durch Carrot2 wurde für den Demonstrator mittels linguistischer Verfahren umgesetzt. Mittels einer sog. syntaktischen Normalisierung wurden bedeutungsschwache Wörter reduziert, durch die morphologischen Lemmatisierungen wurden die verschiedenen grammatikalischen Schreibweisen von Wörtern auf die Grundform zurückgeführt.

Andere Verfahren zur Erkennung von latenter Semantik, wie beispielsweise *Singular Value Decomposition* (SVD) oder *Latent Semantic Analysis* (LSA), können zukünftig prinzipiell ebenfalls eingesetzt werden. Bei einer eventuellen Weiterentwicklung müsste man untersuchen, ob sich ggf. die Anzahl an gefundenen Ergebnissen erhöht (sog. *Recall*), dies aber ggf. mit einer Abnahme der Genauigkeit der Suchergebnisse erkauft wird (sog. *Precision*).

Mittels *(Contextual) Word Embeddings* können weitere Normalisierungen erreicht werden, die ebenso zu einer Verbesserung der Clusterqualität beitragen können.

Auch hier muss die Entwicklung von Precision und Recall im Auge behalten werden.

7.7 Textextraktion

Nachdem die Dokumente aus den verschiedenen Quellen extrahiert und als Verweise (URLs) mit entsprechenden Metadaten in der Metadaten-Datenbank gespeichert wurden, liegen sie in einer weitgehend homogenen Form vor (weitgehend, da sie sich bzgl. der Metadaten geringfügig unterscheiden können), sodass ihre Aufbereitung und Indizierung auf einheitliche und generische Weise erfolgen kann. Hierzu wurden die Dokumente von ihrer Quelle heruntergeladen, mithilfe des Werkzeugs *Apache Tika*[29] analysiert, weitere Metadaten in der Datenbank gespeichert und die Dokumente mit Metadaten und Textfragmenten für die spätere Suche in Elasticsearch gespeichert und indiziert.

Die frei und als Open Source verfügbare Software Bibliothek Apache Tika gestattet die Analyse verschiedenster Dokumenttypen, wie etwa von PDF, HTML oder Microsoft Office Dokumenten. Soweit diese aus dem Dokument ableitbar sind, liefert Tika zu einem Dokument verschiedene administrative und inhaltliche Metadaten. Die administrativen Metadaten, wie z.B. der Mime-Type, die Erzeugungs- und letzte Änderungszeit, Größe, Seitenanzahl etc. werden in der Regel automatisch in den Dokumenten gespeichert und sind deshalb vollständiger als inhaltliche Metadaten, wie Autor, Titel oder Beschreibung, die vom Autor manuell erfasst worden sein müssen. Dies gilt insbesondere für PDF-Dokumente. Webseiten, die als HTML-Dokumente heruntergeladen und analysiert werden, haben z.B. generell keine Seitenzahlen, auch sonstige Metadaten wie Veröffentlichungsdatum sind oft nicht verfügbar.

Tika liefert weiterhin den kompletten reinen Text eines Dokuments. Bei PDF-Dokumenten orientiert sich die Text-Extraktion unmittelbar am optischen Erscheinungsbild des Textes. Zeilenumbrüche und insbesondere dortige automatische Worttrennungen bleiben deshalb 1:1 erhalten. Absatzgrenzen sind in der Regel durch von Tika eingefügte zusätzliche Zeilenumbrüche erkennbar. Bei einigen Fonts (Schriftarten) werden jedoch zwischen einigen Buchstabenpaaren mitten im Wort Leerzeichen eingefügt – nämlich dann, wenn zwischen diesen Buchstaben optisch „zu viel Platz ist“. Diese Wörter konnten für eine spätere Indizierung kaum rekonstruiert werden. Hier müsste man bei einer Weiterentwicklung per Wörterbuch versuchen, aus Text-

[29] https://tika.apache.org/

schnipseln, die keinen Sinn ergeben, wieder ganze Wörter zusammenzusetzen. Dieses Problem bei Schriftarten trat z.B. in einigen Überschriften bei Dokumenten des Eisenbahn-Bundesamtes auf.

Apache Tika bietet darüber hinaus diverse Möglichkeiten, die Textverarbeitung durch Implementation zusätzlicher Funktionen zu erweitern, sofern dies im Rahmen einer möglichen Weiterentwicklung erforderlich und sinnvoll sein sollte.

Der von Apache Tika erzeugte extrahierte Text kann in verschiedenen Formaten abgebildet werden. Neben reinem Text kommt so z.B. auch eine Abbildung als HTML und XML in Betracht, was eine einfachere weitere Verarbeitung ermöglicht.

7.8 Segmentierung

Um insbesondere in großen Dokumenten zielgerichteter thematisch suchen und Fundstellen treffgenauer anzeigen zu können, ist es notwendig, Dokumente in einzelne Fragmente, wie z.B. Kapitel, Unterkapitel, Seiten oder Abschnitte zu zerlegen und diese Fragmente auch einzeln zu indizieren.

7.8.1 Textextraktion mittels Apache Tika

Prinzipiell ist bei PDF-Dokumenten die logische Dokumentstruktur von Kapiteln, Unterkapiteln und Abschnitten nicht direkt ermittelbar, sondern kann lediglich indirekt aus den Textabständen, Schriftgrößen, Nummerierungen, etc. erschlossen werden; dieses Verfahren ist daher unzuverlässig und liefert nur annähernde korrekte Ergebnisse. Apache Tika erzeugt hier aber i.d.R. die technisch bestmöglichen Ergebnisse. Bei HTML-Seiten und Microsoft Word ist die Erkennung der logischen Dokumentstruktur zumindest theoretisch möglich, sofern die Dokumente entsprechende Strukturinformationen wie <h1> oder entsprechende Word-Formatvorlagen beinhalten. Praktisch wird das aber sehr eingeschränkt durch die Vielgestalt und Freiheit bei der Erstellung der Dokumente. Für den im Demonstrator eingesetzten Zweck der Navigation innerhalb von Dokumenten und die Anzeige der relevanten Textstellen innerhalb dieser Dokumente, scheint dieses Ergebnis aber nach den bisherigen Erfahrungen i.d.R. ausreichend zu sein.

PDF- und Microsoft-Word-Dokumente lassen sich per Tika mittels der HTML-Ausgabe zumindest in ihre Seitenstruktur zerlegen, was im Rahmen der Machbarkeitsstudie auch gemacht wurde. Hier gibt es für jede Seite ein Element mit entsprechender CSS-Klasse. Innerhalb dieser Seiten- gibt es für jede Textzeile ein <p> und bei Leerzeilen ein leeres <p>. Anhand dieser Leerzeilen kann man Textblöcke auf einer Seite identifizieren. Eine Dokumentfragmentierung mit diesen Mitteln hat dann folgende Eigenschaften:

1. Absätze sind, falls nicht korrekt, eher etwas zu groß, falls keine trennende Leerzeile vorhanden war.
2. Kein Absatz ist größer als eine Seite. Absätze, die eigentlich über eine Seitengrenze gehen, werden an der Seitengrenze aufgetrennt (dies als Ausnahme zu Punkt 1).
3. Problematisch sind wiederkehrende Kopf- oder Fußtexte auf den Seiten. Die verbleiben in der Regel immer am nächstliegenden Absatz und verfälschen später vielleicht auch Begriffsgewichtungen bei der Indizierung. Eine Möglichkeit wäre, Kopf- oder Fußtexte anhand der auf jeder Seite wiederkehrenden Texte zu er-

kennen und dann bei der Analyse automatisch aus den Seiten auszublenden.

Die Absatztrennung an Leerzeilen ist nur eine grobe Näherung. Zur Verfeinerung des Verfahrens zur Fragmentierung könnte man beispielsweise versuchen, Zwischenüberschriftpräfixe wie z.B. Nummerierungen zu erkennen oder aus einer stark unterdurchschnittlichen Textzeilenlänge auf ein Absatzende zu schließen.

7.8.2 Speicherung und Indizierung von Dokumenten in Elasticsearch

Nach der Metadaten-Extraktion sowie Textanalyse und -fragmentierung durch Tika werden die Dokumente als JSON-Struktur in Elasticsearch gespeichert und indiziert. Ein solches JSON-Dokument enthält dabei die Metadaten, den Verweis auf das Original-Dokument (URL) zur späteren Detailanzeige, den reinen Text des Dokuments als Gesamttext und eine Liste von Textfragmenten, wobei zu jedem Fragment auch die entsprechende Seitenzahl gespeichert wird. Letzteres dient später dem automatischen Scrolling auf die betreffende Seite einer Fundstelle bei der Dokumentanzeige.

Aus dem Gesamttext und den einzelnen Textfragmenten wurden vor dem Speichern und Indizieren alle Zeilenumbrüche entfernt. Da die Texte von PDF-Dokumenten deren Druckansicht widerspiegeln, entstehen bei vielen Zeilenumbrüchen auch Worttrennungen mit Bindestrich. Auch diese Trennungen wurden entfernt, um die Texte anschließend mit ungetrennten Wörtern zu indizieren.

7.9 Semantische Indexierung

Die semantische Indexierung der gespeicherten Dokumente ist der eigentliche zentrale, KI-basierte Kern der Wissensplattform. Hierdurch wird erst aus einer einfachen Menge von gespeicherten Dokumenten aus dem Umweltbereich tatsächlich technisch nutzbares „Wissen“ in dem Sinne, dass der Inhalt der Dokumente analysiert wird, diesen Wissen aus allen Dokumente miteinander verknüpft wird und auf diese Weise tatsächlich eine echte Wissensplattform entsteht, die dann für die Suche nach relevanten Informationen genutzt werden kann.

7.9.1 Grundlage: die BMVI-Ontologie

Die Grundlage für die semantische Indexierung ist die Ontologie, die aus der im Anschluss an den ersten Workshop gemeinsam mit den Teilnehmenden erarbeiteten Vokabularsammlung entwickelt wurde (vgl. 2.2.9). Die Vokabularsammlung enthält eine Liste von Begriffen aus

- der biologischen Systematik für die Tierarten,
- dem Planungswesen, Baurecht und Verwaltungswissen,
- einer Liste wichtiger Gesetze im Planungs-, Verwaltungs- und Umweltrecht,
- wichtiger Begriffe aus Fachartikeln zu Arten- und Umweltschutz sowie
- geografischen Bezeichnungen für Bundesländer, Naturschutz- und Landschaftsschutzgebieten.

Es wurden die Beziehungen der Konzepte untereinander wie Verwandtschaften und Ähnlichkeiten modelliert. Das Ergebnis ist eine Ontologie als RDF-Datei[30] in sog. Turtle-Notation, die mit dem Ontologie-Editor Protegé[31] vom *Stanford Center for Biomedical Informatics Research* eingelesen, angezeigt und geändert werden kann. Die Ontologie enthält ca. 1800 Konzepte (Begriffe) mit insgesamt ca. 7500 Axiomen, d.h. Beziehungen zwischen Konzepten, z.B. „Zauneidechse ist eine Unterart von Reptilien“, „Zauneidechse hat das Synonym L. agilis“.

[30] RDF: Resource Description Format; dies ist eine formale Sprache mittels derer Ontologien erstellt werden können.

[31] Vgl. https://de.wikipedia.org/wiki/Prot%C3%A9g%C3%A9_(Software)

Der Protegé-Editor (Abbildung 10) bietet eine sehr gute Möglichkeit für die Pflege der Vokabularsammlung. Es ist durchaus denkbar, dass nach einer Zeit der Einarbeitung die Ontologie ggf. von den Fachexperten selbst gepflegt werden könnte.

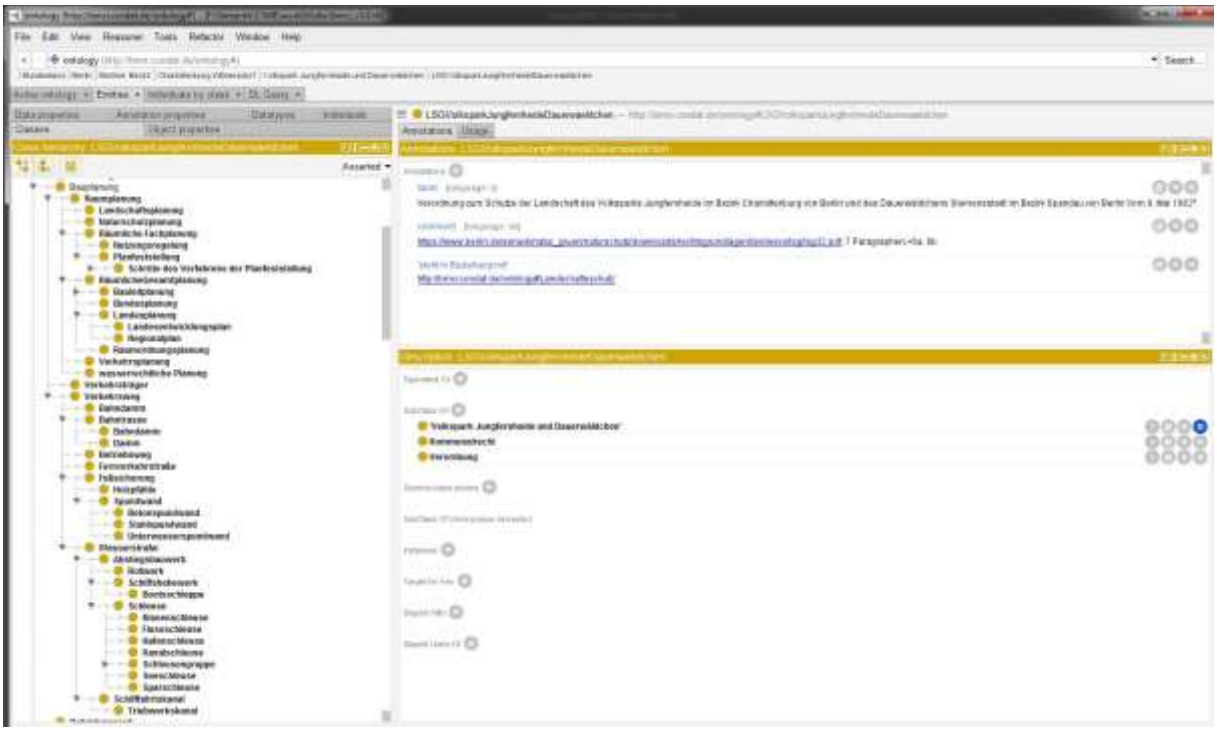

Abbildung 10: Ausschnitt aus der BMVI-Ontologie

7.9.2 Textanalyse

Die Textanalyse erfolgt mit Hilfe des Open Source Tools *Apache OpenNLP*[32] in drei Schritten. Hierbei muss die Sprache des Textes (deutsch oder englisch) bereits bekannt sein. OpenNLP bietet allerdings auch einen *Language Detection Service*, mit dessen Hilfe die Sprache eines Textes ggf. automatisch erkannt werden kann.

Zunächst wird der Text mit dem *OpenNLP-Tokenizer* in einzelne Wörter (*Token*) zerlegt, wofür die Standard-Modelle (englisch und deutsch) von OpenNLP ausreichend sind.

[32] *Die Apache OpenNLP Bibliothek ist ein auf maschinellem Lernen basierendes Toolkit in der Programmiersprache Java für die Verarbeitung von natürlichsprachlichem Text im Bereich Computerlinguistik oder Natural Language Processing.* Quelle : Wikipedia, https://de.wikipedia.org/wiki/Apache_OpenNLP; *s.* https://opennlp.apache.org/

Beispiel: Text: "... für die Haselmäuse …" → Token: "Haselmäuse"

Als zweites werden die Wortarten der gefundenen Token mit Hilfe des *OpenNLP-POSTaggers* ermittelt. Auch hierfür sind die Standardmodelle (englisch und deutsch) von OpenNLP ausreichend, könnten aber bei einer eventuellen Erweiterung der BMVI-Wissensplattform ebenfalls angepasst werden.

Beispiel: Token "Haselmäuse" → Wortart: Substantiv

Anschließend werden die gefundenen Token auf ihre sog. Normalform gebracht ("*Stemming*"). Für deutsche Texte wird hierfür der *DictionaryLemmatizer* von OpenNLP verwendet. Das zugrundeliegende Sprachmodell wurde von der Condat aus dem deutschsprachigen Wiktionary, einem frei verfügbaren Online-Wörterbuch,[33] abgeleitet und ist Bestandteil der SME. Für englische Texte wird der *(Rule-based-)Lemmatizer* und das englische Sprachmodell von OpenNLP verwendet.

Beispiel: Token "Haselmäuse", Wortart Substantiv → normiertes Token: "Haselmaus"

Die so ermittelten normierten Token werden einer umfangreichen Auswahl und Nachbearbeitung unterzogen. Zunächst werden alle Token ausgeschlossen, die nicht als Substantive oder Eigennamen erkannt wurden. Weiterhin werden Konkatenierungen aus Adjektiven, Substantiven und Eigennamen gebildet, wodurch komplexere Token wie "verbindlicher Bauleitplan" oder "echte Eidechse" entstehen.

Die Gesamtheit dieser normierten und erzeugten Token wird *textueller Fingerabdruck* genannt. Er stellt eine Art Kandidatenliste für die Erzeugung des semantischen Fingerabdrucks dar, d.h. zur Erzeugung des semantischen Fingerabdrucks werden die hier gefundenen Token des textuellen Fingerabdrucks herangezogen.

7.9.3 Semantische Analyse

Die gefundenen Token im textuellen Fingerabdruck werden mit den Labeln der Konzepte in der BMVI-Ontologie abgeglichen. Wird ein entsprechendes Label für ein Token gefunden, so wird das dazugehörige Konzept in den semantischen Fingerabdruck des Dokuments aufgenommen. Da die BMVI-Ontologie relativ klein ist, gibt es hierbei keine Probleme mit Mehrdeutigkeiten, daher konnte auf eine ansonsten erfor-

[33] https://de.wiktionary.org/

derliche kontextabhängige Disambiguierung (Auflösung sprachlicher Mehrdeutigkeiten) verzichtet werden.

Beispiel (*deC steht für den deutschen Kategorie-Namensraum)*: normiertes Token: “Haselmaus” → Konzept: “deC:Haselmaus” (über das Label “Haselmaus”@de)

Die einzelnen Einträge des semantischen Fingerabdrucks erhalten noch ein Gewicht, das ihre Relevanz widerspiegelt. Diese Zahl wird aus der Häufigkeit des Auftretens des Konzepts im Dokument, der Häufigkeit im gesamten Dokumentenkorpus und dem Ort des Auftretens berechnet. So wird bspw. ein Konzept, das im Titel eines Dokuments auftaucht, höher gewichtet als wenn es im Detailtext auftauchen würde.

Lateinische Artbezeichnungen

Die Erkennung lateinischer Artbezeichnungen wurde als eine Sonderlösung für die Wissensplattform realisiert anstatt für OpenNLP lateinische Sprachmodelle zu erzeugen. Dies war wesentlich einfacher, da die relevanten Bezeichnungen bekannt sind und es sich nur um gut 100 Token handelte.

Bespiel: Text: “... Muscardinus avellanarius …” → Konzept: “deC:Haselmaus”

Verarbeitung von rechtlichen Texten

Für die Erkennung von Gesetzen und Gesetzesparagraphen wurde ebenfalls eine Speziallösung innerhalb dieser Machbarkeitsstudie entwickelt. Diese basiert auf einem sehr komplexen regulären Ausdruck (“Suchpattern”, über 350 Zeichen), der die am häufigsten verwendeten Formate erkennt.

Beispiel: Text “... BNatSchG § 1 …” → Konzept: “deC:BNatSchG_P_1”

Die bei der Segmentierung erzeugten Dokumente, die die Texte bzw. Metadaten der ursprünglichen Dokumente oder deren Fragmente enthalten werden zusätzlich in die Condat Smart Media Engine importiert und dort gespeichert. Die Attribute *Titel* und *Detailtext* werden außerdem einer textuellen und einer semantischen Analyse unterzogen.

Beispiel für einen semantischen Fingerabdruck für den Artikel "Nieby: Neue Heimat für die Zauneidechse" von shz.de:

Konzept	**Gewicht**
deC:Zauneidechse	3,87
deC:hamburg	1,73
deC:Pflanzenart	1
deC:Habitat	1,73
deC:Verbreitungsgebiet	1
deC:Kommentar	1
deC:Deich	1
deC:Naturschutz	1,41
deC:schleswig_holstein	1,41
deC:Familie	1
deC:Naturschutzplanung	1
deC:Ausnahme	1
deC:Kaeferart	1

7.10 Semantische Suche

Für die semantische Suche wird ein semantischer Fingerabdruck benötigt. Wird die Suche von einem analysierten Dokument aus ausgeführt, liegt dieser bereits vor. Wird die Suche mit einem Stichwort oder einer Liste von Stichwörtern ausgeführt, so wird ähnlich wie bei der semantischen Indexierung aus den Stichwörtern ein seman-

tischer Fingerabdruck abgeleitet, wobei die Gewichte alle auf 1 gesetzt werden. Für das Stichwort “Haselmaus” entsteht z. B. der Basissuchfingerabdruck:

Konzept	**Gewicht**
deC:Haselmaus	1

Der gleiche Fingerabdruck würde entstehen bei der Angabe “Hazel dormouse” oder “Muscardinus avellanarius” aufgrund der mehrsprachigen Ontologie bzw. den dort eingetragenen Synonymen.

7.10.1 Semantische Suchraumerweiterung

Der Basissuchfingerabdruck kann nun mit Hilfe der Informationen in der Ontologie schrittweise erweitert werden um Oberkonzepte, Unterkonzepte oder thematisch verwandte Konzepte. Wie, in welcher Reihenfolge und mit welchen Gewichten, ist komplett konfigurierbar. Obiges Beispiel etwa kann in zwei Schritten auf alle relevanten Nagetierarten erweitert werden:

Beispiel: Erweiterung auf Oberkonzepte mit Gewicht 10%:

Konzept	**Gewicht**
deC:Haselmaus	1
deC:Nagetierart	0,1

Beispiel: Erweiterung auf Unterkonzepte mit Gewicht 10%:

Konzept	**Gewicht**
deC:Haselmaus	1
deC:Nagetierart	0,1
deC:Biber	0,01

Die Behandlung der Gewichte pro Art des Erweiterungsschrittes (Oberkonzept, Unterkonzept, thematisch verwandtes Konzept) kann unterschiedlich eingestellt werden. Weiterhin kann für jeden Einzelaufruf der Services die Stärke der semantischen Suchraumerweiterung insgesamt und für die Teilbereiche der Ontologie *deC:Tierart*, *deC:Bundesland*, *deC:Massnahme* und *deC:Rechtsnorm* (Gesetze, Verordnungen,

etc.) einzeln angesteuert werden. Auch das Abschalten der semantischen Suchraumerweiterung für einzelne Teilbereiche ist möglich. So ist es bspw. möglich, mit den Stichwörtern “Haselmaus” und “Brandenburg” exakt nach “Haselmaus” zu suchen, aber den regionalen Bezug aufzuweichen, auch wenn nach “Brandenburg” vorrangig gesucht wird (größere Relevanz der Ergebnisse). Gesteuert wird dieses Verhalten durch die Einstellung der in Abbildung 6 gezeigten Suchparameter auf der Benutzeroberfläche des Demonstrators.

7.10.2 Ähnlichkeitssuche

Der erweiterte Suchfingerabdruck wird ggf. noch verkleinert (sog. *Dimensionsreduktion*), wenn er sehr groß geworden ist, um eine akzeptable Performanz der Suche zu erreichen. Hierbei werden die Konzepte mit den geringsten Gewichten aus dem Fingerabdruck wieder entfernt. Im Rahmen der Machbarkeitsstudie war dies jedoch nicht notwendig.

Der endgültige Suchfingerabdruck wird mit den Fingerabdrücken aller analysierten Dokumente verglichen und ein Ähnlichkeitsmaß ermittelt. Hierbei werden die jeweiligen Gewichte von übereinstimmenden Konzepten in beiden Fingerabdrücken miteinander multipliziert und über alle Konzepte (auf)addiert. Die Ähnlichkeitsmaße werden schließlich noch normiert, so dass die Ähnlichkeit eines Dokuments mit sich selbst (zumindest bei abgeschalteter Suchraumerweiterung) genau 100% ist.

Beispiel:

Die Ähnlichkeit des obigen Suchfingerabdrucks

Konzept	**Gewicht**
deC:Haselmaus	1
deC:Nagetierart	0,1
deC:Biber	0,01

mit dem Fingerabdruck

Konzept	**Gewicht**
deC:Biber	2

ergibt sich somit durch das Konzept *deC:Biber* zu: 0,01 * 2 = **0,02**.

Bei entsprechender Konfiguration wird bei Suche nach „Haselmaus“ (oder einer dazu äquivalenten englischen oder lateinischen Angabe) auch ein Dokument gefunden in dem nur “Biber” (oder eine andere äquivalente Angabe) steht.

Die gefundenen Dokumente werden absteigend nach dem ermittelten Ähnlichkeits- oder Relevanzmaß sortiert und die Service-Schnittstellen liefern maximal die gewünschte Anzahl an Dokumenten zurück. Das Dokument, mit dem eine Suche ausgelöst wurde, kann und wird in der Regel dabei weggelassen.

8. Zusammenfassung und Fazit

Bei der Planung und Genehmigung von Verkehrsinfrastrukturprojekten sind umfangreiche umwelt-, natur- und artenschutzrechtliche Anforderungen zu beachten, deren Umfang und Komplexität in den vergangenen Jahren stetig angestiegen sind. Um die am Planungs- und Genehmigungsprozess Beteiligten bei der ressourcenintensiven Recherche nach projektrelevanten Informationen zu unterstützen, hat das Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) 2019 die Entwicklung einer softwarebasierten Wissensplattform initiiert, die unter Nutzung selbstlernender Technologien (Künstliche Intelligenz) die Möglichkeit bieten soll, zu jeder fachlichen Frage im Bereich des Artenschutzes verlässliche und zitierfähige Informationen abzurufen.

Im Rahmen der Machbarkeitsstudie sollte geklärt werden, inwieweit eine solche Wissensplattform softwarebasiert und mithilfe von Künstlicher Intelligenz aufgebaut werden kann und inwieweit eine solche Anwendung für die Vorhabenträger und Genehmigungsbehörden einen Mehrwert hätte. Dazu sollte ein Demonstrator entwickelt werden, der die Ergebnisse der Studie anschaulich macht und mit dem erste fachspezifische Suchen durchgeführt werden können. Die Anforderungen an die Wissensplattform resp. den Demonstrator wurden in enger Zusammenarbeit mit potentiellen Anwender/innen aus den Planungs- und Genehmigungsbehörden entwickelt.

Vorgehen/Methodik

Die vom Auftragnehmer Condat AG gemeinsam mit dem Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) als Unterauftragnehmer entwickelte Wissensplattform erbringt folgende Informationsleistungen:

- Die Plattform ermöglicht die Recherche von im Internet zur Verfügung stehenden Informationen zu zehn geschützten Tierarten. Die Suche unterstützt alle üblichen Dateiformate (PDF, MS-Office, HTML etc.) sowie weitere strukturierte Formate, die über Standardschnittstellen abgefragt werden können.
- Die gefundenen Informationen werden mit Hilfe von Methoden der Künstlichen Intelligenz (KI) aufbereitet. Zur Verarbeitung gehört zunächst eine selbstlernende Relevanzbewertung, die auf Basis einer fachspezifischen, gemeinsam mit den künftigen Nutzenden entwickelten Ontologie sicherstellt, dass nur relevante Dokumente in die Wissensplattform übernommen werden. In einem weiteren Schritt werden die importierten Dokumente mittels semantischer KI-Methoden automatisch um Zusatzinformationen angereichert und kategorisiert. Bei diesen Zusatzinformationen handelt es sich um das Bundesland, aus dem das Dokument stammt und um die Art des Dokuments (Fachinformatio-

nen, Rechtliches, Planung). Eine textuelle und inhaltliche Analyse mit Verfahren aus dem Maschinellen Lernen extrahiert relevante Informationen aus den Dokumenten und ermöglicht dadurch Querverweise zwischen Dokumenten und Quellen.
- Mit dem Demonstrator kann die zukünftige Wissensplattform auszugsweise vorgeführt und ihre Leistungsfähigkeit evaluiert werden. Der Demonstrator gibt zudem einen Ausblick auf die mögliche Gestaltung der Benutzeroberfläche, die die explorativen Suchprozesse nutzerfreundlich unterstützen soll.

Ergebnisbewertung / Ausblick

Die in der Plattform eingesetzten Methoden der Künstlichen Intelligenz können den Rechercheaufwand im Vergleich zu einfachen, keyword-basierten Suchen in verteilten Archiven signifikant reduzieren und genauere und umfassendere Ergebnisse liefern. Die Benutzeroberfläche der Plattform stellt vorhandenes, kontextbezogenes Wissen übersichtlich dar und erlaubt ein direktes Sichten der zugrundeliegenden Dokumente mit Sprungmarken zu relevanten Abschnitten.

Im Ergebnis bestätigen Machbarkeitsstudie und Demonstrator, dass eine auf KI-Technologien basierende, selbstlernende Wissensplattform für das skizzierte Anwendungsszenario des BMVI umsetzbar und geeignet ist, um die Effizienz von Planungs- und Genehmigungsprozessen in der Verkehrswegeplanung zu erhöhen.

Es ergaben sich jedoch im Rahmen der Machbarkeitsstudie auch Grenzen.

- Die Inhalte bekannter Quellen konnten nicht vollständig in die Plattform überführt werden, da dies nicht automatisiert erfolgen konnte, sondern alle Quellen händisch angebunden werden mussten.
- Für das Training der KI-basierten Komponenten war die vorhandene Datenmenge teilweise noch zu gering.
- Einige Funktionalitäten wurden nur exemplarisch realisiert, insbesondere dort, wo die technische Machbarkeit außer Frage stand und lediglich eine Frage des eingesetzten Aufwandes darstellte (z.B. Erzeugung von Fragmenten aus PDF-Dokumenten)
- Themen wie Performanz und Skalierbarkeit des Systems auf Hunderte von Nutzenden wurden im Rahmen der Machbarkeitsstudie nicht betrachtet.

Im Hinblick auf die Frage der Vollständigkeit der gefundenen Informationen kann festgehalten werden, dass durch klassische Internet-Suchverfahren alle digital verfügbaren und zugreifbaren Dokumente gefunden werden können, wenn diese durch

Suchanbieter wie Google oder Microsoft Bing indiziert werden. Im Rahmen der Machbarkeitsstudie wurde hierbei durch Mechanismen wie die semantische Suchraumerweiterung auf Basis der erstellten Ontologie die klassische Internet-Suche noch dergestalt erweitert, dass auch relevante Informationen gefunden werden können, die über eine einfache Schlüsselwortsuche nicht gefunden werden können. Gleichwohl müssten diese Suchen in einem echten Produktivbetrieb immer wieder überprüft und durch eine ständige Pflege der Ontologie vervollständigt und angepasst werden.

Andererseits bedeutet „digital verfügbar und zugreifbar“ aber auch, dass Dokumente bzw. Informationen, die nicht digital verfügbar oder zwar digital verfügbar aber nicht zugreifbar sind, nicht direkt gefunden werden können. Das gilt zum Beispiel, wenn sie durch einen Zugangsschutz geschützt bzw. kostenpflichtig sind oder sich in einem Intranet befinden. Technologisch ist die Integration solcher Quellen aber gleichermaßen möglich, z.B. durch Einscannen und OCR-Verfahren (Optical Character Recognition, d.h. Texterkennung) im Falle analog vorhandener Quellen bzw. durch eine direkte Anbindung der Quellen bei bereits bekannten Portalen oder Dokumenten wie z.B. wissenschaftlichen Fachartikel.

Offen geblieben ist das Auffinden bisher gänzlich unbekannter Quellen, die weder über die Internet-Suche noch über die Anbindung geschützter Portale bzw. Datenbanken noch durch das Einscannen von Fachartikeln direkt Eingang in die Plattform finden können. Diese Quellen können im Ergebnis nur *indirekt* gefunden werden, d.h. durch die Analyse der gefundenen Dokumente in Bezug auf evtl. Verweise auf noch nicht bekannte Quellen. Hierzu können insbesondere die Literaturangaben in Fachartikeln herangezogen werden. Auch hierfür wurde mit dem Verfahren der semantischen Indexierung in der Wissensplattform die Grundlage gelegt. Im Zuge einer evtl. Erweiterung der Plattform müsste dieses Verfahren anhand einer größeren Menge an beispielhaften Daten speziell trainiert werden.

Um einen Mehrwert für die Vorhabenträger und Genehmigungsbehörden herzustellen, ist daher eine Weiterentwicklung des Systems empfehlenswert.

Anhang

A. Technische Komponenten des Systems

Im Folgenden werden noch einmal zusammenfassend die verschiedenen Produkte, Komponenten, Software-Bibliotheken und Werkzeuge von Drittanbietern aufgeführt, die im Rahmen der Realisierung des Gesamtsystems eingesetzt wurden.

Komponente/Produkt	**Funktion**
Angular	Software-Bibliothek für die Erstellung der Benutzeroberfläche
Apache OpenNLP	Software-Bibliothek für die Verarbeitung von natürlich sprachlichen Texten
Apache Tika	Software Bibliothek zur Extraktion von Text und Metadaten aus Dokumenten (PDF, HTML, Word, etc.)
Carrot2	Software-Bibliothek für das Clustern von Texten
Elasticsearch	Suchmaschine / Indexierer zur Speicherung der extrahierten Volltexte
jsoup	Software-Bibliothek zum Crawlen von Webseiten und Downloaden des Inhalts
Microsoft Azure Bing API	Schnittstelle für die Internet Suchfunktion
MongoDB	Datenbank für die Speicherung der Metadaten
Protégé	Editor für die Pflege von Ontologien
RabbitMQ	Message Broker Software für die Integration der verschiedenen Komponenten untereinander
Selenium	Software-Bibliothek für das automatisierte, interaktive Ansteuern von Webseiten

Die Entwicklung des Gesamtsystems fand unter Verwendung der Programmiersprachen Python, Java und Typescript statt.

B. Quellensammlung

Name	URL	Imp.	Bemerkung
Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen	http://artenschutz.naturschutzinformationen.nrw.de	ja	Download-Liste diverser Publikationen unter http://artenschutz.naturschutzinformationen.nrw.de/artenschutz/de/downloads
Aarhus-Convention (angeboten vom Unabhängigen Institut für Umweltfragen)	http://www.aarhus-konvention.de	ja	Links zu einigen relevanten Dokumenten
www.amphibien-reptilien.com	http://www.amphibien-reptilien.com	nein	umfangreiche Site mit diversen Infos und Bildmaterial, keine PDF-Dokumente, sehr aufwändig zu crawlen
Verkündungsportal des Saarlandes	http://www.amtsblatt.saarland.de	nein	Suchportal kompliziert zu scrapen, da PDF-Links erst indirekt in Ergebnislisten erreichbar
Bundesamt für Güterverkehr	http://www.bag.bund.de	nein	Suchportal ohne relevante Treffer zu typischen Stichworten
Bayerische Staatsregierung	http://www.bayern.de	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten
Schweizerisches Zentrum für die Kartografie der Fauna	http://www.cscf.ch	ja	diverse Publikationslisten aber fast alles auf Französisch, die "Biberfachstelle" hat deutsche Dokumente, Artenschutzblätter und Rote Listen auch auf Deutsch
private Website "Deutschlands Natur"	http://www.ffh-gebiete.de	nein	Übersicht der FFH-Gebiete, Tier- und Pflanzenarten etc.; kein Suchportal und keine Downloadlisten. nicht downloadbar
Regierung der Oberpfalz	http://www.regierung.oberpfalz.bayern.de	ja	umfangreiches Suchportal
Artenschutzdatenbank des Bundesamt für Naturschutz	http://www.wisia.de	ja	Dokumente wurden einzeln importiert; dies ist eine Namensdatenbank für Tier- und Planzennamen, nicht automatisiert crawlbar

Wasserstraßen-Neubauamt Berlin	http://www.wna-berlin.de	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten
BioConsult SH GmbH & Co. KG	https://bioconsult-sh.de	ja	diverse Publikationslisten
Landesregierung Brandenburg	https://bravors.brandenburg.de	ja	vermutlich nur anderer Zugang zu "recht.brandenburg.de"; 4 relevante Dokumente wurden einzeln importiert
OpenLegalData-Portal	https://de.openlegaldata.io	ja	OpenLegalData-Portal (freier Zugang zu juristischen Informationen); Das Portal bietet eine REST-API zum Abrufen von JSON-Dokumenten zu juristischen Informationen. Diese JSON-Dokumente enthalten diverse Metadaten sowie die eigentlichen juristischen Texte in HTML-Form. Jeder Fall ist über eine eindeutige Nummer identifiziert. Über diese Nummer ist im Portal auch eine URL zur fertig layouteten Anzeige des Dokumentes in HTML-Form aufrufbar.
dejure.org	https://dejure.org	nein	Stichwortsuche nach Gesetzen und Verordnungen, Ergebnisse sehr heterogen mit Weiterleitungen auf Drittportale
digitales Monitoring und Informationssystem des Berliner Energie- und Klimaschutzprogramms	https://dibek.berlin.de	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten
docplayer.org	https://docplayer.org	nein	Suchportal für diverse Dokumente, Download mit Captcha gegen Robots gesichert - dadurch nicht automatisiert scrape-bar
Europäische Kommission	https://ec.europa.eu	ja	diverse Publikationslisten unter https://ec.europa.eu/environment/nature/natura2000/index_en.htm
Internethandbuch zu den Arten der FFH-Richtlinie (Bundesamt für Naturschutz)	https://ffh-anhang4.bfn.de	nein	Internethandbuch zu den Arten der FFH-Richtlinie, keine PDF-Dokumente
Informationsdienst Umweltrecht e.V.	https://idur.de	ja	diverse Publikationslisten zu Umweltrecht
Justizportal des Bundes und der Länder	https://justiz.de	nein	Linksammlung auf Internetangebote der Justiz des Bundes und der Länder, Suchportal ohne relevante Treffer zu typischen Stichworten

Landesamt für Bauen und Verkehr Brandenburg	https://lbv.brandenburg.de	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten
Landesverwaltungsamt Sachsen-Anhalt	https://lvwa.sachsen-anhalt.de	ja	umfangreiches Suchportal; Linksammlung zu Planfeststellung und Gesetzen unter https://lvwa.sachsen-anhalt.de/das-lvwa/wirtschaft-bauwesen-verkehr/planfeststellung, konkrete Planfeststellungen stehen immer nur bis zum Ende des Auslegungszeitraumes zum Herunterladen zur Verfügung
Ministerium für Infrastruktur und Landesplanung Brandenburg	https://mil.brandenburg.de	nein	umfangreiches Suchportal
Landesrecht Brandenburg	https://recht.brandenburg.de	ja	Suchportal für Gesetz- und Verordnungsblätter für das Land Brandenburg
Regierungspräsidien Baden-Württemberg	https://rp.baden-wuerttemberg.de	nein	umfangreiches Suchportal
Freistaat Thüringen	https://thueringen.de	nein	umfangreiches Suchportal
UVP-Portal Niedersachsen	https://uvp.niedersachsen.de/portal	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten
Ministerium für Verkehr Baden-Württemberg	https://vm.baden-wuerttemberg.de	ja	Suchportal mit nur wenigen Treffern zu typischen Stichworten
Land Baden-Württemberg	https://www.baden-wuerttemberg.de	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten
Berlin	https://www.berlin.de	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten, keine PDF-Dokumente
Bundesamt für Naturschutz	https://www.bfn.de	ja	Download-Liste zu Veröffentlichungen unter https://www.bfn.de/infothek/veroeffentlichungen/bfn-skripten/nummerische-sortierung.html
Bundesministerium für Verkehr und digitale Infrastruktur	https://www.bmvi.de	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten
Land Brandenburg	https://www.brandenburg.de	nein	Aktuelles und Allgemeines zu Brandenburg; aber relevantes Suchportal unter https://mil.brandenburg.de (Ministerium für Infrastruktur und Landesplanung)
Bundesverfassungsgericht	https://www.bundesverfassungsgericht.de	nein	Suchportal für Entscheidungen als HTML oder PDF unter https://www.bundesverfassungsgericht.de/SiteGlobals/Forms/Suche/Entscheidungensuche_Formular.html

BUND-Naturschutz in Bayern e.V.	https://www.bund-naturschutz.de	nein	gutes Umwelt-Suchportal, teilweise kostenpflichtiges Material, heterogene Strukturen, darum schwer zu scrapen
Bundesverwaltungsgericht	https://www.bverwg.de	nein	im Suchportal Urteile und Beschlüsse als HTML verfügbar unter z.B. https://www.bverwg.de/suche?q=Artenschutz&db=e&dt=&lim=10&start=1#
CITES - Convention on International Trade in Endangered Species of Wild Fauna and Flora	https://www.cites.org	nein	Veröffentlichungsliste unter https://www.cites.org/eng/resources/publications.php; teilweise kostenpflichtig
Eisenbahn-Bundesamt	https://www.eba.bund.de	ja	Scraping über Suchportal durchgeführt durch Suche nach typischen Stichworten
fachzeitungen.de	https://www.fachzeitungen.de	nein	Suchportal mit wenigen Treffern zu typischen Stichwörtern, kostenpflichtige Angebote, dadurch nicht scrape-bar
Wasserstraßen- und Schifffahrtsverwaltung des Bundes	https://www.gdws.wsv.bund.de	nein	gutes Suchportal, viele Dokumente mit jeweils vielen Anhängen; Suchmöglichkeit nach konkreten Planfeststellungsverfahren unter https://www.gdws.wsv.bund.de/SiteGlobals/Forms/Suche/Planfeststellungsverfahren/Planfeststellungsverfahren_Formular.html (komplex und schwer zu scrapen)
Zentrales Datenmanagement (ZDM) Küstendaten der wsv.bund.de	https://www.kuestendaten.de	ja	Publikationsliste zu diversen Wasserstraßenprojekten im Umweltkontext unter https://www.kuestendaten.de/DE/dynamisch/publikationen/index.php.html
Hamburg	https://www.hamburg.de	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten; spezielles Suchportal für Umweltverträglichkeitsprüfungen unter https://www.hamburg.de/umweltvertraeglichkeitspruefungen-hamburg/
juris.de	https://www.juris.de	nein	kostenpflichtiges Rechtsportal, Account erforderlich
Landesportal Nordrhein-Westfalen	https://www.land.nrw	nein	Suchportal ohne relevante Treffer zu typischen Stichworten
Zeitschrift LOK Report	https://www.lok-report.de	nein	Suchportal verfügbar, einige Treffer zu typischen Stichworten, nur kurze HTML Artikel

Landesanstalt für Umwelt Baden-Württemberg	https://www.lubw.baden-wuerttemberg.de/natur-und-landschaft	ja	neues Dokumentportal unter https://pudi.lubw.de/publikationen
Metadatenverbund	https://www.metaver.de	nein	Umwelt-Metadatenverbund von 7 Bundesländern
Movebank for Animal Tracking Data	https://www.movebank.org	nein	Movebank for Animal Tracking Data, Publikationen, Account erforderlich, vieles kostenpflichtig
Zeitschrift Naturschutz und Landschaftsplanung	https://www.nul-online.de	nein	Suchportal, und Inhaltslisten frei, konkrete Artikel und Hefte nur für Abonnenten mit Anmeldung
Projekt der Anhörungsbehörde im Landesamt für Bauen und Verkehr (LBV) Brandenburg	https://www.o-sp.de/lbvbrandenburg/pfb.php	nein	alle Unterlagen zu konkreten Planfeststellungen, komplexe Struktur, schwer zu scrapen
ResearchGate	https://www.researchgate.net	nein	Account erforderlich
Rheinland-Pfalz	https://www.rlp.de	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten
Saarland	https://www.saarland.de	nein	Suchportal mit nur wenigen Treffern zu typischen Stichworten
Schleswig-Holstein	https://www.schleswig-holstein.de	nein	umfangreiches Suchportal
springer.com	https://www.springer.com	nein	kostenpflichtige Angebote
Unabhängiges Institut für Umweltfragen	https://www.ufu.de	ja	kein Suchportal gefunden, einige Publikationen unter https://www.ufu.de/publikationen/; Jahresberichte wurden importiert
Umwelt in Sachsen	https://www.umwelt.sachsen.de	ja	gutes Umwelt-Suchportal
UVP-Portal des Bundes	https://www.uvp-portal.de	nein	Metadaten zu UVP's von Bundesvorhaben, UVP's selbst liegen nur zu bestimmten Terminen öffentlich aus
UVP-Portal der Länder	https://www.uvp-verbund.de	nein	Suchportal zu Umweltverträglichkeitsprüfungen der Länder

Wolf in Sachsen	https://www.wolf.sachsen.de	nein	Suchportal ohne relevante Treffer zu typischen Stichworten, diverse Publikationen zum Thema "Wolf in Sachsen" verfügbar

C. Vokabularsammlung

Tierart
- Nagetiere (Rodentia)
- Biber (*Castor fiber)*, *Castoridea*
- Haselmaus, *Muscardinus avellanarius*

Reptilien
- Zauneidechse, *Lacerta agilis, L. agilis*

Käfer
- Eremit, *Osmoderma eremita*

Vögel
- Eisvogel, *Alcedo atthis*
- Baumfalke, *Falco subbuteo*

Amphibien
- Kreuzkröte, *Epidalea calamita, Bufo calamita*

Insekten
- Nachtkerzenschwärmer, *Proserpinus proserpina*

Fische
- Rapfen, *Leuciscus aspius, Aspius aspius*

Fledermaus, *Microchiroptera*

Ökologie
- Habitat, Lebensraum
  - Lineares Habitat
- Lebensraumtyp
- Fortpflanzungsstätte
- Ruhestätte
- Verbreitungsgebiet, Vorkommen
- Schutzstatus
- Erhaltungszustand
- Gefährdungen und Beeinträchtigungen
- Zukunftsaussichten
- Handlungsempfehlungen

- Habitatpräferenz
- Population
  - (Teil-) Population
- Individuum
  - adult
  - subadult

juvenil
Individuendichte
aquatisch
semiaquatisch
terrestrisch
tagaktiv
nachtaktiv

Abundanz
Biotop
Feuchtbiotop ...
Bastardisierungen
Biodiversität
Biodiversitätsverlust
klimatische Empfindlichkeit (Vulnerabilität)
naturschutzfachliche Anpassungsszenarien

Vorhabensgebiet
Bahnlinie
Konversionsfläche
Bahndamm
Kanal
Uferböschung
Steinschüttung
Senkrechtufer
Tierausstieg
Spundwand
Zuwegung
Baustelleneinrichtungsfläche
Betriebsweg
Arbeitsfläche
Fluss
Wehr
Schleuse
Bootsschleppe
Fischaufstiegsanlage

Revier
Biberrevier
Biberbauten
Wohnbauten
Wohnkessel
Biberröhren, Fressröhren, Fluchtröhren, Spielröhren
(künstliche) Biberburg, Erdbau, Wall, Knüppelburg
Wohnhöhlen
Biberdamm, Damm
Nahrungsflos

Biberkonflikte
- Fraßschäden, Fraßaktivitäten an Feldfrüchten, Biberfraßgänge
- Fällen von Gehölzen
- Grabaktivitäten
- Schäden durch Unterminierung
- Dammstauaktivitäten,
  - Bibereinstau,
  - Überflutete Wiese nach Bibereinstau,
  - Eingestauter Biberwechsel

Lösungen Biberkonflikte
- Schaffen von ungenutzten Uferflächen
- Lokale Einzelmaßnahmen
- Biberersatzburg
- künstlicher Biberbau
- künstliche Biberburg
- Leiteinrichtung
- Ökologische Durchgängigkeit

Schutzmaßnahme
- Baumschutzmaßnahmen vor Biber
  - Manschette aus Maschendraht, Drahthosen, Schälschutzmittel auf Sandbasis (WÖBRA), Estrichgitter, Umwickeln mit 4-Eck-Geflecht
- Ufersicherungsmaßnahmen
- Dammdrainagen
- Ausleihe von Elektrozaun, oder Fallen
  - Aufbauschema, Aufbauanleitung, Zauntester, Haspel, Grundausstattung für Zaunsysteme 9V, 12V,
- Sicherung von Durchlässen
- Auswahl von Frucht und Fruchtfolge auf landwirtschaftlichen Flächen
- Biberdichte Zäunung (fester Zaun)
- Kurzfristige Vergrämungsmaßnahmen
  - Abtragen und Entfernen von Biberdämmen
- Kanisterketten gegen Biberdämme
- Schutz von Durchlässen mit Drahtgittern
- Beblechen von Brettern in Mönchen und Wehren, Fütterungen
- Entfernen von Verlandungen und Vertiefen von Gewässern
- Umsiedlung
- Vergrämung
- Ausnahmen
- Bauzeitenregelung
- Tierausstieg
- Steinschüttung

Wildtiermanagement

Bibermanagement

Quartier Fledermaus
- gebäudebewohnende Fledermaus

baumbewohnende Fledermaus
Gebäudequartiere - Dachböden, Hohlräume von Zapfenlöchern des Dachgebälk
Winterquartiere (Keller, Stollen, Höhlen)
Sommerquartier
Paarungsquartiere
Tagesquartier
Wochenstuben
Spalten
Baumhöhlen
Bruthöhlen
Höhlung
„Schwärmen" erhöhte Flugaktivität von Fledermäusen vor Höhlen und Stollen als Balz- und Paarungsplatz -> Informationsaustausch

Lebensraum Fledermaus
baumbewohnende Fledermausart
gebäudebewohnende Fledermausart
Siedlungsbereich (Parks, Gartenanlagen, Friedhöfe, Obstbaumanlagen, Obstbaumwiesen, Kirchen, unsanierte Dachstühle und Stallungen)
Jagdgebiete (Wälder, Obstwiesen, Gebüschgruppen, Hecken, insektenreiche Wiesen)-> Linienförmige Strukturen als Leit- und Jagdstrukturen, Gewässer, Straßen und auch Bahnanlagen, Alleen, Hecken
Leitlinien, Leitstrukturen
Biotopverbund

Erhaltungsmaßnahmen
Suche und Schutz von Quartieren
Schutz der Kolonien durch Verwendung fledermausfreundlicher Holzschutzmittel
Erhaltung einer kleinräumigen Kulturlandschaft als Jagdgebiet
Vermeidung der Anlage von Windenergieanlagen in potentiell oder tatsächlich von der Alpenfledermaus besiedelten Gebieten -> nicht nur für Alpenfledermäuse
Zulassen der Einrichtung von Fahrradwegen in Eisenbahntunneln nur dann, wenn diese nachweislich nicht von Fledermäusen als Sommer-, Winter- oder Paarungsquartier genutzt werden
Schaffung von Ersatzquartieren, Fledermausgerechte Sanierung von Bauwerken, Fledermausbunker, „Umhängen“ von Baumhöhlen (Erprobung), Anlegen von Leitlinien zu isolierten Jagdgebieten, Fledermausbretter, Erhalt von Totholz im Wald, Förderung von Laubgehölzen, Förderung und Erhalt von Waldrändern, Flachkästen, Fledermaushohlkästen

Anlagen
Wasserbau
Stauwehr
Triebwerkskanal
(Kläranlage)
Stau- und Hochwasserschutzanlagen
Hochwasserschutz

Damm
Deich
Deichschutzmaßnahmen - Schutz vor Uferabbrüchen, Deichdurchbrüchen
Einbau von Spundwänden an Deichen
Einbau von Dichtwänden
Einbau von Kiessperren
Sicherung mit Steinschüttung
Sicherung mit aufgelegten Gittern

Fördermöglichkeiten
Bund
Blaues Band Deutschland
Erprobungs- und Entwicklungsvorhaben
Forschungs- und Entwicklungsvorhaben
Wasserwirtschaftliche Unterhaltung
Ökologische Durchgängigkeit

Bundesland
Bayern (BY)
Bayerisches Vertragsnaturschutzprogramm (VNP)
Bayerisches Kulturlandschaftsprogram (KULAP)
Bayerischer Naturschutzfonds
Wasserwirtschaftsämter
Ersatzzahlungen
Landschaftspflege- und Naturparkrichtlinien (LNPR)
EU
LIFE

Eingriff, Beeinträchtigung, vorhabenbedingte Beeinträchtigungen

Gewässerarten
Flachwasserzone
Teich
Weiher
See
Kanal
Fluss
Graben
Still und Fließgewässer
Gewässer I. und II. Ordnung ggf. III. Ordnung
Meere (Nord- und Ostsee)
Ggf. Grundwasser
Oberflächenwasser

Vorhaben
Planfeststellungspflichtige Vorhaben

Plangenehmigungen

Organisationen
- Vorhabensträger
- Behörden
- Naturschutzverbände, Umweltschutzverbände
- IUCN, International Union for Conservation of Nature
- Landesamt für Umwelt
- Gesundheit und Verbraucherschutz LUGVLfU
- Genehmigungsbehörden
- Anhörungsbehörden
- Wasserwirtschaftsamt
- Landesumweltamt
- untere Naturschutzbehörde (uNB)
- EU

- MLUL (Ministerium für ländliche Entwicklung und Umwelt)
- Umweltstiftungen und Verbände (NABU, BUND, etc.)
- Forschungseinrichtungen (Helmholz-Zentrum)

Rechtlich
- Gesetze
  - Bundesrecht
    - Naturschutzgesetz, BNatSchG
    - Bundesartenschutzverordnung, BArtSchV
    - Tierschutzgesetz
    - Bundeswasserstraßengesetz (WaStrG)
    - Verwaltungsverfahrensgesetz (VwVfG)
    - Wasserhaushaltsgesetz (WHG)
    - Umweltgesetzbuch (UGB)
    - Umweltgesetzbuch-Entwurf (UGB-E)
    - Umweltgesetzbuch-Kommissionsentwurf
    - (UGB-KomE) Umwelt- und Planungsrecht (UPR) (Zeitschrift)
    - Verwaltungsverfahrensgesetz
    - Planungsvereinfachungsgesetz (PlVereinhG)
      - Planfeststellungsverfahren §72-78
    - Gesetz über die Vermeidung und Sanierung von Umweltschäden
    - Gesetz über die Umweltverträglichkeitsprüfung (UVPG)
    - Umweltschadensgesetz
    - Bundesbodenschutzgesetz (BBodSchG)

  - Landesrecht
    - Artenschutz-Zuständigkeitsverordnung
    - untere Naturschutzbehörde (uNB)

- Richtlinie
  - Fauna-Flora-Habitat-Richtlinie, FFH

FFH-Monitoring
Europäische Vogelschutzrichtlinie
Umweltverträglichkeitsprüfungs-Richtlinie (UVP-RL),
Richtlinie 2003/35/EG des Rates vom 26. Mai 2003, ABl L 156/17
(alt) Richtlinie 85/337/EWG des Rates vom 27. Juni 1985 über die Umweltverträglichkeitsprüfung bei bestimmten öffentlichen und privaten Projekten
WRRL, Richtlinie 2000/60/EG zur Schaffung eines Ordnungsrahmens für Maßnahmen der Gemeinschaft im Bereich der Wasserpolitik - Wasserrahmenrichtlinie
Umweltinformationsrichtlinie

Verordnungen
Verwaltungsgerichtsordnung (VwGO)
Europarecht, Biber, Zugriff, Verbot Verordnung Nr. 3254/91 (Tellereisen-VO)
Internationales Recht, Biber - Berner Konvention
Bundeskompensationsverordnung (BKompV) ?

Leitfäden
Leitfaden Artenschutz und Gewässerunterhaltung
Leitfaden Artenschutz bei Aus- und Neubau von Bundeswasserstraßen
Leitfaden zur FFH Verträglichkeitsprüfung beim Aus- und Neubau von Bundeswasserstraßen
Leitfaden Artenschutz- und Umweltschadensrecht
Leitfaden für Schutzmaßnahmen an Straßen für Amphibien
z.T. spezifische Leitfäden für Vögel / Fledermäuse
z.T. für einzelne Arten (Rotmilan, Mopsfledermaus, Uhu,…)Leitfäden BfG allgemein

Erlass
Fischottererlass
Bibererlass
…

Handlungsempfehlungen
Handlungsempfehlung zum Umgang mit Fledermäusen bei der Planung und Genehmigung von Windenergieanlagen in Brandenburg
Handlungsempfehlungen zur Erhaltung der lokalen Population des Fischotters
…

Arbeitshilfen
Arbeitshilfe Mopsfledermaus
Arbeitshilfe Artenschutz
Arbeitshilfe Fledermäuse
Arbeitshilfe Fischaufstiegsanlagen an BWaStr

…
Ökologische Durchgängigkeit an Bundeswasserstraßen

Textsammlungen
Textsammlung Naturschutzrecht

Gerichte, Gerichtsbarkeiten
Verwaltungsgericht (VG)
Verwaltungsgerichtshof (VGH)
Bundesverwaltungsgericht
Landesrecht
Oberverwaltungsgericht (OVG) - Revisionsinstanz
(BY, BW, HE,AT) Verwaltungsgerichtshof , OVG
Verwaltungsgericht - Erstinstanz

Revisionssenatsnummer
Revisionssenatsnummer - 4 Bau, Boden, Raumordnung, Naturschutz
Revisionssenatsnummer - 7 Umweltrecht, Atom, Wasser- und Wasserverkehrsrecht
Revisionssenatsnummer - 9 Straßen - und Wegerecht Revisionssenatsnummer - 10 Kommunalrecht

Revisionsinstanz
Sprungrevision

Rechtsbehelf
Rechtsmittel
Berufung, Appellation
Revision
Anschlussrevision
Förmliche Rechtsbehelfe
Einspruch/Widerspruch/Einsprache
Beschwerde
Antrag auf richterliche Entscheidung

Aktenzeichen
European Case Law Identifier (ECLI)

Verfahrensart Registerzeichen
A erstinstanzliches Klageverfahren
B Nichtzulassungsbeschwerde
C Revision in Verwaltungsstreitsachen
PKH Anträge Prozesskostenhilfe

Recht auf Zugang zu einem gerichtlichen Überprüfungsverfahren für Entscheidungen über die Zulassung von Vorhaben

Klagebefugnis
Verbandsklagebefugnis
Klagen

Anfechtungsklage,
Mangel
Verbandsklage, Verbandsklagebefugnis Umweltschutzverband
Beteiligungsklage
Bund-Länder-Streit
Klage gegen naturschutzrechtliche Befreiungen
Klage gegen Planfeststellungsbeschlüsse
Klage gegen Plangenehmigungsverfahren mit Öffentlichkeitsbeteiligung
Umweltschadensklage
Zulässigkeit der Umweltschadensklage
Begründetheit der Umweltschadensklage

Zulässigkeitsvoraussetzung
Vorliegen eines Verstoßes gegen Naturschutzrecht
Berührung des satzungsgemäßen Aufgabenbereichs - Vermeidungsgebot
Mitwirkung im vorangegangenen Verwaltungsverfahren - Kompensationspflicht
Artenschutz nach Art. 12 ff. FFH-RL - Ersatzzahlungen
Monitoring- und Berichtspflichten nach Art.11 und Art.17 FFH-RL - Bevorratung von Kompensationsmaßnahmen
Umweltschaden?

Entscheidungstyp
Beschluss
Urteil
Kommentar

Planfeststellungsbeschluss PFB, wasserrechtlicher
Planfeststellungsbeschluss/-verfahren gemäß § 29 I Nr. 4 BNatSchG a.F.
Antragsbefugnis, gerichtliche Kontrollbefugnis
Planrechtfertigung, Prognose, Abwägungsfehler, vorläufiger Rechtsschutz

Einstweilige Anordnung
Baustopp verhängt
Beteiligungsrechte des Landesnaturschutzverbandes
Rahmenbetriebsplan
Planfeststellungsergänzungsbeschluss

Verfahren der Planfeststellung -Schritte
Planerstellung
Einreichen des Planes bei zuständiger Anhörungsbehörde (§73 Abs.1 VwVfG)
Öffentliche Auslegung (§73 Abs. 3 VwVfg)
Frist: innerhalb 3 Wochen für die Dauer eines Monats (Auslegungsfrist)
Betroffene können Einwendungen einreichen
Gelegenheit zur Planungseinsicht
Ablauf der Einwendungsfrist
Präklusionswirkung
Veränderungssperre

Erörterungstermin
Anhörungsverfahren (§ 73 Abs. 2 VwVfG)
Erörterung (§ 73 Abs. 6 VwVfG)
Weiterleitung der Anhörungsergebnisse (§ 73 Abs.9 VwVfG)
Planfeststellungsbeschluss (§ 74 VwVfG)
Umfassendes Planungsermessen
Abwägungsgebot
Durchführung des Plans
Unanfechtbarkeit
Rechtswirkung

Erhebliche Gewässerschäden gemäß§ 90WHG
Umweltschadensantrag

Fachplanung
Raumbedeutsame Planung
Verkehrsplanung
Verkehrsträger
Umweltplanung
Wasserrechtlicher Planung
Bauleitplan
Einwendungsausschluss
Beteiligungsverfahren
Wahl des Beteiligungsverfahrens
Umgehung des Beteiligungsverfahrens

Erhaltungszustand
Standarddatenbogen (SDB)

ungünstig-unzureichend , U1
Sukzession
Besonderer Artenschutz
Artenschutzrechtlicher Fachbeitrag (AFB)
Landschaftspflegerischer Begleitplan (LBP)
Umweltverträglichkeitsstudie (UVS)
Umweltverträglichkeitsuntersuchung (UVU)
Natura 2000-Verträglichkeitsuntersuchungen (FFH-, SPA-VU)
Landschaftspflegerische Ausführungsplanung (LAP)
Ausführungsplanung
(artenschutzrechtliches) Zugriffsverbot
Ausnahmetatbestand
Befreiungstatbestand

LANA 2010
BVerwG 2011
Zugriffsverbot
Tötungsverbot

Beschädigungs-, Zerstörungs-, Entnahme-, Verletzungs- oder Fangverbot
Störungsverbot
Verbot der Beschädigung des Standortes von Wildpflanzen
Besitz- und Vermarktungsverbote

Baubedingter Verlust
Anlagebedingte Verluste
Betriebsbedingter Verlust
Bauzeitliche, temporär
dauerhaft
Signifikant erhöhtes Tötungsrisiko
Störungsverbot
Vermeidung
Vermeidungsgebot
Vermeidungsmaßnahme
Minderung
Minderungsmaßnahme
Querungshilfe
Lärmschutzwall
Baugrubensicherung

Bauzeiten unter Berücksichtigung der Phänologie
Bauzeitenregelung
Vergrämung, Vergrämungsmaßnahme
Ausnahmegenehmigung

Entwicklungszeit Habitat
Prädationsrisiko
Prädator
Verbotswidrige Beeinträchtigung
Risikobewertung
Kartierung
Potenzialabschätzung
Eingriffsregelung
Legalausnahme
Ausnahme, artenschutzrechtliche Ausnahme
Zumutbare Alternative
Öffentliches Interesse
Erhaltungszustand (keine Verschlechterung)
Eignung
Eingriffsintensität
Naturschutzfachliche Bewertung
Ampel-Bewertungsverfahren
Erhaltungszustand,
Lokal
Regional
Überregional
Wirksamkeit (von Vermeidungsmaßnahmen)

Verbotstatsbestand
Zulassungsentscheidung
Genehmigung
Ausnahmegenehmigung
Genehmigung zum Aussetzen im Rahmen einer Umsiedlung
Gefährdung von Biotopen, Ökosystemen, Arten
Verfälschung
Gefährdung einzelner Populationen
Vorlauf
Bauvorbereitende Maßnahme
Zwischenhälterung
Schutzkonzept
Vermeidungs- und Verminderungsmaßnahmen
Minimierungsmaßnahme
CEF-Maßnahmen
Artenschutzfachliche Ausgleichsmaßnahmen
Erfolgskontrolle
Monitoring
Risikobewertung

Fachlich

Literatur
Blanke 2010

Maßnahmen
Ersatzhabitat
Lebensraum-Vergrößerung
Lebensraum-Neuschaffung
Ausgleichsmaßnahme
Vorgezogene Ausgleichsmaßnahme CEF-Maßnahme
Funktionserhaltende Maßnahme,
FCS-Maßnahme, Maßnahme zur Sicherung des Erhaltungszustands der Population
Wiederherstellung
Wiederbesiedlung, Wiedereinbürgerung
Umsiedlung
Fang (Hand-, Schlingenfang…)
Risikomanagement, Monitoring
Vorübergehende Verschlechterung
Fangen
Umsetzen
Umsiedlung, Translocation
Zielstandort
Anforderungen
Verhinderung Abwanderung, Zäunung
Zwischenhälterung,
Gehege

Umsetzung, Verlagerung
Maßnahmenerfolg
Korrekturmaßnahme
Vorsorgemaßnahme
Ersatzmaßnahme
Kompensationsmaßnahme
Schutzmaßnahme
Gestaltungsmaßnahme
Kohärenzsicherungsmaßnahme
Schadensbegrenzungsmaßnahmen
Ersatzgeldzahlung
Ökokonto
Flächenpool
Produktionsintegrierte Kompensationsmaßnahmen (PiK)

Kartierung, Erfassung
Begehung
Sichtbeobachtung
Verhören
Kescherfang

Bewertung
Zufallsfunde
Potenzial
Habitatqualität

Erfolgskontrolle, Monitoring
Korrigierende Maßnahmen
Zwischenbericht
Endbericht
Plankontrolle

Schutzstatus
Arten des Anhang II, Arten des Anhang IV
Streng geschützte Art
Besonders geschützte Art
Rote Liste
Charakteristische Art

Geoentitäten
Baden-Württemberg,
Bayern
Etc.
Natura2000-Gebiete
Schutzgebiete
Naturräume

Landschaften
Kulturlandschaften
Regional bedeutsame Landschaften
Kreis

Gemeinden

D. Referenzen

Leistungsbeschreibung für die Vergabe einer Machbarkeitsstudie zur Entwicklung eines Demonstrators für eine Wissensplattform im Umweltbereich, Bundesministerium für Verkehr und Digitale Infrastruktur, Stand 17.05.2019

AP1 Bericht - Workshop zur Ermittlung der Nutzeranforderungen 09.09.2019, Condat AG, Stand 16.10.2019

AP2/AP3 Bericht - Workshop zur Präsentation und Diskussion des Zwischenstands 12.11.2019, Condat AG, Stand 19.12.2019

BMVI Machbarkeitsstudie 2019

Aufbau einer Wissensplattform im Umweltbereich zur Planungsbeschleunigung von Infrastrukturmaß-nahmen

Abschlussbericht
Teil B: IT-Anforderungen

Stand: 31.03.2020

Auftraggeber
Bundesministerium für Verkehr und digitale Infrastruktur (BMVI)
Referat G10
Invalidenstr. 44
10115 Berlin

condat®

Inhaltsverzeichnis

1 Ziele des Systems 3
2 Funktionale Anforderungen 3
2.1 Suche von Informationen 3
2.1.1 Allgemeine Anforderungen 3
2.1.2 Suche im Internet 3
2.1.3 Suche in bekannten Quellen 5
2.1.4 Suche in bzw. nach nicht digital verfügbaren Quellen 5
2.2 Verarbeitung der Informationen 6
2.2.1 Relevanzbewertung 6
2.2.2 Anreicherung mit Metadaten 7
2.2.3 Analyse 7
2.3 Speicherung der Informationen 8
2.4 Benutzeroberfläche 8
2.5 Nicht funktionale Anforderungen 9
2.5.1 Usability und Softwareergonomie 9
2.5.2 Barrierefreiheit 9
2.5.3 Browserkompatibilität 10
2.5.4 Verfügbarkeit 10
2.5.5 Überwachung 10
2.5.6 Mengengerüste 10
2.5.7 Performanz 11
2.5.8 Modularität 12
2.5.9 Erweiterbarkeit 12
2.5.10 Wartbarkeit 12
2.5.11 IT-Sicherheit / IT-Grundschutz 12
2.5.12 Datenschutz 12
2.5.13 Möglichkeit der Regelungen des Zugriffschutzes, Rollen- und Benutzerkonzept 13
3 Technische Randbedingungen 13
3.1 IT-Architektur 13
3.1.1 Web-Applikation 13
3.1.2 Style-Guide 13
3.1.3 Technologien 13

3.2 Betriebsinfrastruktur 13
3.3 Integration Active Directory und SSO 14
3.4 IT-Arbeitsplätze 14

1 Ziele des Systems

Das System soll einschlägige, rechtssichere und zitierfähige umwelt-, natur- und artenschutzrechtlicher Informationen für die Verwendung bei Planungen und Genehmigungen von Verkehrswegen ermitteln und den Nutzenden bereitstellen. Dabei soll das System den Suchaufwand im Abgleich zu einer händischen Suche reduzieren, einen besseren Überblick über die relevanten Daten gewährleisten, eine umfassende Entscheidungsgrundlage schaffen und mögliche Probleme beim Umgang mit gefährdeten Arten frühzeitig erkennbar machen.

2 Funktionale Anforderungen

2.1 Suche von Informationen

Das System muss sowohl Dokumente aus (noch) unbekannten Quellen im Internet als auch aus bekannten Quellen ermitteln und passend zur Suche durch die Nutzenden aufbereiten und bereitstellen.

2.1.1 Allgemeine Anforderungen

Sicherstellung einer möglichst langfristigen Verfügbarkeit der gefundenen Informationen und eines langfristigen Zugangs zu diesen (im Hinblick ggfs. eines Aufbaus einer eigenen Datenbank oder/und einer Verknüpfung zu bestehenden Datenbanken).

2.1.2 Suche im Internet

Das System muss eine automatisierte Suche nach relevanten Informationen im Internet realisieren.

Hierbei sind folgende Eigenschaften zu erfüllen:

- Die Suche muss vollständig sein. Alle relevanten im Internet durch eine manuelle Suche über Suchmaschinen wie Google, Bing, etc. auffindbaren Informationen müssen auch über die automatische Suche durch die Wissensplattform gefunden werden. Dies beinhaltet z.B. für deutschsprachige Quellen auch Quellen aus der Schweiz, Österreich, etc..

- Die Suche muss automatisch in regelmäßigen Abständen (vorzugsweise nächtlich) durchgeführt werden.

- Die Suche muss das Finden aller relevanten Dokumente zu allen geschützten Tier- und perspektivisch auch Pflanzenarten zum Ergebnis haben. Hierbei ist zu beachten, dass der Schutzstatus zeitlich und regional unterschiedlich sein kann und auch jederzeit veränderbar ist. Dies sollte, wenn möglich, automatisch von der Wissensplattform angepasst werden, zumindest aber durch eine administrative Anpassung erfolgen können.

- Ausgangspunkt für die Auswahl der Schlagworte, die in den automatisierten Suchen verwendet werden, sollte darüber hinaus das als Ergebnis der Machbarkeitsstudie entstandene Vokabular sein.

- Die in der Suche angewandten Schlagworte müssen administrativ gepflegt werden können.

- Die Suche sollte eine Blacklist von gesperrten bzw. zu ignorierenden Webseiten berücksichtigen können. Die Blacklist muss administrativ gepflegt werden können.

  Die Suche kann unter Verwendung (kommerzieller) Suchmaschinen wie z.B. Google Search API oder Microsoft Bing Search API realisiert werden.

- Eine Erweiterung der Suche von Dokumenten in anderen Sprachen sollte möglich sein.

2.1.3 Suche in bekannten Quellen

Das System muss die Suche(n) in bekannten Quellen durch

- Crawling und Scraping oder
- dedizierte Anbindungen (z.B. via API)

unterstützen.

- Es sollen alle bekannten Quellen angebunden werden. Falls es sich hierbei um zugriffsgeschützte und ggf. kostenpflichtige Quellen handelt, müssen die Quellen angebunden werden, sofern der Auftraggeber entsprechende Zugänge bereitstellt.

- Die Liste der Quellen sollte kontinuierlich weiterentwickelt werden. Die Anbindung von neuen Quellen muss möglich sein und per Konfiguration bzw. Administration unterstützt werden.

2.1.4 Suche in bzw. nach nicht digital verfügbaren Quellen

Das System soll perspektivisch auch das Finden nicht digital verfügbarer Informationen unterstützen, wie z.B. Artikel aus Fachzeitschriften, die nur im Print-Format vorliegen. Um dies zu ermöglichen, soll das System folgende Möglichkeiten bieten:

- Finden möglicher nicht-digitaler Quellen durch die entsprechende Verarbeitung und Auswertung der digital gefundenen Quellen, d.h. insbes. Erkennen und Verarbeiten von Quellenangaben, Literaturhinweisen in Fachartikeln, Kommentaren zu Gerichtsurteilen etc. Diese nicht-digitalen Quellen sollen zum einen als Grundlage für eine verfeinerte Suche im Internet dienen. Zum anderen sollen Hinweise auf nicht digitale Quellen durch die Wissensplattform den Nutzenden zur Verfügung gestellt werden.

- Integration von eingescannten Dokumenten in die Plattform, d.h. es ist eine Upload-Möglichkeit für die Administratoren des Systems mit anschließender OCR-Erkennung und Verarbeitung (soweit technisch möglich) zu ermöglichen, so dass das Dokument dann auch digital in der Wissensplattform verfügbar ist.

- Die Upload-Möglichkeit soll das Hochladen einer Menge von Dokumenten, z.B. von allen Dokumenten eines Verzeichnisses (z.B. eines Dateiordners im

Windows Explorer) durch die Administratoren des Systems unterstützen.

- Es muss gewährleistet sein, dass die entsprechenden Rechte für die Upload-Dokumente vorliegen. Dies könnte bspw. durch einen Warnhinweis seitens des Systems unterstützt werden, bzw. vom Administrator digital bestätigt werden müssen.

2.2 Verarbeitung der Informationen

Das System muss im Internet gefundene und aus bekannten Quellen ermittelte Informationen mit Hilfe von KI-Methoden geeignet verarbeiten und aufbereiten, um die Informationen dann den Nutzenden in der bestmöglichen Weise zur Verfügung stellen zu können.

Als Quellenformate müssen alle üblichen Formate, wie HTML, PDF, Microsoft Office, aber auch strukturierte Formate über APIs oder z.B. aus Linked Open Data-Quellen unterstützt werden.

Die Verarbeitung umfasst dabei zumindest die folgenden Schritte:

2.2.1 Relevanzbewertung

Die über die normale Internet-Suche gefundenen Dokumente müssen vor dem Import einer Relevanzbewertung unterzogen werden, um sicherzustellen, dass nur relevante Dokumente in die Wissensplattform übernommen werden.

Z.B. soll verhindert werden, dass bei der Suche nach „Fledermaus“ Suchergebnisse aufgenommen werden, die von der Operette von J. Strauß handeln. Gleichzeitig dürfen aber relevante Suchergebnisse nicht herausgefiltert werden.

Durch Anwendung von selbstlernenden KI-Verfahren sollte dieses Verfahren nach einer initialen Trainingsphase selbständig das optimale Verhältnis zwischen Anzahl („Recall“) und Genauigkeit der Ergebnisse („Precision“) finden und sich laufend verbessern.

Dieses Verfahren muss hierbei gleichzeitig transparent in dem Sinne sein, dass ausgeschlossene und eingeschlossene Suchergebnisse vom System protokolliert werden, sodass seitens der Systemadministratoren bzw. Systemadministratorinnen (oder Nutzenden) regelmäßig überprüft werden kann, ob evtl. relevante Ergebnisse

ausgeschlossen wurden, bzw. umgekehrt. Das Ergebnis muss dann jeweils in das KI-Verfahren einfließen, sodass dessen Berechnung mit der Zeit immer exakter wird.

Als Ausgangspunkt für die Relevanzbewertung kann die als Ergebnis der Machbarkeitsstudie entstandene Ontologie dienen (bzw. das Vokabular).

2.2.2 Anreicherung mit Metadaten

Alle Informationen, die als relevant bewertet wurden und in die Wissensplattform importiert werden, sollen für die weitere Analyse automatisch mit Metadaten angereichert werden. Folgende Metadatenanreicherungen sind sinnvoll:

- **Regionalzuordnung**: Für jedes Dokument mit einem klaren Regionalbezug sollte dieser automatisch dokumentiert werden; für Quellen aus dem Bereich der Bundesrepublik Deutschland ist dies das jeweilige Bundesland, ansonsten das Land (z.B. Schweiz, Österreich) .

- **Kategoriezuordnung**: Die einzelnen Quellen sollten unterschiedlichen inhaltlichen Kategorien zugeordnet werden können, z.B. den Kategorien Planung, Rechtliches, Fachinformation. Das System soll diese Zuordnung durch Anwendung von KI-Verfahren automatisch vornehmen. Die Festlegung der Kategorien soll flexibel erfolgen können. Hierbei wären für die jeweilige zu bildende Kategorie z.B. exemplarische Dokumente seitens des Auftraggebers bzw. der Nutzenden zu definieren („Trainingsset“), die sodann als Ausgangspunkt für das Training der automatischen Kategoriezuordnung dienen können. Dokumente sollen hierbei auch mehreren Kategorien zugeordnet werden können, sofern sie thematisch unter mehrere Kategorien fallen. Dabei ist ein unterschiedlicher Rangplatz in der Sortierung in den verschiedenen Kategorien möglich.

Weitere sinnvolle Metadatenanreicherungen sind denkbar und im Entwicklungsprozess ggf. entsprechend zu berücksichtigen.

2.2.3 Analyse

Alle in die Wissensplattform importierten Dokumente müssen einer textuellen und inhaltlichen Analyse unterzogen werden. Hierbei sollten Verfahren aus dem Maschinellen Lernen (ML), der KI und der Natürlichsprachigen Verarbeitung (NLP) Anwendung finden.

Die wesentlichen Ziele der Analyse sind:

- Extraktion der relevanten Informationen aus den Dokumenten
- Verknüpfung der Informationen und Quellen untereinander, sodass seitens der Nutzenden durch die Informationen der Wissensplattform navigiert werden kann und die Informationen in strukturierter Weise dargestellt werden können.

Als Verfahren sollte hierbei, wie im Rahmen der Machbarkeitsstudie geschehen, eine „Semantische Indizierung“ mittels „Named Entity Recognition“ erfolgen. Grundlage der Semantischen Indizierung ist wiederum die spezielle Domänen-Ontologie, die als Ergebnis der Machbarkeitsstudie entstanden ist. Hierbei sollten alle Begriffe und Beziehungen, die in dieser Ontologie definiert wurden, erkannt und den jeweiligen Textstellen zugeordnet werden.

Das Ergebnis der Analyse sollte entsprechend der Erkenntnisse der vorliegenden Machbarkeitsstudie der Aufbau eines „Knowledge Graph“ sein, der die einfache Definition von Abfragen und die automatische Generierung strukturierter Informationen zu jeder geschützten Tier- und perspektivisch zu jeder geschützten Pflanzenart erlaubt. Diese Art und Weise des Vorgehens sollte es z.B. auch ermöglichen, alle Dokumente zu einem bestimmten Vorhaben zusammenzustellen und dem Nutzenden auf der Benutzeroberfläche zur Verfügung zu stellen. Hierbei ist wesentlich, dass die Abfragen nicht vorher abschließend definiert werden können, sondern diese im späteren Betrieb dynamisch und flexibel erstellbar sein müssen.

2.3 Speicherung der Informationen

Die gefundenen relevanten und aufbereiteten Informationen müssen in geeigneter Weise durch das System gespeichert werden, um sie den Nutzenden über die Benutzeroberfläche insbesondere mittels einer Suchfunktion bereitzustellen. Im Rahmen der Machbarkeitsstudie und der Entwicklung des Demonstrators erfolgte die Speicherung der Daten in einer MongoDB und in Elasticsearch (s.a. Abschnitt 3.1 Abschlussbericht Teil A – Systembeschreibung).

2.4 Benutzeroberfläche

Das System muss eine Benutzeroberfläche bereitstellen, die den Nutzenden auf intuitive Weise ein effizientes, zielgerichtetes Finden von Informationen ermöglicht:

- Die Benutzeroberfläche muss eine Suchfunktion über die im System gespeicherten, konsolidierten und aufbereiteten Informationen bereitstellen.

- Die Benutzeroberfläche soll eine Suchfunktion für die Live-Suche im Internet (z.B. über Google oder Bing) bereitstellen.

- Der Aufbau und die Gestaltung der Benutzeroberfläche sollen einen explorativen Suchprozess ermöglichen.

- Die Benutzeroberfläche muss konfigurierbar und personalisierbar sein.

- Es soll ein Feedback-Mechanismus bzgl. der Relevanz von Informationen realisiert werden.

2.5 Nicht funktionale Anforderungen

2.5.1 Usability und Softwareergonomie

Das wesentliche Ziel bei der Gestaltung der Benutzerschnittstelle ist die optimale Unterstützung des explorativen Suchprozesses.

Um dieses Ziel zu erreichen, sollte der Designprozess iterativ unter Beteiligung von Nutzenden erfolgen. Die Einbeziehung möglicher Nutzenden in den Entwicklungs- und Designprozess hat sich im Rahmen der Machbarkeitsstudie als sinnvoll erwiesen.

Bei der Entwicklung des Systems muss allgemein den Anforderungen an die Gebrauchstauglichkeit nach Norm EN ISO 9241 Teil 11, den Grundsätzen der Dialoggestaltung nach EN ISO 9241-110 und den Anforderungen an die Informationsdarstellung nach EN ISO 9241-12 entsprochen werden.

2.5.2 Barrierefreiheit

Das System muss die Kriterien der Barrierefreiheit nach der „Barrierefreie Informationstechnik-Verordnung“ (BITV) erfüllen[1].

1 Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung - BITV 2.0), https://www.gesetze-im-internet.de/bitv_2_0/index.html

2.5.3 Browserkompatibilität

Das System muss mit den aktuellen, marktgängigen Internet-Browsern nutzbar sein, die in den Vorschriften für den Bundesclient freigegeben sind.

Der im Rahmen der Machbarkeitsstudie bereitgestellte Demonstrator wurde für die Nutzung in den Browsern Google Chrome, Mozilla Firefox und Microsoft Edge entwickelt.

2.5.4 Verfügbarkeit

Das System muss hoch verfügbar sein: Seine Funktion darf zur Hauptbetriebszeit nur minimal oder innerhalb festgelegter Zeiten unterbrochen werden.[2]

2.5.5 Überwachung

Das System muss für seine Architektur und Komponenten die Einbindung in gängige Überwachungswerkzeuge (z.B. Nagios) ermöglichen.

2.5.6 Mengengerüste

Das System muss hinsichtlich seiner interaktiven Komponenten (Benutzerschnittstelle) für eine noch zu definierende Vielzahl gleichzeitiger Benutzenden ausgelegt sein. Der Demonstrator war für keine definierte Nutzeranzahl ausgelegt, da er lediglich funktionale Anforderungen demonstrieren und ansonsten keinerlei definierte nicht-funktionale Anforderungen erfüllen musste.

Das System muss hinsichtlich der Anbindung von Datenquellen (periodische Internet-Suchen, Crawling und Scraping, dedizierte Anbindung über APIs) und der Vorverarbeitung der importierten Dokumente horizontal linear skalierbar sein.

Für die Datenhaltung des Systems müssen mindestens folgende Mengengerüste unterstützt werden:

[2] AEC-2 („High Availability“) nach Harvard Research Group-Klassifizierung, https://de.wikipedia.org/wiki/Hochverf%C3%BCgbarkeit#Availability_Environment_Classification

- Metadaten zu Quellen, Suchergebnissen und Dokumenten: Mindestens 25 Mio. Dokumente

- Volltexte zu heruntergeladenen Dokumenten: Mindestens 5 Mio. Dokumente[3]

- Semantischer Index: Mindestens 5 Mio. Dokumente

Es ist nicht ausgeschlossen, dass die Mengengerüste deutlich höher sein werden.

2.5.7 Performanz

Das System **soll** für seine interaktiven Komponenten (Benutzerschnittstelle) folgendes Antwortzeitverhalten[4] aufweisen:

- <= 1 Sekunde
  - Einfache Interaktionen und Dialoge
  - Anpassungen an der Struktur oder Darstellung der Benutzerschnittstelle.
- <= 4 Sekunden (ggf. geeignete Anzeige eines Statusindikators)
  - Komplexe Interaktionen (exkl. Suche)
  - Detailansicht eines Dokuments
  - Einfache Suche in der Wissensplattform
- <= 10 Sekunden (geeignete Anzeige eines Status- und Fortschrittindikators)
  - Komplexe Suche in der Wissensplattform
  - Einfache synchrone Internet-Suche
- <= 60 Sekunden (geeignete Anzeige eines Status- und Fortschrittsindikators)
  - Komplexe Internet-Suche bzw. Internet-Suche mit synchroner komplexer semantischer Anreicherung der Suchergebnisse (Aufgrund der Erfahrungen im Rahmen der Machbarkeitsstudie wird davon ausgegangen, dass ein Antwortzeitverhalten von unter 10 Sekunden mit angemessenem Aufwand nicht realisierbar ist. Um ein relevantes Ergebnis zu erhalten, sollten mindestens die ersten 1.000 Treffer einer Internet-

[3] Berechnung: Rund 500 geschützte Arten, mindestens 10.000 Dokumente je Art. Der Dokumentenbestand des Demonstrators beträgt rund 70.000 Dokumente für 10 Arten.

[4] Gesamtantwortzeit aus Benutzersicht (d.h. im Browser).

Suche verarbeitet werden. Die Anreicherung dauert dann entsprechend. Im Demonstrator wurde die Verarbeitung aus Performance-Gründen auf 50 oder 100 Ergebnisse beschränkt. Da die Internet-Suche nur eine von mehreren parallelen Suchen ist und die Anreicherung für den Nutzenden entsprechenden Mehrwert liefert, ist eine längere Antwortzeit hinnehmbar. Dabei sollte, wie gefordert, eine geeignete Anzeige eines Status- und Fortschritt-Indikators erfolgen.)

2.5.8 Modularität

Das System sollte komponentenbasiert entwickelt werden und offene, wohldefinierte Schnittstellen nutzen. Die Komponenten sollten austauschbar sein.

2.5.9 Erweiterbarkeit

Das System soll für seine Architektur und Komponenten über den Entwicklungszeitraum und Lebenszyklus eine Erweiterbarkeit zum Hinzufügen neuer Funktionen oder eine Anpassung bestehender Funktionen über definierte und offene Schnittstellen im Sinne einer evolutionären Entwicklung ermöglichen.

2.5.10 Wartbarkeit

Das System soll für seine Architektur und Komponenten eine gute Wartbarkeit aufweisen. Dies soll durch Dokumentation, Spezifikation von Schnittstellen, Modularität, Parametrisierbarkeit auch im Hinblick auf eine längere Verwendungsdauer unter sich fortentwickelnden Anforderungen aufrechterhalten werden.

2.5.11 IT-Sicherheit / IT-Grundschutz

Für das System und die enthaltenen, verarbeiteten Daten müssen die Anforderungen des IT-Grundschutzes (siehe www.bsi.bund.de/gshb/) erfüllt werden.

Insbesondere für die aus dem Internet heruntergeladenen Dokumente müssen geeignete Sicherheitsmaßnahmen (u.a. Virenschutz) umgesetzt werden.

2.5.12 Datenschutz

Das System muss den Anforderungen der DSGVO entsprechen.

2.5.13 Möglichkeit der Regelungen des Zugriffschutzes, Rollen- und Benutzerkonzept

Für das System müssen Mechanismen für den Zugriffsschutz und ein anpassbares Benutzer- und Rollenkonzept umgesetzt werden.

Siehe auch Abschnitt 3.3 Integration Active Directory und SSO.

3 Technische Randbedingungen

3.1 IT-Architektur

3.1.1 Web-Applikation

Der Client muss als Web-Applikation realisiert werden.

Die Web-Applikation kann als Single-Page-Anwendung (SPA) oder auch als Progressive Web-Anwendung (PWA) realisiert werden.

3.1.2 Style-Guide

Etwaige Vorgaben im Hinblick auf zu verwendende Style Guides für die Benutzeroberfläche müssen eingehalten werden.

3.1.3 Technologien

Etwaige Vorgaben bzw. Einschränkungen im Hinblick auf Software, Bibliotheken etc. (bspw. technisch, seitens der Lizenzbedingungen) müssen eingehalten werden.

Auch etwaige Vorgaben bzw. Einschränkungen bzgl. der Systemvoraussetzungen müssen eingehalten werden.

3.2 Betriebsinfrastruktur

Etwaige technische Rahmenbedingungen der Betriebsinfrastruktur (On-Premise, Cloud etc.), sind zu beachten. Aufgrund der im Rahmen der Machbarkeitsstudie gemachten Erfahrungen wäre es empfehlenswert, zumindest einen hybriden Ansatz

zuzulassen, da sich unter Umständen mit den Möglichkeiten des Serverless Computing bei Cloud-Dienstleistern wie AWS, Google oder Azure insbesondere aufwändig zu rechnende KI-Funktionen hoch-skalierbar aber kostengünstig realisieren lassen.

3.3 Integration Active Directory und SSO

Das System muss die Anbindung an bestehende Verzeichnissysteme und eine Integration in übergeordnete Infrastrukturen zur Authentisierung und Autorisierung unterstützen.

3.4 IT-Arbeitsplätze

Technische Spezifikation der IT-Arbeitsplätze bei den Nutzenden des Systems müssen im Vorfeld einer Weiterentwicklung des Systems geklärt werden.

BMVI Machbarkeitsstudie 2019

Aufbau einer Wissensplattform im Umweltbereich zur Planungsbeschleunigung von Infrastrukturmaß-nahmen

Dokumentation des Quellcodes

Stand: 31.03.2020

Auftraggeber
Bundesministerium für Verkehr und digitale Infrastruktur (BMVI)
Referat G10
Invalidenstr. 44
10115 Berlin

condat®

Inhaltsverzeichnis

1 Benutzeroberfläche 4
1.1 Quellen 4
1.2 Installation und Betrieb 6
1.2.1 Voraussetzungen 6
1.3 Installation des Quellcodes zum Bearbeiten 7
1.4 Erzeugen der Dateien für die Veröffentlichung 7
2 Backend 8
2.1 Systemvoraussetzungen 8
2.1.1 Software 8
2.1.2 Systemumgebung 8
2.1.3 Verzeichnisstruktur 8
2.1.4 Beschreibung der Dateien 9
2.2 Installationshinweise 12
2.2.1 Systemvoraussetzungen 12

Dokumenthistorie

27.02.2020	V0.1	J. Thomsen (Condat AG)	Dokument erstellt
28.02.2020	V0.2	A. Sarioglu (Condat AG)	Dokumentation der Benutzer-oberfläche
02.03.2020	V1.0	J. Thomsen	Diese Version

1 Benutzeroberfläche

1.1 Quellen

Der Quellcode für die Benutzeroberfläche beinhaltet die im Folgenden aufgeführten Komponenten. Diese sind für den Ablauf der Anwendung verantwortlich. Diese umfassen auch jeweils eine HTML-Datei als Template und SCSS-Datei für die Style-Informationen. Die Benutzeroberfläche wurde auf Basis des Frameworks Angular realisiert.

Die einzelnen Dateien und ihre Funktion werden nachfolgend kurz beschrieben.

`index.html`

Dies ist die Einstiegsseite in die Anwendung. Sie ruft die Hauptkomponente auf.

`src/app/app.component.ts`

Dies ist Hauptkomponente der Anwendung. Sie ruft das Module app-routing auf.

`src/app/app-routing.module.ts`

Leitet weiter auf die SearchComponent.

`src/app/search/search.component.ts`

Beinhaltet die Topsuche und diese ruft die einzelnen Suchen auf.

`src/app/cards/card-title.component.ts`

Kopfbereich der einzelnen Suche.

`src/app/cards/card-body/card-body.component.ts`

Such- und Ergebnisbereich der einzelnen Suchen.

`src/app/cards/card-dialog/card-dialog.component.ts`

Komponente für die Detail- und PDF-Ansicht.

Die Verzeichnisstruktur des Quellcodes ist wie folgt:

```
│   index.html
├───dist
│   └───bmvi
│       └───assets
│           ├───flags
│           └───icons
├───doc
├───e2e
│   └───src
└───src
    ├───app
    │   │   app-routing.module.ts
    │   │   app.component.ts
    │   ├───cards
    │   │   │   card-title.component.ts
    │   │   ├───card-body
    │   │   │       card-body.component.ts
    │   │   └───card-dialog
    │   │           card-dialog.component.ts
    │   ├───core
    │   │       api.ts
    │   │       bundeslaender.ts
    │   │       data.service.spec.ts
```

```
│   │       data.service.ts
│   │       Utils.ts
│   ├───search
│   │       search.component.ts
│   └───shared
│       ├───maintenance-dialog
│       ├───popup
│       ├───searchparams-dialog
│       └───settings-dialog
├───assets
│   ├───flags
│   └───icons
└───environments
```

1.2 Installation und Betrieb

1.2.1 Voraussetzungen

Das Frontend des Demonstrators basiert auf dem Framework Angular in der Version 8.2.14. Für die Installation einer Angular-Arbeitsumgebung wird Node.js (https://nodejs.org/de/) vorausgesetzt. Dies(e?) muss zunächst installiert werden. Die Anleitung zur Installation von Angular befindet sich unter https://angular.io/guide/setup-local.

1.3 Installation des Quellcodes zum Bearbeiten

Im Verzeichnis, in dem sich die Quellcodes befinden, werden durch Aufruf von `npm install` die benötigten Bibliotheken im Verzeichnis node_modules erzeugt.

Zum Bearbeiten des Quellcodes und Betrachten der Anwendung kann eine lokale Instanz mit `npm run start` aufgerufen werden. Diese startet einen lokalen Webserver und öffnet die Anwendung in den voreingestellten Browser. Bei Änderungen am Quellcode werden diese im Browser sofort angezeigt.

1.4 Erzeugen der Dateien für die Veröffentlichung

Durch Aufruf von `npm run buildprod` wird der Quellcode übersetzt und für die Installation auf einem Produktivsystem im Verzeichnis `dist/bmvi` bereitgestellt. Die Dateien, die in dem Ordner `dist/bmvi` erzeugt wurden, können dann in ein Verzeichnis eines Webservers (z.B. Apache, nginx, etc.) kopiert werden.

2 Backend

Das Backend umfasst die Komponenten, die zum Betrieb der prototypischen Implementierung der Wissensplattform Umwelt und für die Interaktion über die Benutzeroberfläche erforderlich sind.

2.1 Systemvoraussetzungen

2.1.1 Software

Das Backend setzt folgende Software-Installationen bzw. vorhandenen Zugänge zu cloudbasierten Diensten voraus. Hierbei werden jeweils die Versionsnummern angegeben, unter denen das aktuelle System läuft. Unter älteren oder aktuelleren Versionen ist das System ggf. ebenfalls lauffähig, müsste aber im Einzelfall getestet werden.

- RabbitMQ Version 3.6.6
- MongoDB Version 3.2.10
- Elasticsearch Version 7.1
- Microsoft Azure Bing API
- Condat Smart Media Engine Release 1.0.1 (Anbindung über AWS - Amazon Web Services)
- Python 3.7+
- Docker

2.1.2 Systemumgebung

Das System läuft bei Condat AG unter Betriebssystem Linux CentOS 7.2. Prinzipiell sollten alle Komponenten auch unter anderen Betriebssystemen lauffähig sein, im Einzelnen können ggf. Anpassungen erforderlich werden.

2.1.3 Verzeichnisstruktur

```
├── <root>
│   │   bmvi_dispatcher.py
│   │   bmvi_es_connector.py
│   │   bmvi_restserver.py
│   │   bmvi_sme_connector.py
│   │   config.py
│   │   requirements.txt
```

```
│   │   startup.sh

├── classes
│   │   RabbitMQ.py
│   │   Search.py
│   │   SMEConnector.py


├── clustering
│    │   wissensplattform-umwelt-1.0.jar
│    │   README-md


├── documenttype_detection
│     | app.py
│     | consts.py
│     | textCategorization.py
│     | webApi.py
│     | words.txt

├── lib
│     | dfki.py
│     | sme.py
│     | utilities.py

├── ontology
│     | bmvi.ttl

├── restapi
│     | dfki.py
│     | sme.py
```

2.1.4 Beschreibung der Dateien

Das Backend besteht im Wesentlichen aus drei Teilen:

- Den externen Komponenten (MongoDB, RabbitMQ, Elasticsearch, etc., wie in 2.1.1 beschrieben)
- Eine Reihe von Programmen, die die Kommunikation mit diesen Komponenten abwickeln; vorwiegend auf Python-Basis, in Einzelfällen auch auf Java-Basis
- Das REST-API, das eine einheitliche Schnittstelle auf Dienste und Daten bereitstellt, insbes. für die Benutzeroberfläche

Die Funktion der einzelnen Dateien wird im Folgenden kurz erläutert

`bmvi_dispatcher.py`

RabbitMQ Consumer Client; zentrale Komponente, die den Ablauf steuert. Empfängt neue Suchresultate aus der Search-Komponente, ruft verschiedene Dienste zur Anreicherung mit Metadaten wie Region, Dokumenttyp auf, speichert das Dokument in der MongoDB und leitet es zur weiteren Verarbeitung über RabbitMG Messages an die Komponenten bmvi_es_connector.py und bmvi_sme_connector.py weiter.

`bmvi_es_connector.py`

RabbitMQ Consumer Client, der die Kommunikation mit dem Elasticsearch API abwickelt; über den `bmvi_es_connector` werden Dokumente in Elasticsearch gespeichert und dort die Segmentierung durchgeführt.

`bmvi_sme_connector.py`

RabbitMQ Consumer Client, der die Kommunikation mit der Smart Media Engine abwickelt. Über den bmvi_sme_connector werden Dokumente in die SME zur Semantischen Indizierung exportiert.

`bmvi_restserver.py`

Komponente, die den REST-Server für den Zugriff auf die Daten und Dienste realisiert. Beruht auf dem Python-basierten Web-Framework Flask. Teile der REST-API sind in Dateien separiert, die unter im Verzeichnis `restapi` liegen.

`config.py`

Konfigurationsdatei.

`requirements.txt`

Listet alle Python-Frameworks auf, die von der Anwendung benötigt werden.

`startup.sh`

Ein Startskript, das die Komponenten des Backends startet.

`classes/RabbitMQ.py`

Komponente, die alle benötigten Basisfunktionen für RabbitMQ-Clients zur Verfügung stellt.

`classes/Search.py`

Komponente, die die Suchfunktionen bereitstellt, insbes. die Suche über das BingAPI.

`classed/SMEConnector.py`

Komponente, die die Funktionen zur Kommunikation mit der Smart Media Engine bereitstellt (außer die Funktionen für das REST API).

`clustering/wissensplattform-umwelt-1.0.jar`

Stellt das Clustering und die Bundeslanderkennung bereit.

```
documenttype_detection/{app.py,consts.py,
          textCategorization.py,webApi.py,words.txt}
```

Umfasst die einzelnen Programme, die für die Dokumenttyperkennung benötigt werden.

`lib/dfki.py`

Stellt Hilfsfunktionen zur Kommunikation mit den Komponenten clustering und `documenttype_detection` bereit.

`lib/sme.py`

Stellt Hilfsfunktionen zur Kommunikation der Smart Media Engine bereit.

`lib/utilities.py`

Bibliothek mit generellen Hilfsfunktionen.

`restapi/dfki.py`

Stellt die REST Endpoints für die Kommunikation mit Komponenten `clustering` und `documenttype_detection` bereit.

`restapi/sme.py`

Stellt die REST Endpoints für die Kommunikation mit der SME bereit.

`ontology/bmvi.ttl`

Beinhaltet die BMVI Ontologie.

2.2 Installationshinweise

2.2.1 Systemvoraussetzungen

Software

Die unter 2.1.1. beschriebenen Software-Voraussetzungen müssen als Mindestanforderungen erfüllt sein, damit die Anwendung installations- und betriebsfähig ist.

Betriebssystem

Das System läuft unter Linux CentOS 7.2. Prinzipiell sollten alle Komponenten aber auch unter anderen Betriebssystemen lauffähig sein, im Einzelnen können ggf. Anpassungen erforderlich werden.

Für eine reine interne Testinstallation sollte eine einzelne Server-Instanz mit mindestens 32GB RAM ausreichen. Für einen Betrieb mit Echtdaten und mehreren Nutzenden ist dringend zu empfehlen, die Komponenten wie MongoDB und Elasticsearch auf eigenen Servern installieren, bzw. Cloud-Dienste zu nutzen. Für einen echten Produktivbetrieb ist das System nicht ausgelegt.

Hardware

Das System läuft unter Linux CentOS 7.2. Prinzipiell sollten alle Komponenten aber auch unter anderen Betriebssystemen lauffähig sein, im Einzelnen können ggf. Anpassungen erforderlich werden.

Machbarkeitsstudie zur Entwicklung einer Wissensplattform im Umweltbereich

Ausgangslage/Zielsetzung

Bei der Planung und Genehmigung von Verkehrsinfrastrukturprojekten sind umfangreiche umwelt-, natur- und artenschutzrechtliche Anforderungen zu beachten, deren Umfang und Komplexität in den vergangenen Jahren stetig angestiegen sind. Um die am Planungs- und Genehmigungsprozess Beteiligten bei der ressourcenintensiven Recherche nach projektrelevanten Informationen zu unterstützen, hat das Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) 2019 die Entwicklung einer softwarebasierten Wissensplattform initiiert, die unter Nutzung selbstlernender Technologien (Künstliche Intelligenz) die Möglichkeit bieten soll, zu jeder fachlichen Frage im Bereich des Artenschutzes verlässliche und zitierfähige Informationen abzurufen.

Im Rahmen der Machbarkeitsstudie sollte geklärt werden, inwieweit eine solche Wissensplattform softwarebasiert und mithilfe von Künstlicher Intelligenz aufgebaut werden kann und inwieweit eine solche Anwendung für die Vorhabenträger und Genehmigungsbehörden einen Mehrwert hätte. Dazu sollte ein Demonstrator entwickelt werden, der die Ergebnisse der Studie anschaulich macht und mit dem erste fachspezifische Suchen durchgeführt werden können. Die Anforderungen an die Wissensplattform resp. den Demonstrator wurden in enger Zusammenarbeit mit potentiellen Anwender/innen aus den Planungs- und Genehmigungsbehörden entwickelt.

Vorgehen/Methodik

Die von dem Auftragnehmer Condat AG gemeinsam mit dem Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) als Unterauftragnehmer entwickelte Wissensplattform erbringt folgende Informationsleistungen:

- Die Plattform ermöglicht die Recherche von im Internet zur Verfügung stehenden Informationen zu zehn geschützten Tierarten. Die Suche unterstützt alle üblichen Dateiformate (PDF, MS-Office, HTML etc.) sowie weitere strukturierte Formate, die über Standardschnittstellen abgefragt werden können.
- Die gefundenen Informationen werden mit Hilfe von Methoden der Künstlichen Intelligenz (KI) aufbereitet. Zur Verarbeitung gehört zunächst eine selbstlernende Relevanzbewertung, die auf Basis einer fachspezifischen, gemeinsam mit den künftigen Nutzenden entwickelten Ontologie sicherstellt, dass nur relevante Dokumente in die Wissensplattform übernommen werden. In einem weiteren Schritt werden die importierten Dokumente mittels semantischer KI-Methoden automatisch um Zusatzinformationen angereichert und kategorisiert. Bei diesen Zusatzinformationen handelt es sich um das Bundesland, aus dem das Dokument stammt und um die Art des Dokuments (Fachinformationen, Rechtliches, Planung). Eine

textuelle und inhaltliche Analyse mit Verfahren aus dem Maschinellen Lernen extrahiert relevante Informationen aus den Dokumenten und ermöglicht dadurch Querverweise zwischen Dokumenten und Quellen.

- Mit dem Demonstrator kann die zukünftige Wissensplattform auszugsweise vorgeführt und ihre Leistungsfähigkeit evaluiert werden. Der Demonstrator gibt zudem einen Ausblick auf die mögliche Gestaltung der Benutzeroberfläche, die die explorativen Suchprozesse nutzerfreundlich unterstützen soll.

Fazit/Ausblick

Die in der Plattform eingesetzten Methoden der Künstlichen Intelligenz können den Rechercheaufwand im Vergleich zu einfachen, keyword-basierten Suchen in verteilten Archiven signifikant reduzieren und genauere und umfassendere Ergebnisse liefern. Die Benutzeroberfläche der Plattform stellt vorhandenes, kontextbezogenes Wissen übersichtlich dar und erlaubt ein direktes Sichten der zugrundeliegenden Dokumente mit Sprungmarken zu relevanten Abschnitten.

Im Ergebnis bestätigen Machbarkeitsstudie und Demonstrator, dass eine auf KI-Technologien basierende, selbstlernende Wissensplattform für das skizzierte Anwendungsszenario des BMVI umsetzbar und geeignet ist, um die Effizienz von Planungs- und Genehmigungsprozessen in der Verkehrswegeplanung zu erhöhen.

Es ergaben sich jedoch im Rahmen der Machbarkeitsstudie auch Grenzen. Die Inhalte bekannter Quellen konnten nicht vollständig in die Plattform überführt werden, da dies nicht automatisiert erfolgen konnte, sondern alle Quellen händisch angebunden werden mussten. Für das Training der KI-basierten Komponenten war die vorhandene Datenmenge teilweise noch zu gering. Um einen Mehrwert für die Vorhabenträger und Genehmigungsbehörden herzustellen, ist daher eine Weiterentwicklung des Systems notwendig.