
Themenheft 72: Mixed Methods im Spannungsfeld. Weiterentwicklung digitaler Forschungsmethoden zwischen KI, NLP, Big Data und transdisziplinärer Forschung. Herausgegeben von Jana Heinz, Andrea Hense und Christian Janßen

Digitale Ressourcen als informelle Lerngelegenheiten: Ein KI-gestütztes Mixed-Methods-Design

Webscraping, promptbasierte Analysepipeline und Validierung am Beispiel der Katastrophenvorsorge

Sophie Lacher-May¹  und Jan May²

¹ Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau

² Jan May IT Solutions

Zusammenfassung

Der Beitrag untersucht deutschsprachige Online-Ressourcen zur Katastrophenvorsorge aus erwachsenenpädagogischer Perspektive als potenzielle Strukturen informeller, selbstgesteuerter Lerngelegenheiten. Auf der Gegenstandsebene steht damit nicht die Beobachtung individueller Lernprozesse im Zentrum, sondern die Frage, welche Inhalte, Adressierungen, medialen Repräsentationsformen und Handlungsorientierungen Webseiten für mögliche Aneignungsprozesse bereitstellen. Auf der Forschungsebene entwickelt der Beitrag ein sequenzielles Mixed-Methods-Design der digitalen Methodenintegration, das qualitativ entwickelte Kategorien mit Webscraping und automatisierter LLM-gestützter Klassifikation in einer skalierbaren Analysepipeline verschränkt. Ergänzend wurden explorative Textanalysen triangulierend genutzt. Empirisch wurden 7.305 deutschsprachige Webseiten ausgewertet. Die Ergebnisse zeigen eine plurale, jedoch stark institutionell und marktförmig gerahmte Angebotslandschaft: Kommerzielle Anbieter:innen und staatliche Stellen dominieren, während zivilgesellschaftliche und individuelle Ressourcen seltener vorhanden sind. Die untersuchten Lerngelegenheitsstrukturen erscheinen überwiegend als statische Informationsseiten. Instruktive, interaktive oder kollaborative Formate bleiben randständig. Darüber hinaus diskutiert der Beitrag, wie sich theoriegeleitete Interpretation, KI-basierte Skalierung und Validierung so verschränken lassen, dass automatisierte Codierungen nachvollziehbar und ihre epistemischen Grenzen sichtbar werden. Darauf aufbauend werden Gütekriterien für digitale Mixed-Methods-Designs herausgearbeitet, insbesondere hinsichtlich Transparenz, Reproduzierbarkeit sowie der Dokumentation von Prompts und Code.

Digital Resources as Opportunities for Informal Learning: An AI-Supported Mixed-Methods Research Design. Web Scraping, Prompt-Based Analysis and Validation in the Context of Disaster Preparedness

Abstract

This article examines German-language online resources on disaster preparedness from the perspective of adult education, considering them as potential sources of informal, self-directed learning. Therefore, at the level of the empirical object, the article does not investigate individual learning processes, but rather asks which content, forms of address, media representations and action orientations websites provide for possible processes of appropriation. In terms of research design, the article takes a sequential mixed-methods approach to integrating digital methods. This approach combines categories developed through qualitative analysis with web scraping and automated, large language model (LLM)-supported classification in a scalable analysis pipeline. Exploratory text analyses were also employed for triangulation purposes. A total of 7,305 German-language websites were analysed empirically. The findings reveal a diverse, yet heavily institutional and market-shaped landscape of online resources, with commercial providers and government agencies dominating and civil society and individual resources being less prevalent. The analysed learning opportunity structures predominantly appear to be static information pages, with instructional, interactive or collaborative formats remaining marginal. Furthermore, the article discusses how theory-driven interpretation, AI-based scaling and validation can be integrated to make automated coding processes traceable and their epistemic limitations visible. On this basis, it identifies quality criteria for digital mixed-methods research designs, particularly with regard to transparency, reproducibility, and prompt and code documentation.

1. Einleitung

Katastrophenvorsorge erhält im deutschsprachigen Raum aktuell mehr Aufmerksamkeit von Politik und Öffentlichkeit als zuvor. Wiederkehrende Extremereignisse und Krisendebatten verdeutlichen, dass Krisenbewältigung nicht allein institutionell organisiert werden kann, sondern auch Selbstschutzkapazitäten in den privaten Haushalten voraussetzt (BBK 2010; BMI 2016; UNISDR 2015). Zugleich haben sich digitale Medien als wichtige Infrastrukturen der Informationssuche etabliert: Bürger:innen nutzen Suchmaschinen, Webseiten und soziale Medien, um sich über Vorbereitung und Handeln im Ereignisfall zu informieren (Lacher und Rohs 2024).

Aus erwachsenenpädagogischer Perspektive lässt sich diese Online-Informationssuche als potenzieller Kontext informellen selbstgesteuerten Lernens rahmen. Lernen meint hier keine in dieser Studie empirisch erfasste individuelle Kompetenzentwicklung

oder Verhaltensänderung, sondern potenzielle anlassbezogene Aneignungsprozesse, in denen Erwachsene digitale Ressourcen zur Orientierung, Bewertung und Vorbereitung nutzen können (Livingstone 2001; Marsick und Watkins 2001; Park und He 2024). Für die Bildungsforschung stellt sich damit die Frage, welche Inhalte, Adressierungen, medialen Repräsentationsformen und Handlungsorientierungen solche Ressourcen für mögliche Aneignungsprozesse bereitstellen.

Digital unterstützte Forschungsmethoden eröffnen die Möglichkeit, solche Lernlandschaften systematisch zu erfassen, stehen jedoch im Spannungsfeld von Skalierung und Interpretation. Die Kombination aus Webscraping, Text-Mining und LLM-gestützter Klassifikation ermöglicht die Analyse umfangreicher Webkorpora, erfordert zugleich aber theoriegeleitete Kategorien, eine transparente Dokumentation und systematische Validierung, um die Nachvollziehbarkeit und epistemischen Grenzen automatisierter Codierung sichtbar zu machen (Baden et al. 2020; Kuckartz und Rädiker 2022).

Vor diesem Hintergrund entwickelt der Beitrag ein sequenzielles Mixed-Methods-Design der digitalen Methodenintegration.¹ Am Beispiel deutschsprachiger Online-Ressourcen zur Katastrophenvorsorge wird gezeigt, wie heterogene Webseiten – von staatlichen Informationsangeboten über kommerzielle und zivilgesellschaftliche Ressourcen bis hin zu privaten Beiträgen – als potenzielle *Lerngelegenheitsstrukturen* im Web entlang erwachsenenpädagogisch begründeter Dimensionen rekonstruiert werden können und welche Validierungsanforderungen, Fehlermuster und Grenzen sich bei LLM-gestützter Klassifikation ergeben.

Leitend ist die Frage, wie sich qualitative Kategorienentwicklung, suchmaschinenbasiertes Sampling und promptbasierte LLM-Klassifikation in ein sequenzielles Mixed-Methods-Design integrieren und über Plausibilitätschecks methodisch reflexiv absichern lassen.

2. Theoretischer Rahmen: Informelles Lernen, digitale Ressourcen und Katastrophenvorsorge

Der Beitrag verortet Katastrophenvorsorge als erwachsenenpädagogisch relevantes Handlungsfeld und fasst Webseiten als *Arenen informeller selbstgesteuerter Aneignung*. Diese Perspektive knüpft an Arbeiten zum informellen und selbstgesteuerten Lernen an, die Lernen als anlassbezogene, häufig nicht explizit als Lernen markierte Nutzung von Ressourcen zur Bearbeitung konkreter Handlungsprobleme beschreiben (Livingstone 2001; Marsick und Watkins 2001; Rohs 2016; Park und He 2024). Webseiten werden vor diesem Hintergrund als digitale Deutungs- und Angebotsräume verstanden, in denen Inhalte, Adressierungen, mediale Repräsentationsformen und Handlungsaufforderungen

¹ Ein bereits publizierter englischsprachiger Beitrag berichtet die empirischen Ergebnisse der Studie ausführlicher (Lacher und Rohs 2026). Der vorliegende Beitrag fokussiert demgegenüber die digitale Methodenintegration, insbesondere Design, Operationalisierung, Prompt-Logik und Validierung.

bereitgestellt werden. Dabei wird kein einheitlicher Lehrauftrag vorausgesetzt: Die untersuchten Ressourcen können staatliche Informationsangebote, kommerzielle Produkt- und Beratungsseiten, zivilgesellschaftliche Hinweise oder private Erfahrungs- und Szenebeiträge umfassen. Als Lerngelegenheitsstrukturen werden sie daher nicht aufgrund einer expliziten pädagogischen Intention der Anbieter:innen verstanden, sondern aufgrund ihres potenziellen Beitrags zu informeller Orientierung, Bewertung und Vorbereitung. Im Zentrum steht nicht die Beobachtung individueller Lernprozesse oder tatsächlicher Nutzungspraktiken, sondern die Analyse von *Lerngelegenheitsstrukturen* im Web. Untersucht wird, wie Online-Ressourcen mögliche Aneignungsprozesse ermöglichen, rahmen oder begrenzen.

2.1 Informelles und selbstgesteuertes Lernen

In bildungspolitischen und erwachsenenpädagogischen Debatten wird Lernen häufig zwischen formalen, non-formalen und informellen Formen unterschieden (Kommission der Europäischen Gemeinschaften 2000; Eurostat 2016). Während formales und non-formales Lernen organisatorisch gerahmt sind, verlaufen informelle Lernprozesse ausserhalb organisierter Bildungsarrangements, werden häufig selbst initiiert und nicht explizit als «Lernen» markiert (Livingstone 2001; Marsick und Watkins 2001; Rohs 2016). Selbstgesteuertes Lernen akzentuiert darüber hinaus, dass Lernende Ziele, Ressourcen und Strategien eigenständig bestimmen (Garrison 1997; Knowles 1950). Im vorliegenden Beitrag wird daher von informeller, selbstgesteuerter Aneignung gesprochen, um Kontext (ausserhalb institutioneller Angebote) und Steuerung zusammenzuführen.

Empirisch überlagern sich Lernformen häufig; sinnvoller als eine fixe Gliederung erscheint eine Perspektive auf *Lernen als Kontinuum unterschiedlicher Grade von Strukturierung und Selbststeuerung* (Colley et al. 2003; Eraut 2004, 2010). Entscheidend ist, wie Nutzer:innen Situationen deuten und Ressourcen zur Bearbeitung konkreter Handlungsprobleme nutzen (Illeris 2003; Lave und Wenger 1991). Im digitalen Raum erweitern Suchmaschinen, Webseiten und Online-Plattformen Möglichkeiten anlassbezogener Informationssuche, wobei Grenzen zwischen Information, Kommunikation und Lernressourcen verschwimmen (Couldry und Hepp 2013; Lundby 2014). Ein Review zum Online-Lernen Erwachsener zeigt, dass informelles und selbstgesteuertes Lernen eng mit webbasierter Informationsnutzung verbunden ist (Park und He 2024).

Für den vorliegenden Beitrag folgt daraus eine analytische Konsequenz: Die Studie rekonstruiert nicht individuelles Lernen, sondern beschreibt Online-Angebote als *Lerngelegenheitsstrukturen*, die problemorientierte Suchbewegungen adressieren, wie sie für informelles Lernen prototypisch beschrieben werden (Marsick und Watkins 2001; Tough 1979), und Handlungswissen in unterschiedlicher Form bereitstellen. Damit wird informelles Lernen über die Strukturen digitaler Ressourcen analytisch zugänglich gemacht.

2.2 Katastrophenvorsorge als Feld der Erwachsenen- und Weiterbildung

Katastrophenvorsorge ist im Zivil- und Bevölkerungsschutz verankert, lässt sich aus erwachsenenpädagogischer Perspektive jedoch zugleich als Bildungs- und Lernbereich beschreiben, in dem Erwachsene Wissen, Orientierungen und Kompetenzen im Umgang mit Risiken und Ausnahmesituationen entwickeln (Karutz, Geier, und Mitschke 2017; Kitagawa 2017; Lacher und Rohs 2023; Preston 2012). Lacher und Rohs (2023) rahmen Katastrophenvorsorge im deutschsprachigen Raum als emergentes Handlungsfeld, in dem formale, non-formale und informelle Bildungsprozesse ineinandergreifen.

Für die vorliegende Studie ist dabei eine Unterscheidung zwischen verschiedenen Akteur:innen und Adressat:innen zentral. *Erstens* existieren institutionalisierte Qualifizierungsstrukturen für haupt- und ehrenamtliche Akteur:innen des Zivil- und Bevölkerungsschutzes (Guerrero Lara und Gerhold 2020; Posingies, Hammerl, und Karutz 2025). *Zweitens* richten sich staatliche, kommunale und zivilgesellschaftliche Informations- und Unterstützungsangebote an die Bevölkerung und adressieren diese als potenziell vorsorgende Haushalte, Familien oder Einzelpersonen (z. B. BBK 2025). *Drittens* umfasst das digitale Feld auch kommerzielle, journalistische, private und szenespezifische Ressourcen, die nicht notwendigerweise einem pädagogischen Auftrag folgen, aber dennoch Orientierungswissen, Handlungsempfehlungen oder Deutungsangebote bereitstellen können (u. a. Ressourcen sogenannter Prepper:innen) (Luy 2025; Mdl-MV 2018).

Damit ist das Feld durch unterschiedliche Grade von Institutionalisierung, unterschiedliche Produzent:innen und unterschiedliche Adressierungen geprägt. Neben formalen und non-formalen Bildungsangeboten stehen Online-Ressourcen, deren Funktionen zwischen Information, Beratung, politischer Kommunikation, Produktvermarktung, Erfahrungsbericht und Handlungsanleitung changieren. Gerade diese heterogene Angebotslandschaft ist für informelle Aneignungsprozesse bedeutsam, wurde in der Erwachsenenbildungsforschung bislang jedoch vergleichsweise selten systematisch untersucht (Lacher und Rohs 2023; Lacher 2025). Der Beitrag knüpft an diese Lücke an, indem er Katastrophenvorsorge explizit als Feld informeller Aneignung im digitalen Raum adressiert.

2.3 Online-Ressourcen als Arenen informeller Aneignung: Analytische Dimensionen

Um die digitale Angebotslandschaft bildungswissenschaftlich zu erfassen, werden Online-Ressourcen nicht als neutrale Informationsspeicher, sondern als Konstellationen von Inhalten, Gestaltungsformen und Adressierungen konzeptualisiert, in denen Deutungsangebote, Handlungsmodelle und Subjektpositionen bereitgestellt werden. Nutzer:innen können solche Ressourcen anlassbezogen aufsuchen und – je nach Vorwissen, Lebenssituation und Handlungskontext – in praktische Orientierungen und Vorsorgeentscheidungen übersetzen.

Für die Analyse werden vier Dimensionen unterschieden: (1) Inhalte und Themen (adressierte Risiken, Szenarien und Vorsorgemassnahmen) (Gerhold et al. 2019); (2) die mediale Aufbereitung und Repräsentationsformen (von beschreibend bis anleitend; z. B. Checklisten, Infografiken, Videos und ggf. interaktive Elemente), wobei handlungsnahe Darstellungen den Übergang von Wissen zu Verhalten unterstützen können (Becker et al. 2012); (3) die Adressierungen und Zielgruppen (implizite und explizite Zielgruppen sowie Rollen- und Verantwortungszuschreibungen) (Goersch und Werner 2011; Lacher und Rohs 2023); (4) die Handlungsorientierung (Grad der Anleitung konkreter Selbstschutzpraktiken von allgemeinen Empfehlungen bis zu Schrittfolgen) (Becker et al. 2012; Müller und Winkler 2023). Diese Dimensionen bilden die konzeptionelle Brücke zur digitalen Methodenintegration: Sie werden in ein Kategoriensystem und promptbasierte Output-Schemata überführt, um Online-Ressourcen qualitativ interpretierbar und auf Korpusebene vergleichbar auszuwerten.

3. Methodisches Vorgehen: Mixed Methods und digitale Methodenintegration

Gegenstand der Untersuchung sind deutschsprachige Online-Ressourcen zur Katastrophenvorsorge, die als potenzielle Lerngelegenheitsstrukturen im Web analysiert werden. Methodisch leitend ist die Frage, wie sich qualitative Kategorienentwicklung, suchmaschinenbasiertes Sampling und promptbasierte LLM-Klassifikation in ein sequenzielles Mixed-Methods-Design integrieren und über Plausibilitätschecks reflexiv absichern lassen. Kapitel 3 skizziert dazu Forschungsdesign, Analysepipeline und Validierung.

3.1 Mixed Methods, digitale Methodenintegration und Bildungsforschung

Das Webangebot zur Katastrophenvorsorge stellt ein dynamisches, algorithmisch strukturiertes und akteur:innenseitig heterogenes Feld dar. Für die Erwachsenenbildung ist daran zentral, dass mögliche informelle Aneignung hier nicht primär als Teilnahme an organisierten Bildungsangeboten sichtbar wird, sondern als anlassbezogene Nutzung digitaler Informations-, Orientierungs- und Handlungsressourcen. Analytisch rückt damit die Frage in den Vordergrund, *wie* potenzielle Lerngelegenheiten im Web dargestellt, gerahmt und begrenzt werden. Online-Ressourcen werden in diesem Beitrag deshalb als *Lerngelegenheitsstrukturen* sichtbar gemacht, ohne vorauszusetzen, dass ihre Anbieter:innen damit eine explizite pädagogische Intention verfolgen. Vielmehr wird untersucht, wie unterschiedliche Anbieter:innen – etwa staatliche Stellen, Hilfsorganisationen, Unternehmen, Vereine, Gemeinschaften oder Privatpersonen – Vorsorgewissen, Adressierungen, Handlungsorientierungen und mediale Repräsentationsformen bereitstellen und dadurch Selbststeuerung ermöglichen, aber auch durch normative Rahmungen, Anbieter:inneninteressen und algorithmische Sichtbarkeit begrenzen.

Letztere verweist auf medienpädagogische Fragen digitaler Ungleichheit, da Such- und Plattformlogiken mitbestimmen, welche Ressourcen auffindbar und damit potenziell aneignungsrelevant werden (Verständig et al. 2016).

Mixed-Methods-Designs bieten einen Rahmen, um solche Angebotsstrukturen sowohl interpretativ als auch systematisch auf Korpusebene zu erfassen. Digitale Verfahren wie Webscraping, Text-Mining und LLM-gestützte Klassifikation ermöglichen Skalierung (Androutsopoulos 2024), ersetzen aber nicht die bildungswissenschaftliche Deutungs- und Rahmungsperspektive, denn ohne theoriegeleitete Kategorien und Validierung drohen Kontext, normative Rahmungen und Machtverhältnisse in datengetriebenen Abstraktionen zu verschwinden (Baden et al. 2020; boyd und Crawford 2012; Kennedy et al. 2015). Der Beitrag folgt einem sequenziellen Mixed-Methods-Design der digitalen Methodenintegration. Die Rolle der Forschenden verschiebt sich dabei von manuellen Codierer:innen einzelner Texte zu Gestalter:innen von Forschungsarchitekturen, in denen interpretative Deutungshoheit und methodische Transparenz gesichert werden (Kuckartz und Rädiker 2022).

3.2 Integration der qualitativen und KI-gestützten Analyse

Die Methodenintegration folgt einem explorativen sequenziellen Mixed-Methods-Design, in dem qualitative und automatisierte Analysen in einer definierten Abfolge verschränkt werden (Creswell und Plano Clark 2018; Kuckartz und Rädiker 2022). *Zuerst* wurden für die *Baselining-Phase* 30 Expert:innen aus Katastrophenvorsorge, Zivil-/Bevölkerungsschutz sowie Erwachsenen- und Weiterbildung gebeten, bis zu zehn Webressourcen zu benennen, die sie der Allgemeinbevölkerung als Informations- und Lernangebote empfehlen würden. Die Gruppe war hinsichtlich institutioneller Kontexte (staatlich/administrativ, zivilgesellschaftlich, privatwirtschaftlich sowie nicht-institutionell) bewusst heterogen zusammengesetzt, um unterschiedliche Perspektiven auf öffentliche Vorsorgekommunikation abzubilden. Um nicht nur Startseiten, sondern typische Navigationspfade von Nutzer:innen zu erfassen, wurden zusätzlich verlinkte Unterseiten (eine Hierarchieebene tiefer) erhoben, um doppelte Einträge bereinigt und anschliessend nach inhaltlicher Relevanz gefiltert (Baseline-Korpus: n = 344). Dieses Korpus wurde manuell anhand eines deduktiv-induktiv entwickelten Kategoriensystems mithilfe der inhaltlich-strukturierenden qualitativen Inhaltsanalyse (Kuckartz und Rädiker 2022) ausgewertet. Die Baseline diente damit der Exploration, Präzisierung von Kategorien und Entscheidungsregeln sowie der Kalibrierung der nachfolgenden Skalierungs- und Validierungsschritte (ebd.).

Im zweiten Schritt wurde das qualitativ entwickelte Kategoriensystem in die KI-gestützte Analysepipeline überführt. Die Prompts und Output-Schemata wurden so formuliert, dass Kategorien und Definitionen der qualitativen Phase regelgeleitet auf den Webkorpus angewendet werden konnten. KI-Klassifikation fungiert damit nicht als autonome Interpretation, sondern als Skalierungsinstrument entlang theoriegeleiteter Dimensionen (vgl. «Hybrid Content Analysis», Baden et al. 2020).

Im dritten Schritt wurden die automatisierten Zuordnungen stichprobenbasiert manuell geprüft. Zufällig gezogene Dokumente wurden mit ihren Modell-Outputs verglichen und Diskrepanzen ausgewertet. Dieser Schritt diente der Qualitätskontrolle sowie der Identifikation typischer Fehlermuster und Grenzfälle, die im weiteren Verlauf für Prompt-Nachschärfungen und die reflexive Interpretation der Ergebnisse genutzt wurden.

3.3 Ableitung des Suchstrings und suchmaschinenbasiertes Webscraping

Webscraping wird als automatisierte Extraktion öffentlich zugänglicher Webinhalte verstanden und erlaubt Datenerhebung auch ohne API-Zugänge (Brown et al. 2025; Krotov et al. 2021; Landers et al. 2016). Ziel war, ein Korpus öffentlich zugänglicher deutschsprachiger Online-Ressourcen zur Katastrophenvorsorge zu erfassen, das den Such- und Zugangslogiken der Allgemeinbevölkerung möglichst nahekommt. Damit wird Sampling nicht nur als technische Auswahl verstanden, sondern als Annäherung an selbstgesteuerte Suchpraktiken, in denen Suchmaschinenrankings Lerngelegenheiten mitstrukturieren.

3.3.1 Suchstring-Entwicklung

Ausgehend von Expert:innenempfehlungen und der daran anschliessenden qualitativen Analyse einer initialen Stichprobe wurden zentrale Begriffe und Formulierungen des Felds der Katastrophenvorsorge identifiziert und in eine systematische Suchstrategie überführt. Ziel war nicht, einzelne Begriffe normativ zu privilegieren oder bestimmte Akteur:innengruppen zu unterstützen, sondern im Feld vorhandene unterschiedliche Suchlogiken und Selbstbeschreibungen abzubilden. Dies betrifft auch szenespezifische oder politisch ambivalente Begriffe wie «Preppen», die hier ausschliesslich als empirisch relevante Such- und Feldbegriffe verwendet werden.

Die Suchstrategie folgte einem kombinatorischen Ansatz, nach dem Begriffsmengen aus drei Wortlisten systematisch als kartesisches Produkt verknüpft wurden $L = L_1 \times L_2 \times L_3$. Die Wortlisten umfassen (1) Krisentypen/Szenarien (L_1 , $n = 18$; z. B. «Notfall», «Blackout», «Hochwasser»), (2) handlungsbezogene Suchbegriffe (L_2 , $n = 18$; z. B. «Vorsorge», «Vorbereitung», «Preppen», «Ausrüstung») sowie (3) informations- und formatbezogene Suchbegriffe (L_3 , $n = 22$; z. B. «Information», «Beispiel», «Tutorial», «Checkliste») (vgl. Tabelle 1).

Wort 1	Wort 2	Wort 3
Wortgruppe 1A: Ereignis / Umfeld / Fokus	Wortgruppe 2A: Tätigkeit – Auswahl A	Wortgruppe 3: informations- & formatbezogen
Krise	Vorsorge / vorsorgen	Information / informieren
Notfall	Vorbereitung / vorbereiten	zeigen
Krieg	Schutz / schützen	Übung / üben
Pandemie	Versorgung / versorgen	Beispiel
Unwetter	Management / managen	Maßnahme
Hochwasser	Planung / planen	Konzept
Erdbeben	Kommunikation	Best Practice
Feuer	Survival	Weiterbildung
Wildnis	Selbstschutz	Tutorial
Stromausfall	Preppen / Prepping	Erklärung
Blackout	Ausrüstung / ausrüsten	Beschreibung
		Definition
Wortgruppe 1B: Oberbegriffe	Wortgruppe 2B: Tätigkeit – Auswahl B	Erläuterung
Bevölkerungsschutz	Vorsorge / vorsorgen	Demonstration
Katastrophenschutz	Vorbereitung / vorbereiten	Ausführung / ausführen
Prepper	Planung / planen	nutzen
Lebensmittel	Survival	Gestaltung / gestalten
Wasser	Selbstschutz	prüfen
Erste Hilfe	Preppen / Prepping	entwickeln
Gefahrenabwehr	Ausrüstung / ausrüsten	Skills
		Fähigkeiten
		Kompetenzen

Tab. 1: Wortlisten zur Suchstring-Generierung.

Anmerkung: Suchanfragen: $L = L_1 \times L_2 \times L_3$. Mehrdeutige Terme wurden ausgeschlossen oder angepasst, z. B. «Brand».

Daraus wurden $n = 7.128$ spezifische Suchstrings generiert. Mehrdeutigkeiten wurden in Pretests identifiziert und angepasst bzw. ausgeschlossen (z. B. «Brand» aufgrund deutsch-englischer Mehrdeutigkeit). Die Wortlisten sind in Tabelle 1 dargestellt; die technischen Verfahrensschritte und Ausschlusskriterien sind im Anhang dokumentiert.

3.3.2 Technische Implementierung des Webscrapings

Für die technische Umsetzung des Webscrapings wurde eine modulare Softwarearchitektur entwickelt. Die Basis bildet eine asynchrone Node.js-basierte Anwendung in TypeScript, die über eine spezialisierte Scraping-Bibliothek (Firecrawl)² generisch generierte Webseiten erfassen kann. Die Suchmaschinenautomatisierung wurde mittels Selenium³ implementiert, einem Framework zur Browserautomatisierung, das die Interaktion mit der Google-Suchmaschine programmatisch steuert. Für jede generierte Suchanfrage wurde die erste Ergebnisseite erfasst (je ca. zehn bis zwölf Treffer). Die dort gelisteten Strings der jeweiligen URLs wurden extrahiert. Die Beschränkung auf die erste Ergebnisseite folgt einerseits forschungspragmatischen Erwägungen, reflektiert andererseits aber auch das typische Suchverhalten von Nutzer:innen, die überwiegend die ersten Suchergebnisse konsultieren (Jansen und Spink 2005). Die damit verbundene Abhängigkeit von Suchmaschinenrankings wird im Diskussionsteil (Kap. 5) als Limitation reflektiert.

² <https://www.firecrawl.dev/> [Version 1.9.2].

³ <https://www.selenium.dev/> [Version 4.29.0].

Die extrahierten URLs wurden anschliessend einem mehrstufigen Filterprozess unterzogen. Zunächst erfolgte eine Deduplizierung, bei der identische oder funktional äquivalente URLs zusammengeführt wurden. Anschliessend wurden URLs gefiltert, die auf strukturelle Seiten ohne inhaltliche Relevanz verwiesen (z. B. Impressum, Datenschutzerklärung, Login-Seiten, Kontaktformulare). Ein Katalog von 80 Ausschlussmustern stellte sicher, dass möglichst nur inhaltlich relevante Ressourcen in die Analyse eingingen.

Die gescrapten Inhalte wurden in einer dokumentenorientierten Datenbank (MongoDB)⁴ persistiert. Jedes Dokument enthält den Hauptinhalt im Markdown-Format, den Quellcode sowie über 50 Metadaten, darunter Titel, Beschreibung, Sprache, Open-Graph-Daten, Zeitstempel und technische Statusinformationen. Die technische Ablauf- und Konfigurationsdokumentation des Webscrapers, einschliesslich Suchmaschinenabfrage, URL-Extraktion, Filterlogik, Scraping-Parameter, Fehlerbehandlung und Datenbankschema, ist im Anhang dokumentiert.

3.3.3 Datenbasis und Qualitätssicherung

Insgesamt wurden 8.238 URLs identifiziert, davon 7.305 erfolgreich erfasst; Ausfälle resultierten u. a. aus Nichterreichbarkeit, robots.txt-Restriktionen und Extraktionsfehlern. Standortabhängigkeiten der Google-Suchergebnisse wurden in Pretests untersucht. Die finale Erhebung erfolgte über einen festen Internetzugang aus Rheinland-Pfalz. Die damit verbundene Standortbindung wird als Limitation dokumentiert, da die Ergebnisse nicht unabhängig von algorithmischen und regionalen Kontextualisierungen der Suchmaschine zu verstehen sind.

3.3.4 Ethische und technische Compliance

Das Webscraping folgte forschungsethischen Standards für die automatisierte Datenerhebung im Internet (Brown et al. 2025; Krotov et al. 2020). Robots.txt-Dateien wurden als maschinenlesbare Crawling-Richtlinien automatisiert host- bzw. domainbezogen geprüft und auf die jeweiligen URL-Pfade angewendet. URLs, deren Abruf durch entsprechende Regeln ausgeschlossen war oder bei denen technische Abruf- bzw. Extraktionsprobleme auftraten, wurden nicht in das finale Korpus aufgenommen. Zusätzlich wurde ein mehrstufiges Rate-Limiting implementiert: Ein Token-Bucket-Algorithmus begrenzte die Anfragen an die Scraping-API auf maximal 100 Requests pro Minute. URLs wurden randomisiert in Batches verarbeitet. Ein Crawl-Delay zwischen 1 und 15 Sekunden wurde eingehalten, sofern robots.txt entsprechende Vorgaben enthielt. Diese Massnahmen dienten der Minimierung von Serverlast und der Einhaltung von Crawling-Vorgaben der Webseitenbetreibenden.

4 <https://www.mongodb.com/> [Version 8.11.0].

3.4 KI-gestützte Analysepipeline und ergänzende explorative Textanalysen

Die inhaltliche Analyse des Webkorpus erfolgte mit einer KI-gestützten Analysepipeline, die grosse Sprachmodelle (LLMs) zur automatisierten Klassifikation entlang erwachsenenpädagogisch begründeter Dimensionen nutzt. Ziel ist nicht, Interpretation an das Modell zu delegieren, sondern theoriegeleitete Kategorien auf Korpusebene skalierbar anzuwenden und dabei Validierung und Nachvollziehbarkeit sicherzustellen. Aus erwachsenenpädagogischer Sicht wird die Klassifikation daher als strukturierte Beschreibung von Lerngelegenheiten verstanden, nicht als Messung individueller Kompetenzen oder Lernleistungen.

3.4.1 KI-Modelle und Einsatzlogik

Die eingesetzten Modelle basieren auf der Transformer-Architektur (Vaswani et al. 2017) und können auf Basis von Prompts strukturierte Ausgaben generieren. Verwendet wurden GPT-4.1-mini⁵, GPT-4o-mini⁶ und GPT-4o⁷, die über die OpenAI API-Schnittstelle angesprochen wurden. Die meisten Klassifizierungen erfolgten aus Kosten- und Effizienzgründen mit GPT-4o-mini. GPT-4o wurde hauptsächlich für Web-Search-Funktionen zur Identifikation von Webseitenbetreibern genutzt. Der Kategorien-Analysator verwendete GPT-4.1-mini, da dieses Modell in unseren Evaluierungen bessere Ergebnisse als GPT-4o-mini erzielte und dabei dennoch kosteneffizient war (0,40\$/1M Tokens). Pro Analyseauftrag wurde die Eingabelänge pipeline-seitig auf maximal 8.000 Tokens begrenzt, was je nach Textstruktur ungefähr 25.000 Zeichen entspricht. Sehr lange Webseiten wurden auf den extrahierten Hauptinhalt und relevante Metadaten begrenzt; dies wird als methodische Limitation berücksichtigt.

Jüngere Forschung zeigt, dass LLMs bei Textklassifikationsaufgaben mit sorgfältig entwickelten Prompts Ergebnisse erzielen können, die mit manueller Annotation vergleichbar oder dieser teilweise überlegen sind (Bunt et al. 2025; Gilardi et al. 2023; Törnberg 2024). Zugleich werden die Grenzen dieser Verfahren, etwa zur Erfassung subtiler Bedeutungsnuancen oder kontextabhängiger Interpretationen, in der Methodoliteratur kritisch diskutiert (Demszky et al. 2023; Pangakis et al. 2023). Darüber hinaus handelt es sich um ein dynamisches methodisches Feld, in dem sich Leistungsfähigkeit und Problemmuster mit der rapiden Modellentwicklung verschieben: So zeigen jüngste Studien, dass subtile Implementierungsentscheidungen die Klassifikationsergebnisse erheblich beeinflussen können und etablierte Prompting-Strategien wie *Chain of Thought* in fortgeschrittenen Modellen an Wirksamkeit verlieren können (Carlson und Burbano 2025).

5 <https://platform.openai.com/docs/models/gpt-4.1-mini>.

6 <https://platform.openai.com/docs/models/gpt-4o-mini>.

7 <https://platform.openai.com/docs/models/gpt-4o>.

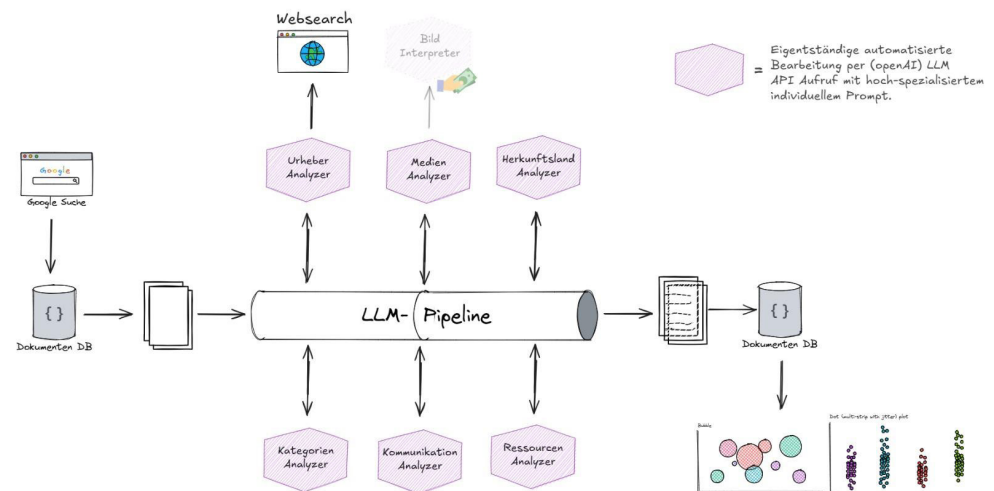


Abb. 1: Automatisierte Pipeline inkl. Analysatoren mit jeweils individuellem Prompt [Quelle: eigene Darstellung].

3.4.2 Architektur der Analysepipeline

Die Analysepipeline verarbeitet jedes Dokument des Webkorpus entlang modularer Analysatoren. Ausgangspunkt ist jeweils ein in MongoDB gespeichertes Dokument mit URL, Titel, Beschreibung, extrahiertem Hauptinhalt im Markdown-Format sowie technischen und medialen Metadaten. Diese Dokumentinformationen werden analysatorspezifisch in strukturierte Prompts überführt; die Outputs werden als JSON gespeichert, mit kurzen Begründungen versehen und anschliessend für die Korpusauswertung aggregiert. Die vollständigen Promptvorlagen, Output-Schemata, Modellparameter und Entscheidungsregeln sind im Anhang dokumentiert.

Die Pipeline umfasst sechs Analysatoren (vgl. Abbildung 1):

- (1) Der *Kategorien-Analysator* verarbeitet URL, Titel, Beschreibung und extrahierten Haupttext. Er ordnet jedes Dokument einer oder mehreren Primär- und Unterkategorien des Kategoriensystems zu (z. B. Hauptkategorie: Notfallvorsorge themenübergreifend, Subkategorie: Checklisten & Anleitungen; Hauptkategorie: Energie & Infrastruktur, Subkategorie: Blackout-Vorsorge). Randthemen und beiläufige Erwähnungen werden gemäss konservativer Codierregel nicht berücksichtigt. Der Output umfasst Primärkategorien, Unterkategorien und eine kurze Begründung. Das System wurde deduktiv aus Theorie und Vorstudie abgeleitet und im Prozess materialgestützt erweitert.
- (2) Der *Medien-Analysator* verarbeitet automatisch extrahierte Hinweise zu Medienformaten, insbesondere Anzahl und Metadaten von Bildern, Videos und Audiodateien sowie Kontextinformationen wie Alt-Texte, Bildunterschriften und Überschriften. Er klassifiziert, welche Medienformate den Hauptinhalt wesentlich unterstützen,

unterscheidet dekorative von informationstragenden Elementen und erfasst Hinweise auf interaktive Medienbestandteile. Ergänzend zur textbasierten Medienanalyse werden Bilder in einem separaten visionsbasierten Durchlauf einzeln analysiert. Als Input dienen das jeweilige Bild sowie Dateiname, Alt-Text, Bildunterschrift und umgebender Text, soweit verfügbar. Klassifiziert werden u. a. erklärende oder anleitende Funktion, Visualisierungstyp, Selbsterklärungsfähigkeit und kognitive Komplexität. Die Einzelbildanalysen werden anschliessend pro Webseite aggregiert.

- (3) Der *Kommunikations-Analysator* prüft, ob und in welcher Form Interaktionsmöglichkeiten erkennbar sind. Die Klassifikation erfolgt entlang einer vierstufigen Skala (einseitig; eingeschränkt interaktiv; interaktiv; sonstige). Als Evidenz dienen u. a. Hinweise auf Kommentarbereiche, Formulare, Feedbackfunktionen, Foren, Chats oder kollaborative Werkzeuge.
- (4) Der *Ressourcen-Analysator* typisiert die Webseite als Selbstlerneinheit, Instruktion, Kollaborationsplattform, Informationsressource oder sonstige Ressource. Ausschlaggebend sind Strukturierungsgrad, Handlungsorientierung, Sequenzierung, Interaktionsmöglichkeiten und Hinweise auf Lern- oder Anwendungsschritte. Die Kategorie «Instruktion» wird nur vergeben, wenn eine Schrittfolge oder Handlungssequenz den Hauptteil der Seite prägt; andernfalls wird die Seite konservativ als Informationsressource codiert.
- (5) Der *Urheber-Analysator* ordnet den Webseitenbetreibenden einen Organisationstyp zu (staatlich; nicht-staatliche Hilfsorganisation; sonstiger Verein; internationale Organisation, Verband; Gemeinschaft; Unternehmen; Privatperson). Die Klassifikation erfolgt zweistufig: Domain-Caching (bereits klassifizierte Domains) und – falls nötig – ergänzende Recherche über das Modell via Web-Search zur Einordnung des Betreibers.
- (6) Der *Herkunfts-Analysator* nimmt anhand der Top-Level-Domain eine grobe geografische Zuordnung vor. Länderspezifische Domains wie .de, .at und .ch werden Deutschland, Österreich beziehungsweise der Schweiz zugeordnet. Bei generischen Domains wie .org, .com und .net ist dagegen keine eindeutige nationale Zuordnung möglich. Dies stellt eine Limitation der Analyse dar.

Die einzelnen Outputs werden nicht als autonome Interpretation des Modells verstanden, sondern als standardisierte Zuordnungen entlang zuvor definierter Kategorien. Sie bilden die Grundlage für die deskriptive Korpusauswertung, wurden in einem weiteren Schritt stichprobenbasiert plausibilisiert und für die Ergebnisinterpretation verdichtet. Die vollständigen Modellkonfigurationen, Output-Schemata, Promptvorlagen und Entscheidungsregeln sind in den Abschnitten 3–6 des elektronischen Zusatzmaterials dokumentiert.

3.4.3 Prompt-Design und Reproduzierbarkeit

Alle Prompts wurden in deutscher Sprache verfasst und folgen einer einheitlichen Struktur aus System-Prompt und User-Prompt. Die Klassifikation folgt dem Prinzip konservativer Codierung: Kategorien werden nur bei möglichst klarer Evidenz vergeben. Randthemen und beiläufige Erwähnungen bleiben unberücksichtigt. Zusätzlich erzeugt das Modell kurze Begründungen, um manuelle Plausibilitätsprüfungen und Grenzfallanalysen zu unterstützen.

Um die Variabilität der Modellausgaben zu reduzieren, wurde der Temperaturparameter, der den Grad der Zufallsvariation bei der Generierung beeinflusst, auf 0,0 gesetzt. Zudem wurde mit JSON (JavaScript Object Notation) ein standardisiertes, maschinenlesbares Ausgabeformat vorgegeben. Diese Einstellungen erhöhen die Konsistenz der Ausgaben, gewährleisten jedoch keine vollständige Reproduzierbarkeit.

3.4.4 Rolle der Forschenden

Im beschriebenen Design verschiebt sich die Rolle der Forschenden gegenüber klassisch-qualitativen Studien: Sie codieren nicht primär einzelne Texte, sondern gestalten Kategoriensystem, Prompt- und Schema-Design, Pipeline-Konfiguration sowie Validierung und Fehleranalyse (Kennedy, Poell, und van Dijck 2015; Knappertsbusch, Langfeldt, und Kelle 2021). Die interpretative Deutungshoheit verbleibt so bei den Forschenden. KI-Systeme wenden vordefinierte Kategorien auf grosse Datenmengen an, fungieren aber nicht als eigenständige Interpret:innen.

Diese Rollentransformation eröffnet die Möglichkeit, Lernlandschaften auf Korpusebene sichtbar zu machen, die manuell kaum erschliessbar wären. Zugleich sind Kategorien, Sampling, Prompt-Formulierungen, Modellparameter und Validierungsschritte als epistemische Setzungen digitaler Forschungspraxis zu reflektieren (boyd und Crawford 2012; Kennedy, Poell, und van Dijck 2015). Dies gilt besonders für LLM-gestützte Designs, deren Ergebnisse von technischen und implementierungsbezogenen Entscheidungen abhängen (Carlson und Burbano 2025; Törnberg 2024). Zudem strukturieren Suchmaschinenrankings Sichtbarkeit und Auswahlwahrscheinlichkeit von Webressourcen (Lewandowski und Schultheiß 2023; Westerwick 2013).

4. Ergebnisse: Online-Ressourcen als Arenen informeller Aneignung

Die Ergebnisse beziehen sich auf das finale Webkorpus von 7.305 deutschsprachigen Webseiten zur Katastrophenvorsorge. Sie werden entlang der im theoretischen Rahmen entwickelten Dimensionen – Akteur:innenstruktur, thematische Ausrichtung, Ressourcentypen sowie Medienformate und Repräsentationsmodi – als Strukturmerkmale potenzieller Lerngelegenheiten dargestellt. Ausführlichere Ergebnisdarstellungen (u. a. Detailwerte und ergänzende Deskriptionen des Korpus) sind in Lacher und Rohs (2026) dokumentiert; der vorliegende Beitrag berichtet die Befunde verdichtet und fokussiert

ihre methodische Einordnung. Die berichteten Anteile basieren auf den in Kapitel 3 beschriebenen stichprobenbasiert manuell geprüften LLM-Zuordnungen.

4.1 Akteur:innenlandschaft und Ressourcentypen

Die Analyse der Urheberkategorien zeigt, dass Katastrophenvorsorge im deutschsprachigen Web von einer Vielzahl unterschiedlicher Akteur:innen adressiert wird. Dominant sind kommerzielle Akteur:innen (43,5 %; z. B. Online-Shops, Sicherheits-, Outdoor- oder Vorsorgeanbieter), gefolgt von staatlichen Einrichtungen (29,7 %; z. B. Bundes-, Landes-, Kreis- oder kommunale Behörden). Kleinere Anteile entfallen auf nicht-staatliche Hilfsorganisationen (9,3 %; z. B. Rotes Kreuz, Johanniter, Malteser, Arbeiter-Samariter-Bund, DLRG), Privatpersonen (7,1 %) sowie Vereine und Gemeinschaftsinitiativen (6,0 %). Für einen Restanteil (4,4 %) war keine eindeutige Zuordnung möglich. Damit ist die sichtbare Angebotslandschaft plural, zugleich aber deutlich durch staatliche und kommerzielle Akteur:innen geprägt; community-basierte und individuelle Stimmen bleiben vergleichsweise selten sichtbar.

Die folgenden Ressourcentypen wurden nicht induktiv aus dem Gesamtkorpus gebildet, sondern auf Basis der theoretischen Dimensionen und der qualitativen Voranalyse vorab als promptbasierte Klassifikationskategorien definiert und anschliessend automatisiert auf das Webkorpus angewendet. Die Promptvorlagen des Ressourcen-Analysators und die zugehörigen Entscheidungsregeln sind im elektronischen Zusatzmaterial (Abschnitte 5.2 und 6.1–6.4) dokumentiert. Mit Blick auf die Ressourcentypen zeigt die KI-gestützte Klassifikation ein deutliches Übergewicht klassisch informierender Formate: 93,8 Prozent der Webseiten werden als allgemeine Informationsressource ohne ausgeprägte Strukturierung im Sinne von Anleitung, Sequenzierung oder Interaktion eingeordnet. Instruktive Formate wie Schritt-für-Schritt-Anleitungen, Tutorials oder Checklisten machen einen deutlich kleineren Anteil aus (3,5 %); kollaborative Ressourcen mit Foren- oder Community-Funktionen bleiben selten (1,7 %). Auch explizit als Selbstlernereinheiten konzipierte Online-Kurse oder Lernmodule sind nur vereinzelt vorhanden (0,2 %). Der Rest entfällt auf Residual-/Sonstige-Kategorien bzw. unklare Zuordnungen (0,8 %). Das Online-Angebot ist damit primär durch Informations- und Ratgeberangebote geprägt, weniger durch strukturierte Lernpfade, anleitende Formate oder dialogische Austauschmöglichkeiten.

4.2 Thematische Schwerpunkte und adressierte Vorsorgemassnahmen

Die thematische Klassifikation basiert auf dem im elektronischen Zusatzmaterial dokumentierten Kategoriensystem mit seinen Primär- und Unterkategorien (Abschnitte 7.1 und 7.2); die zugehörigen Promptvorlagen des Kategorien-Analysators sind in Abschnitt 5.1 zu finden. Die Prozentwerte beziehen sich auf den Anteil der Webseiten, denen die

jeweilige Primärkategorie zugeordnet wurde; Mehrfachzuordnungen pro Webseite waren möglich, sodass die Kategorien nicht als gegenseitig ausschliessend zu verstehen sind. Am häufigsten treten übergreifende Kategorien wie *Notfallvorsorge themenübergreifend* (25,6%), *persönlicher Schutz & Gefahrenabwehr* (22,2%) sowie *Notfallplanung & -ausrüstung* (21,7%) auf. Diese Ressourcen bündeln grundlegende Hinweise, Orientierungswissen und niedrigschwellige Handlungsanregungen. Sie bilden damit den thematischen Kern des Korpus.

Ergänzend dazu finden sich risikobezogene Themenfelder wie Naturkatastrophen (17,6%; z. B. Hochwasser, Sturmereignisse, Hitze) und medizinische Versorgung (11,4%). Sie konkretisieren allgemeine Vorsorgeprinzipien und adressieren spezifische Gefährdungen, bleiben quantitativ aber hinter den generalisierten *All-Hazards*-Angeboten zurück.

Seltener werden demgegenüber Themenbereiche wie Lebensmittelbevorratung (7,1%), baulicher Schutz von Gebäuden und Infrastruktur (5,6%) sowie kriegs- und terrorbezogene Risiken (4,6%) aufgegriffen. Katastrophenvorsorge wird im untersuchten Online-Angebot damit vor allem über grundlegende, relativ niedrigschwellige Schutz- und Vorsorgemassnahmen gerahmt; komplexere, ressourcenintensive oder konflikthaft diskutierte Aspekte von Vorsorge sind deutlich weniger präsent. Damit bildet das Online-Angebot zwar ein breites, aber thematisch ungleich verteiltes Spektrum ab; stärker voraussetzungsreiche Kontext- und Aushandlungsdimensionen von Vorsorge bleiben im Korpus nur begrenzt sichtbar.

4.3 Medienformate und Repräsentationsmodi

Die Auswertung der medialen Merkmale vertieft die Perspektive auf Online-Ressourcen als potenzielle *Lerngelegenheitsstrukturen*. Bei 50,6 Prozent der Webseiten ist Text der primäre Medientyp. Primär video- (5,8%) und audiobasierte (0,8%) Angebote kommen wesentlich seltener vor, obwohl sie davon abweichende Zugänge zu Informationen und Handlungsorientierungen eröffnen können.

Visuelle Elemente mit erklärend-instruktiver Funktion – etwa Checklisten, Beschriftungen, Karten oder Schrittfolgen – finden sich nur auf einem Teil der Webseiten: 20,2 Prozent (n = 1.474) der analysierten Seiten enthalten entsprechende Bilder. Insgesamt wurden von 61.824 identifizierten Bildern lediglich 5,8 Prozent (n = 3.593) als erklärend-instruktiv klassifiziert; der überwiegende Teil der Visuals erfüllt primär dekorative oder aufmerksamkeitssteuernde Funktionen. Gleichzeitig sind erklärend-instruktive Bilder, wenn sie vorkommen, meist selbsterklärend (95%), bleiben jedoch überwiegend einfach und statisch (97% zeigen Einzelobjekte oder Einzelschritte), sodass sie eher punktuelle Orientierung und basale Handlungsimpulse unterstützen als komplexere prozedurale Zusammenhänge darzustellen. Damit bleibt das Potenzial multimodaler Gestaltung im untersuchten Korpus insgesamt begrenzt.

Vor dem Hintergrund informeller Aneignung verweist diese Befundlage auf eine Angebotslandschaft, die zwar vielfältige Möglichkeiten zur Informationssuche bereitstellt, jedoch überwiegend einem konsumtiven Informationsmodell folgt. Unterstützungsstrukturen für Einordnung, Anwendung und Austausch, etwa interaktive, dialogische oder sequenziell aufgebaute Lernformate, sind im Korpus nur punktuell zu finden.

4.4 Qualitätssicherung der KI-gestützten Zuordnung

Die KI-gestützte Klassifikation wurde über manuelle codebuchgeleitete Plausibilitätschecks abgesichert, indem in einem iterativen Verfahren zufällig ausgewählte Maschinenoutputs mit manuellen Zuordnungen des Forschungsteams verglichen wurden. Währenddessen wurden Grenzfälle dokumentiert und Prompts bzw. Entscheidungsregeln nachgeschärft. Ziel war keine quantifizierte Modellgüte im Sinne einer *Accuracy* auf Basis eines vollständig gelabelten Ground-Truth-Datensatzes (Deng und Zheng 2020), sondern eine transparente Einordnung der Tragfähigkeit und Grenzen automatisierter Zuordnungen. Die hierfür relevanten Modellkonfigurationen, Output-Schemata, Promptvorlagen und Entscheidungsregeln sind im elektronischen Zusatzmaterial (Abschnitte 3–6) dokumentiert.

5. Diskussion: Mixed Methods zwischen Automatisierung und Interpretation

Die Ergebnisse zeigen eine umfangreiche Online-Angebotslandschaft zur Katastrophenvorsorge, die jedoch nur begrenzt anleitende, interaktive und dialogische Unterstützungsformen bereitstellt. Sie ist durch institutionelle, marktförmige und technische Logiken strukturiert und macht Spannungen zwischen informeller Aneignung, Verantwortungszuschreibungen und digitaler Methodenintegration exemplarisch sichtbar.

5.1 Informelle Aneignung zwischen Zugang und Handlungsmacht

Die thematische Breite des Korpus eröffnet grundsätzlich niedrigschwellige Zugänge: Die untersuchten Ressourcen können anlassbezogene Suchbewegungen, Orientierung und grundlegende Handlungsoptionen unterstützen. Gleichzeitig dominiert ein konsumtives Informationsmodell. Überwiegend textbasierte, deskriptive Ressourcen stehen wenigen instruktiven, interaktiven oder kollaborativen Formaten gegenüber. Damit bleibt die Übersetzung von «wissen, dass» in «wissen, wie» weitgehend Aufgabe der Nutzer:innen; Unterstützungsstrukturen für Einordnung, Anwendung und Austausch sind selten. Diese Ambivalenz digitaler Informalität ist anschlussfähig an Befunde zu ungleich verteilten Kompetenzen der Informationssuche, -bewertung und -anwendung (Morris und Rohs 2023; Park und He 2024).

Die Akteur:innenstruktur verstärkt diese Logik: Staatliche Stellen und kommerzielle Anbieter:innen prägen die sichtbaren Ressourcen, während zivilgesellschaftliche und individuelle Stimmen sowie dialogische Formate randständig bleiben. Bildungstheoretisch lässt sich dies als ambivalente Responsibilisierung lesen: Vorsorge wird als individuelle Aufgabe adressiert, während strukturelle Fragen von Risiko, Ungleichheit und kollektiver Verantwortung im Online-Angebot seltener ausgearbeitet werden. In diesem Sinne tendiert die Online-Kommunikation eher zu einer *«pedagogy for the public»* (Biesta 2014) als zu partizipativen Formen einer *«pedagogy of the public»* (ebd.).

Damit rückt ein Spannungsfeld in den Blick, das erwachsenenpädagogische Debatten zur politischen Bildung, zu Resilienz und Sicherheit seit Längerem durchzieht: Katastrophenvorsorge kann sowohl als emanzipatorische Stärkung von Handlungsfähigkeit gelesen werden als auch als Verschiebung von Verantwortung auf Individuen, ohne dass entsprechende Ressourcen und Mitbestimmungsmöglichkeiten bereitgestellt werden (Rådestad und Larsson 2020; Roth und Prior 2014).

5.2 Methodische Implikation: Notwendigkeit der reflexiven Einbettung

Methodisch zeigt die Studie am konkreten Fall der hier verwendeten Pipeline, wie qualitative Kategorienentwicklung, suchmaschinenbasiertes Sampling und LLM-gestützte Klassifikation zur Analyse eines grossen, heterogenen Webkorpus verschränkt werden können. Die Aussagen beziehen sich auf die zum Erhebungs- und Analysezeitpunkt verfügbaren technischen Mittel, Modelle und Implementierungsentscheidungen, sind daher nicht ohne Weiteres auf andere Modelle, Plattformen oder künftige technische Entwicklungsstände übertragbar. Zugleich bleiben Blindstellen: Automatisierte Zuordnungen sind besonders stark bei expliziten Textsignalen, aber begrenzt bei impliziten Adressierungen, Grenzfällen hybrider Seitenformate (z. B. Info-Text plus kurze Checkliste) und bei Interaktionselementen, die dynamisch nachgeladen oder im Extrakt nicht sichtbar sind. Vor diesem Hintergrund sind die Operationalisierungen – einschliesslich des Samplings über Suchmaschinenrankings – als epistemische Setzungen zu verstehen. Entsprechend sind Makromuster (Akteur:innendominanz; Übergewicht statischer Informationsseiten; geringe Interaktivität) belastbarer interpretierbar, während Feindifferenzierungen medialer, anleitender und interaktiver Unterstützungsformen vorsichtiger zu lesen und bei Bedarf qualitativ zu vertiefen sind. Der hohe Anteil kommerzieller Anbieter:innen (43,5%) kann als Artefakt von SEO-Optimierung diskutiert werden. Kommerzielle Seiten investieren systematisch in Suchmaschinenranking und können so im Sample überproportional sichtbar werden (Lewandowski et al. 2021). Bei einer Datenerhebung entlang dieser Rankings steigt die Wahrscheinlichkeit, dass solche Anbieter:innen die Stichprobe dominieren. Damit ist nicht belegt, dass Katastrophenvorsorge insgesamt *«kommerzialisiert»* wäre; eher zeigt sich die *Koppelung der Online-Sichtbarkeit an SEO-getriebene Distributionsmechanismen*. Ob und inwiefern damit eine

breitere Kommerzialisierung von Katastrophenvorsorge verbunden ist, bleibt eine weiterführende Forschungsfrage.

Insgesamt unterstreicht die Diskussion, dass digitale Mixed-Methods-Ansätze ihr Potenzial dort entfalten, wo sie nicht als Ersatz, sondern als Erweiterung bildungswissenschaftlicher Perspektiven verstanden werden: Sie machen Strukturen und Muster digitaler Angebotslandschaften sichtbar und erfordern zugleich, die eigenen Deutungen, Kategorien und methodischen Entscheidungen immer wieder kritisch zu prüfen.

6. Fazit und Ausblick

Die Studie macht Online-Ressourcen zur Katastrophenvorsorge als Arenen informeller Aneignung rekonstruierbar und zeigt eine Angebotslandschaft, deren Sichtbarkeit stark durch staatliche und kommerzielle Akteur:innen geprägt ist. Zivilgesellschaftliche und individuelle Beiträge sind deutlich seltener. Thematisch dominieren allgemeine Vorsorge, persönlicher Schutz und basale Notfallplanung; ressourcenintensive oder konflikthaft diskutierte Aspekte wie längerfristige Versorgung oder kriegsbezogene Szenarien bleiben randständig. Auch medial überwiegt ein konsumtives Informationsmodell: Statische Informationsseiten sind die Regel, während Instruktionen, Interaktion und Kollaboration Ausnahmen bleiben; Bilder sind überwiegend dekorativ, erklärend-instruktive Visualisierungen sind selten.

Methodisch zeigt der Beitrag, wie qualitative Kategorienentwicklung, suchmaschinenbasiertes Sampling und LLM-gestützte Klassifikation in einem sequenziellen Mixed-Methods-Design verschränkt und durch codebuchgeleitete Plausibilitätschecks reflexiv abgesichert werden können. Die Analysepipeline eröffnet damit Möglichkeiten, digitale Angebotslandschaften auf Korpusebene sichtbar zu machen, bleibt aber an die verwendeten Modelle, Prompts, Datenzugänge und Validierungsschritte gebunden.

Für Praxis und Politik ergibt sich, Online-Ressourcen stärker als anwendungsorientierte Informations- und Orientierungsangebote zu gestalten – etwa durch Schrittfolgen, erklärend-instruktive Visualisierungen oder niederschwellige Interaktionsmöglichkeiten – und partizipative sowie communitybasierte Formate auszubauen. Forschungsseitig erscheinen länder- und sprachvergleichende Designs, Längsschnittstudien sowie die Verbindung mit Nutzungs- und Aneignungsperspektiven (Interviews/Surveys/Ethnografie) als besonders anschlussfähig; multimodale Ansätze versprechen, mediale, anleitende und interaktive Unterstützungsformen jenseits des Textes präziser zu erfassen.

Literatur

- Androutsopoulos, Jannis. 2024. «Scaling as method: A three-stage, mixed-methods approach to digital discourse analysis». *Discourse, Context & Media* 62:100817. <https://doi.org/10.1016/j.dcm.2024.100817>.
- Baden, Christian, Neta Kligler-Vilenchik, und Moran Yarchi. 2020. «Hybrid Content Analysis: Toward a Strategy for the Theory-driven, Computer-assisted Classification of Large Text Corpora». *Communication Methods and Measures* 14 (3): 165–83. <https://doi.org/10.1080/19312458.2020.1803247>.
- BBK. 2010. *Neue Strategie zum Schutz der Bevölkerung in Deutschland*. 2. Aufl. Wissenschaftsforum Band 4. Bonn: Bundesamt für Bevölkerungsschutz und Katastrophenhilfe. http://www.bbk.bund.de/SharedDocs/Downloads/BBK/DE/Publikationen/Wissenschaftsforum/Band-4_Neu-eStrategie.pdf?__blob=publicationFile.
- BBK – Bundesamt für Bevölkerungsschutz und Katastrophenhilfe. 2025. *Vorsorgen für Krisen und Katastrophen*. Bonn: Bundesamt für Bevölkerungsschutz und Katastrophenhilfe.
- Becker, Julia S., Douglas Paton, David M. Johnston, und Kevin R. Ronan. 2012. «A model of household preparedness for earthquakes: how individuals make meaning of earthquake information and how this influences preparedness». *Natural Hazards* 64 (1): 107–37. <https://doi.org/10.1007/s11069-012-0238-x>.
- Biesta, Gerd. 2014. «Making Pedagogy Public: For the Public, of the Public, or in the Interest of Publicness?». In *Problematizing Public Pedagogy*, herausgegeben von Jake Burdick, Jennifer A. Sandlin, und Michael P. O'Malley, 15–25. Hoboken: Taylor and Francis.
- BMI. 2016. «Konzeption Zivile Verteidigung (KZV)». https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/themen/bevoelkerungsschutz/konzeption-zivile-verteidigung.pdf;jsessionid=CE3E0B53E2722B0C1D868F7DD3071883.1_cid373?__blob=publicationFile&v=2.
- boyd, danah, und Kate Crawford. 2012. «CRITICAL QUESTIONS FOR BIG DATA». *Information, Communication & Society* 15 (5): 662–79. <https://doi.org/10.1080/1369118X.2012.678878>.
- Brown, Megan A., Andrew Gruen, Gabe Maldoff, Solomon Messing, Zeve Sanderson, und Michael Zimmer. 2025. «Web scraping for research: Legal, ethical, institutional, and scientific considerations». *Big Data & Society* 12 (4). <https://doi.org/10.1177/20539517251381686>.
- Bunt, Hannah L., Alex Goddard, Tom W. Reader, und Alex Gillespie. 2025. «Validating the use of large language models for psychological text classification». *Frontiers in Social Psychology* 3. <https://doi.org/10.3389/frsps.2025.1460277>.
- Carlson, Natalie A., und Vanessa Burbano. 2025. «The Use of LLMs to Annotate Data in Management Research: Foundational Guidelines and Warnings». *Strategic Management Journal*, 1–27. <https://doi.org/10.1002/smj.70023>.
- Colley, Helen, Phil Hodgkinson, und Janice Malcolm. 2003. *Informality and formality in learning: A report for the Learning and Skills Research Centre*. London: Learning and Skills Research Centre.
- Couldry, Nick, und Andreas Hepp. 2013. «Conceptualizing Mediatization: Contexts, Traditions, Arguments». *Communication Theory* 23 (3): 191–202. <https://doi.org/10.1111/comt.12019>.

- Creswell, John W., und Vicki L. Plano Clark. 2018. *Designing and Conducting Mixed Methods Research*. 3rd edition. Los Angeles, Calif, London, New Delhi, Singapore, Washington DC, Melbourne: SAGE.
- Demszky, Dorottya, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margaret Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel Jones Mitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, und James W. Pennebaker. 2023. «Using large language models in psychology». *Nature Reviews Psychology* 2: 688–701. <https://doi.org/10.1038/s44159-023-00241-5>.
- Deng, Weijian, und Liang Zheng. 2020. «Are Labels Always Necessary for Classifier Accuracy Evaluation?». <https://arxiv.org/pdf/2007.02915>.
- Eraut, Michael. 2004. «Informal learning in the workplace». *Studies in Continuing Education* 26 (2): 247–73. <https://doi.org/10.1080/158037042000225245>.
- Eraut, Michael. 2010. «Non-Formal Learning and Tacit Knowledge in Professional Work». *The British journal of educational psychology* 70 (1): 113–36. <https://doi.org/10.1348/000709900158001>.
- Eurostat. 2016. *Classification of Learning Activities (CLA): Manual*. 2016 edition. Luxembourg: Publications Office. <https://doi.org/10.2785/874604>.
- Garrison, D. R. 1997. «Self-Directed Learning: Toward a Comprehensive Model». *Adult Education Quarterly* 48 (1): 18–33. <https://doi.org/10.1177/074171369704800103>.
- Gerhold, Lars, Stefanie Wahl, und Wolf R. Dombrowsky. 2019. «Risk perception and emergency food preparedness in Germany». *International Journal of Disaster Risk Reduction* 37:101183. <https://doi.org/10.1016/j.ijdrr.2019.101183>.
- Gilardi, Fabrizio, Meysam Alizadeh, und Maël Kubli. 2023. «ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks». *Proceedings of the National Academy of Sciences of the United States of America* 120 (30): e2305016120. <https://doi.org/10.1073/pnas.2305016120>.
- Goersch, Henning G., und Ute Werner. 2011. *Empirische Untersuchung der Realisierbarkeit von Maßnahmen zur Erhöhung der Selbstschutzzfähigkeit der Bevölkerung*. Forschung im Bevölkerungsschutz Bd. 15. Bonn: Bundesamt für Bevölkerungsschutz und Katastrophenhilfe.
- Guerrero Lara, Anna, und Lars Gerhold. 2020. *Bildung im Bevölkerungsschutz: Teil 1: Bildungsatlas Bevölkerungsschutz – strukturelle Merkmale der Bildung im Bevölkerungsschutz*. Forschung im Bevölkerungsschutz, Band 26. Bonn: Bundesamt für Bevölkerungsschutz und Katastrophenhilfe.
- Illeris, Knud. 2003. «Towards a contemporary and comprehensive theory of learning». *International Journal of Lifelong Education* 22 (4): 396–406. <https://doi.org/10.1080/02601370304837>.
- Jansen, Bernard J., und Amanda Spink. 2005. «An analysis of Web searching by European AlltheWeb.com users». *Information Processing & Management* 41 (2): 361–81. [https://doi.org/10.1016/S0306-4573\(03\)00067-0](https://doi.org/10.1016/S0306-4573(03)00067-0).
- Karutz, Harald, Wolfram Geier, und Thomas Mitschke. 2017. «Einführung». In *Bevölkerungsschutz*, herausgegeben von Harald Karutz, Wolfram Geier, und Thomas Mitschke, 1–27. Berlin, Heidelberg: Springer.

- Kennedy, Helen, Thomas Poell, und Jose van Dijck. 2015. «Data and agency». *Big Data & Society* 2 (2). <https://doi.org/10.1177/2053951715621569>.
- Kitagawa, Kaori. 2017. «Situating preparedness education within public pedagogy». *Pedagogy, Culture & Society* 25 (1): 1–13. <https://doi.org/10.1080/14681366.2016.1200660>.
- Knappertsbusch, Felix, Bettina Langfeldt, und Udo Kelle. 2021. «Mixed-Methods and Multimethod Research». In *Soziologie – Sociology in the German-Speaking World: Special Issue Soziologische Revue 2020*, herausgegeben von Betina Hollstein, Rainer Greshoff, Uwe Schimank, und Anja Weiß, 261–72. Berlin, Boston: De Gruyter Oldenbourg. <https://doi.org/10.1515/9783110627275-018>.
- Knowles, Malcolm S. 1950. *Informal Adult Education: A Guide for Administrators, Leaders, and Teachers*. New York: Association Press.
- Kommission der europäischen Gemeinschaften. 2000. *Memorandum über Lebenslanges Lernen*. Brüssel: Kommission der europäischen Gemeinschaften. https://www.die-bonn.de/espid/dokumente/doc-2000/EU00_01.pdf.
- Krotov, Vlad, Leigh Johnson, und Leiser Silva. 2020. «Tutorial: Legality and Ethics of Web Scraping». *CAIS* 47:539–63. <https://doi.org/10.17705/1CAIS.04724>.
- Kuckartz, Udo, und Stefan Rädiker. 2022. *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung: Grundlagentexte Methoden*. 5. Auflage. Grundlagentexte Methoden. Weinheim, Basel: Beltz Juventa.
- Lacher, Sophie. 2025. «Exploring disaster preparedness: a scoping review of adult and continuing education approaches in international civil protection research». *Journal of Risk Research*, 1–17. <https://doi.org/10.1080/13669877.2024.2447252>.
- Lacher, Sophie, und Matthias Rohs. 2023. «Civil protection through adult and continuing education in Germany. A scoping review of an emerging research field». *International Journal of Lifelong Education*, 1–18. <https://doi.org/10.1080/02601370.2023.2263651>.
- Lacher, Sophie, und Matthias Rohs. 2024. «Erwachsenenbildung als Beitrag zu einer Katastrophenvorsorge: Eine empirische Untersuchung zu Informations- und Lernangeboten für den Hochwasserschutz». In *Erwachsenenbildung und Nachhaltigkeit: Sondierungen und Forschung zwischen Anspruch und Wirklichkeit*, herausgegeben von Malte Ebner von Eschenbach, Bernd Käßlinger, Maria Kondratjuk, Katrin Kraus, Matthias Rohs, Bernhard Schmidt-Hertha, Karin J. Rott, und Veronika Thalhammer, 129–41: Barbara Budrich.
- Lacher, Sophie, und Matthias Rohs. 2026. «Preparedness Without Pedagogy? An AI-Assisted Web Scraping Analysis of Informal Online Disaster Preparedness Resources for the Public». *Education Sciences* 16 (1): 146. <https://doi.org/10.3390/educsci16010146>.
- Landers, Richard N., Robert C. Brusso, Katelyn J. Cavanaugh, und Andrew B. Collmus. 2016. «A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data from the Internet for Use in Psychological Research». *Psychological Methods* 21 (4): 475–92. <https://doi.org/10.1037/met0000081>.
- Lave, Jean, und Étienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Learning in doing. Cambridge: Cambridge Univ. Press. <http://www.loc.gov/catdir/description/cam024/91010450.html>.

- Lewandowski, Dirk, und Sebastian Schultheiß. 2023. «Public awareness and attitudes towards search engine optimization». *Behaviour & Information Technology* 42 (8): 1025–44. <https://doi.org/10.1080/0144929X.2022.2056507>.
- Lewandowski, Dirk, Sebastian Sünkler, und Nurce Yagci. 2021. «The influence of search engine optimization on Google's results». In *13th ACM Web Science Conference 2021*, 12–20. New York: ACM. <https://doi.org/10.1145/3447535.3462479>.
- Livingstone, D. W. 2001. «Adults' Informal Learning: Definitions, Findings, Gaps, and Future Research». <https://tspace.library.utoronto.ca/retrieve/4484/>.
- Lundby, Knut, Hrsg. 2014. *Mediatization of Communication*. Handbooks of Communication Science [HoCS] 21. Berlin: de Gruyter Mouton. <https://www.degruyterbrill.com/isbn/9783110272215>. <https://doi.org/10.1515/9783110272215>.
- Luy, Mischa. 2025. *Die erwartete Katastrophe*. Unter Mitarbeit von Jürgen Straub. Gießen: Psycho-sozial-Verlag. Dissertation.
- Marsick, Victoria J., und Karen E. Watkins. 2001. «Informal and Incidental Learning». *New Dir Adult Contin Educ* 2001 (89): 25–34. <https://doi.org/10.1002/ace.5>.
- MdI-MV – Ministerium für Inneres und Europa Mecklenburg-Vorpommern. 2018. *Abschlussbericht der Kommission zur Beleuchtung der Prepper-Szene in Mecklenburg-Vorpommern*. Schwerin: Ministerium für Inneres und Europa Mecklenburg-Vorpommern.
- Morris, Thomas Howard, und Matthias Rohs. 2023. «The potential for digital technology to support self-directed learning in formal education of children: a scoping review». *Interactive Learning Environments* 31 (4): 1974–87. <https://doi.org/10.1080/10494820.2020.1870501>.
- Müller, Bianca, und Gertrud Winkler. 2023. «Ernährung in Krisenzeiten: Untersuchung zum Lebensmittel-Notvorrat in privaten Haushalten». https://haushalt-wissenschaft.de/wp-content/uploads/2023/09/HUW_13_2023_Winkler_Notvorrat.pdf.
- Pangakis, Nicholas, Samuel Wolken, und Neil Fasching. 2023. «Automated Annotation with Generative AI Requires Validation». arXiv preprint arXiv:2306.00176. <https://doi.org/10.48550/arXiv.2306.00176>.
- Park, Yujin, und Dan He. 2024. «An Integrative Review of Adult Learners' Informal Learning in Online Settings». In *Handbook of Research in Online Learning*, herausgegeben von Trey Martindale, Tonya B. Amankwatia, Lauren Cifuentes, und Anthony A. Piña, 168–95. BRILL. https://doi.org/10.1163/9789004702813_007.
- Posingies, Corinna, Angela Hammerl, und Harald Karutz. 2025. «Aus-, Fort- und Weiterbildungsangebote für Lernende im Bevölkerungsschutz». *Notfall Rettungsmed*. <https://doi.org/10.1007/s10049-025-01564-x>.
- Preston, John. 2012. *Disaster Education*. Rotterdam: Sense. <https://doi.org/10.1007/978-94-6091-873-5>.
- Rådestad, Carl, und Oscar Larsson. 2020. «Responsibilization in contemporary Swedish crisis management: expanding 'bare life' biopolitics through exceptionalism and neoliberal governmentality». *Critical Policy Studies* 14 (1): 86–105. <https://doi.org/10.1080/19460171.2018.1530604>.

- Rohs, Matthias. 2016. «Genese informellen Lernens». In *Handbuch informelles Lernen*, herausgegeben von Matthias Rohs, 3–38. Springer Reference Sozialwissenschaften. Wiesbaden: Springer VS.
- Roth, Florian, und Tim Prior. 2014. «The Boundaries of Building Societal Resilience: Responsibilization and Swiss Civil Defense in the Cold War». *BEHEMOTH – A Journal on Civilisation* 7 (2): 103–23. <https://doi.org/10.6094/behemoth.2014.7.2.835>.
- Törnberg, Petter. 2024. «Best Practices for Text Annotation with Large Language Models». *Sociologica* 18 (2). <https://doi.org/10.6092/issn.1971-8853/19461>.
- Tough, Allen M. 1979. *The adult's learning projects: A fresh approach to theory and practice in adult learning*. 2nd ed. Research in education series / Ontario Institute for Studies in Education no. 1. Toronto: Ontario Institute for Studies in Education.
- UNISDR – United Nations International Strategy for Disaster Reduction. 2015. «Sendai Framework for Disaster Risk Reduction 2015–2030». https://www.preventionweb.net/files/43291_sendaiframeworkfordrren.pdf.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, und Illia Polosukhin. 2017. «Attention Is All You Need». https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- Verständig, Dan, Alexandra Klein, und Stefan Iske. 2016. «Zero-Level Digital Divide: Neues Netz und neue Ungleichheiten». *Si:So. Analysen – Berichte – Kontroversen* 21 (1): 50–55.
- Westerwick, Axel. 2013. «Effects of Sponsorship, Web Site Design, and Google Ranking on the Credibility of Online Information». *Journal of Computer-Mediated Communication* 18 (2): 194–211. <https://doi.org/10.1111/jcc4.12006>.

Anhang

Prompt-, Output- und Konfigurationsdokumentation

Dieses Dokument stellt die für die Reproduzierbarkeit relevanten Prompt- und Schema-Elemente, Modellkonfigurationen sowie zentralen Entscheidungsregeln der Analysepipeline bereit. Die Promptvorlagen werden in der Form dokumentiert, in der sie im Analyseprozess verwendet wurden. Begriffe wie «didaktisch» werden dabei als Bestandteil der ursprünglichen Prompt-Sprache wiedergegeben; im Haupttext werden diese Konzepte präziser als erklärend-instruktive, anleitende oder mediale Unterstützungsformen gefasst.

1. Überblick über Datenbasis und Pipeline

Korpus: n = 7.305 deutschsprachige Webseiten zur Katastrophenvorsorge

Sampling: suchmaschinenbasiert (erste Ergebnisseite pro Suchanfrage bei Google)

Suchstrings: n = 7.128, generiert als Kombination dreier Wortlisten

Identifizierte URLs: n = 8.238

Erfolgreich erfasste Webseiten: n = 7.305

Analysedimensionen: Inhalte/Themen; Ressourcentyp; Interaktivität; mediale Repräsentationsformen und erklärend-instruktive Elemente; Urheber*typ; Herkunft (D/A/CH)

LLM-basierte Klassifikation: modulare Analysatoren mit JSON-Outputs; konservative Codierung; stichprobenbasierte Plausibilitätschecks

Zweistufige Medienanalyse: (1) textbasierte Medienerfassung, (2) visuelle Bildanalyse mittels GPT-4o-mini Vision

2. Webscraping-Dokumentation

2.1 Ziel und Grundlogik des Webscrapings

Ziel des Webscrapings war es, ein Korpus öffentlich zugänglicher deutschsprachiger Online-Ressourcen zur Katastrophenvorsorge aufzubauen, das den Such- und Zugangslogiken der Allgemeinbevölkerung möglichst nahekommt. Das Sampling erfolgte daher suchmaschinenbasiert über Google. Für jeden generierten Suchstring wurde die erste Ergebnisseite erfasst. Dieses Vorgehen folgt der Annahme, dass Nutzer:innen in selbstgesteuerten Suchsituationen überwiegend frühe Suchergebnisse konsultieren und Suchmaschinenrankings damit die Sichtbarkeit potenzieller Lerngelegenheiten mitstrukturieren.

Die finale Erhebung erfolgte über einen festen Internetzugang aus Rheinland-Pfalz im Zeitraum vom 23.03.2025 bis 09.04.2025. Standortabhängigkeiten der Suchergebnisse wurden in Pretests geprüft und im Haupttext als Limitation reflektiert.

2.2 Softwarearchitektur

Die technische Umsetzung basierte auf einer modularen Node.js-/TypeScript-Anwendung. Die Suchmaschinenabfrage und Browserinteraktion wurden mit Selenium¹ automatisiert. Für die anschließende Extraktion der Webseiteninhalte wurde Firecrawl² eingesetzt, um auch gerenderte Webseiteninhalte erfassen zu können. Die extrahierten Daten wurden in einer dokumentenorientierten MongoDB-Datenbank³ gespeichert.

Die Pipeline umfasst folgende technische Schritte:

1. Generierung der Suchstrings aus drei Wortlisten
2. automatisierte Google-Abfrage pro Suchstring
3. Extraktion der Ziel-URLs aus der ersten Ergebnisseite
4. Normalisierung, Bereinigung um doppelte URLs und Filterung der URLs
5. Übergabe der verbleibenden URLs an Firecrawl
6. Speicherung von Hauptinhalt, Quellcode und Metadaten in MongoDB
7. Dokumentation von Ausfällen, Fehlern und Ausschlussgründen

2.3 Suchmaschinenabfrage und URL-Extraktion

Für jede Suchanfrage wurde die erste Google-Ergebnisseite automatisiert geöffnet und ausgewertet. Pro Suchstring wurden in der Regel ca. 10 bis 12 Treffer erfasst. Aus den Suchergebnissen wurden die Ziel-URLs der organischen Treffer extrahiert und für die weitere Verarbeitung gespeichert. Nicht-inhaltliche Google-Elemente, interne Suchmaschinenlinks oder offensichtlich nicht als Webseitenressource auswertbare Treffer wurden nicht als eigenständige Ressourcen in den Korpus aufgenommen.

2.4 URL-Normalisierung und Bereinigung um doppelte Einträge

Die extrahierten URLs wurden vor dem Scraping normalisiert und um doppelte Einträge bereinigt. Dabei wurden identische URLs zusammengeführt. Zusätzlich wurden funktional äquivalente Varianten, soweit eindeutig erkennbar, bereinigt, etwa durch Vereinheitlichung technischer URL-Varianten. Ziel war es, Mehrfacherfassungen derselben Ressource zu vermeiden, ohne inhaltlich unterschiedliche Unterseiten einer Domain auszuschliessen.

Die Bereinigung erfolgte damit auf URL-Ebene; Unterseiten wurden nur dann zusammengeführt, wenn sie als identisch oder funktional äquivalent erkannt wurden. Inhaltlich eigenständige Unterseiten blieben im Korpus erhalten.

1 <https://www.selenium.dev/> [Version 4.29.0]

2 <https://www.firecrawl.dev/> [Version 1.9.2]

3 <https://www.mongodb.com/> [Version 8.11.0]

2.5 Filterlogik und Ausschlussmuster

Nach der Bereinigung um doppelte Einträge wurden URLs anhand definierter Ausschlussmuster gefiltert. Ausgeschlossen wurden insbesondere strukturelle oder administrative Seiten ohne unmittelbare inhaltliche Relevanz für Katastrophenvorsorge, z. B. Impressum, Datenschutzerklärung, Login-Seiten, Kontaktformulare, Warenkorbseiten oder rein technische Seiten. Insgesamt wurde ein Katalog von 80 Ausschlussmustern verwendet. Ziel der Filterung war es, möglichst nur inhaltlich relevante Informations-, Orientierungs- oder Handlungsressourcen in das Webkorpus aufzunehmen.

2.6 Scraping-Prozess und Datenbankoutput

Die nach der URL-Bereinigung und Filterung verbleibenden URLs wurden an Firecrawl übergeben. Für erfolgreich erfasste Webseiten wurden der extrahierte Hauptinhalt im Markdown-Format, der Quellcode bzw. HTML-Inhalt sowie technische und inhaltliche Metadaten gespeichert. Jedes MongoDB-Dokument enthält mindestens folgende Felder bzw. Feldgruppen:

- URL und normalisierte URL
- Titel und Beschreibung
- extrahierter Hauptinhalt im Markdown-Format
- HTML-/Quellcode bzw. technischer Seiteninhalt
- Sprache bzw. Sprachhinweise
- Open-Graph- und weitere Metadaten
- Zeitstempel der Erfassung
- technischer Status der Erfassung
- Hinweise auf Medienobjekte, soweit automatisch erfassbar
- Fehler- bzw. Ausschlussinformationen, sofern zutreffend

2.7 Fehlerbehandlung

Nicht alle identifizierten URLs konnten erfolgreich erfasst werden. Ausfälle wurden dokumentiert und nach Fehlerarten unterschieden. Dazu zählen insbesondere Nichterreichbarkeit, Timeouts, Weiterleitungs- oder Abruffehler, robots.txt-bezogene Ausschlüsse sowie Extraktionsfehler. Insgesamt wurden 8.238 URLs identifiziert; davon konnten 7.305 erfolgreich erfasst und analysiert werden.

2.8 Robots.txt-Prüfung und Zugriffsvorgaben

Robots.txt-Dateien wurden als maschinenlesbare Crawling-Richtlinien verstanden und host- bzw. domainbezogen geprüft. Für eindeutige Host-/Domain-Kombinationen wurde automatisiert geprüft, ob eine robots.txt-Datei verfügbar war und ob die jeweiligen

URL-Pfade für automatisierte Abrufe zugelassen oder ausgeschlossen waren. Die Ergebnisse wurden für weitere URLs derselben Host-/Domain-Kombination wiederverwendet, um wiederholte Abrufe der robots.txt-Dateien zu vermeiden.

Robots.txt wurde dabei nicht als rechtlich bindende oder technische Zugriffsbeschränkung im Sinne serverseitiger Zugriffskontrollen verstanden, sondern als forschungsethisch relevante Crawling-Vorgabe. URLs, deren automatisierter Abruf durch entsprechende Regeln ausgeschlossen war oder bei denen technische Abruf- bzw. Extraktionsprobleme auftraten, wurden nicht in den finalen Korpus aufgenommen. Die Anzahl eindeutiger Hosts/Domains betrug 2.839.

2.9 Rate-Limiting und Serverlast

Zur Minimierung von Serverlast und zur Berücksichtigung von Crawling-Vorgaben wurde ein mehrstufiges Rate-Limiting implementiert. Ein Token-Bucket-Algorithmus begrenzte die Anfragen an die Scraping-API auf maximal 100 Requests pro Minute. URLs wurden randomisiert in Batches verarbeitet. Ein Crawl-Delay zwischen 1 und 15 Sekunden wurde eingehalten, sofern robots.txt entsprechende Vorgaben enthielt. Diese Massnahmen dienten der kontrollierten Durchführung des Webscrapings und der Reduktion technischer Belastungen für Websitebetreibende.

3. Modellkonfiguration und Laufzeitparameter

Alle LLM-basierten Analysatoren wurden mit Temperatur 0.0 und festem JSON-Ausgabeformat ausgeführt. Pro Analyseaufruf wurde die Eingabelänge pipeline-seitig auf maximal 8.000 Tokens begrenzt, was je nach Textstruktur ungefähr 25.000 Zeichen entspricht. Sehr lange Webseiten wurden auf den extrahierten Hauptinhalt und relevante Metadaten begrenzt.

Analysator	Modell	Temperatur	Output
Kategorien-Analysator	GPT-4.1-mini	0.0	JSON
Ressourcen-Analysator	GPT-4o-mini	0.0	JSON
Kommunikations-Analysator	GPT-4o-mini	0.0	JSON
Medien-Analysator (Text)	GPT-4o-mini	0.0	JSON
Bild-Analysator (Vision)	GPT-4o-mini	0.0	JSON
Urheber-Analysator	GPT-4o	0.0	JSON
Herkunfts-Analysator	regelbasiert	---	JSON

Tab. A1: Modell- und Laufzeitkonfigurationen

4. Schematische Output-Strukturen (JSON)

Für jeden Analysator wird ein strukturiertes JSON-Objekt erzeugt. Die folgenden Darstellungen zeigen die erwarteten Felder und zulässigen Werte.

4.1 Kategorien-Analysator

```
{ "kategorien": { "primärkategorien": ["Kategorie1", "Kategorie2"],  
  "unterkategorien": { "Kategorie1": ["Unterkategorie1", "Unterkategorie2"],  
    "Kategorie2": ["Unterkategorie3", "Unterkategorie4"] } }, "begründung": "Kurze  
Begründung für die Kategorisierung" }
```

4.2 Ressourcen-Analysator

```
{ "ressourcenart":  
  "SELBSTLERNEINHEIT|INSTRUKTION|KOLLABORATION|INFORMATIONSSRESSOURCE|SONSTIGE",  
  "begründung": "Begründung für die Zuordnung", "didaktische_merkmale": ["Merkmal1",  
    "Merkmal2"] }
```

4.3 Kommunikations-Analysator

```
{ "kommunikationsart": "EINSEITIG|EINGESCHRÄNKT INTERAKTIV|INTERAKTIV|SONSTIGE",  
  "begründung": "Begründung für die Zuordnung", "interaktionselemente": ["Element1",  
    "Element2"] }
```

4.4 Medien-Analysator (Textbasiert)

```
{ "medienformate": ["textbasiert", "Bilder", "Video", "Audio", "Multimedia",  
  "Sonstige"], "beschreibung": "Beschreibung der gefundenen Medien",  
  "informationsgehalt": "hoch|mittel|niedrig", "besonderheiten": "Besonderheiten  
der Medien", "interaktive_elemente": true|false, "interaktive_beschreibung":  
  "Beschreibung interaktiver Elemente" }
```

4.5 Bild-Analysator (Visuell)

Die Bildanalyse erfolgt in einem zweiten Durchlauf mittels GPT-4o-mini Vision. Jedes Bild wird einzeln analysiert und anschliessend aggregiert.

Einzelbild-Schema:

```
{ "filename": "image.jpg", "md5Hash": "abc123...", "hasDidacticValue": true|false,  
  "confidenceScore": 0.0-1.0, "contentDescription": "Kurze Beschreibung des  
Bildinhalts", "educationalElements": ["Element1", "Element2"], "visualType":  
  "photo|diagram|chart|infographic|illustration|screenshot|text|map|icon|other",  
  "didacticFunctions":  
  ["instruction|explanation|warning|demonstration|visualization|overview|example|comp  
arison"], "selfSufficiency": "self_explanatory|text_dependent|context_dependent",  
  "cognitiveComplexity": "simple|moderate|complex", "evidence": ["Evidenz 1",
```

```
"Evidenz 2"], "contextUsed": "none|filename|alt_text|surrounding_text|caption",  
"analyzedAt": "ISO-8601 timestamp", "isReused": false}
```

Aggregations-Schema (Summary):

```
{"summary": {"totalImages": 4, "analyzedImages": 4, "didacticImages": 1,  
"didacticPercentage": 25.0, "averageConfidence": 0.65, "overallClassification":  
"decorative|mixed|didactic", "aggregatedAt": "ISO-8601 timestamp", "reusedCount":  
0}, "processingStatus": "completed|partial|failed", "lastProcessedAt": "ISO-8601  
timestamp"}
```

4.6 Urheber-Analysator

```
{"urheber": "staatlich|nicht staatliche Hilfsorganisation|sonstige  
Vereine|Organisationen|Gemeinschaften|Unternehmen|Privatperson", "begründung":  
"Kurze Begründung für die Einordnung", "quellen": ["Quelle 1", "Quelle 2"],  
"confidence": 0.0-1.0, "method": "database|web_search|error"}
```

4.7 Herkunfts-Analysator

```
{"nationalitaet": "Deutschland|Österreich|Schweiz|unbekannt", "tld":  
"de|at|ch|...", "domain": "example", "subdomain": "www"}
```

5. Prompt-Vorlagen

Die folgenden Prompt-Vorlagen dokumentieren die tatsächlich im Code verwendeten Prompts.

5.1 Kategorien-Analysator

Systemprompt:

```
Du bist ein Analysesystem für Webseiteninhalte. Du antwortest ausschließlich im JSON-Format. Du  
analysierst den Inhalt und gibst deine Antwort in folgendem JSON-Format: {"primärkategorien":  
["Kategorie1", "Kategorie2"], "unterkategorien": {"Kategorie1": ["Unterkategorie1", "Unterkategorie2"],  
"Kategorie2": ["Unterkategorie3", "Unterkategorie4"]}, "begründung": "Kurze Begründung für die  
Kategorisierung"}
```

Userprompt:

```
Analysiere den folgenden Webseiteninhalt und ordne ihn Kategorien zu. URL: {url} Titel: {title}  
Beschreibung: {description} 1. Kategorisierung: Ordne den Inhalt einer oder mehreren dieser  
Kategorien zu: – Primärkategorien: {PRIMARY_CATEGORIES} – Die Kategorien sollten den Inhalt  
des Textes widerspiegeln und die wichtigen Hauptthemen abdecken. – Randthemen / nebenläufige  
Erwähnungen sollen nicht berücksichtigt werden. – lieber weniger Kategorien als zu viele. 2.  
Unterkategorisierung: Ordne jeder ausgewählten Primärkategorie eine oder mehrere passende  
Unterkategorien zu: {SECONDARY_CATEGORIES mapping} 4. Begründung: Gib eine kurze Begrün-  
dung für deine Kategorisierung. Hier ist der Inhalt: {text} Antworte ausschließlich im JSON-  
Format.
```

5.2 Ressourcen-Analysator

Systemprompt:

Du bist ein Analysesystem für Ressourcentypen in Webseiten. Du antwortest ausschließlich im JSON-Format. Du analysierst den Ressourcentyp und gibst deine Antwort in folgendem JSON-Format: {"ressourcenart": "SELBSTLERNEINHEIT/INSTRUKTION/KOLLABORATION/INFORMATIONSSRESSOURCE/SONSTIGE", "begründung": "Begründung für die Zuordnung", "didaktische_merkmale": ["Merkmal1", "Merkmal2"]}

Userprompt:

Analysiere diesen Webseiteninhalt nach seinem didaktischen Format und Ressourcentyp. RES-SOURCENART: Klassifiziere die Website nach ihrem didaktischen Format: A) "SELBSTLERNEINHEIT": • Strukturiertes Lernpfad mit Lernzielen/Modulen • Enthält Übungen oder Wissensüberprüfungen • Beispiele: E-Learning-Kurs, Wissensplattform mit Kapiteln B) "INSTRUKTION": • Schrittweise Anleitungen oder Erklärungen • Fokus auf praktische Umsetzung mit Demonstrationen • Beispiele: How-To-Guides, Tutorials, Notfallpläne C) "KOLLABORATION": • Plattform für gemeinsames Arbeiten/Lernen • Austausch oder Diskussionsräume vorhanden • Beispiele: Wiki-Plattform, Bildungsforum D) "INFORMATIONSSRESSOURCE": • Bereitstellung von Fakten ohne didaktische Struktur • Primär zur Information, nicht zum gezielten Lernen • Beispiele: Nachrichtenwebsite, Lexikon, Blog E) "SONSTIGE": Nur wenn keine andere Kategorie zutreffend Achte auf Struktur, Navigation, Übungsanteile und Interaktionsmöglichkeiten. Hier ist der Inhalt: {text} Antworte ausschließlich im JSON-Format.

5.3 Kommunikations-Analysator

Systemprompt:

Du bist ein Analysesystem für Kommunikationsarten in Webseiten. Du antwortest ausschließlich im JSON-Format. Du analysierst die Kommunikationsart und gibst deine Antwort in folgendem JSON-Format: {"kommunikationsart": "EINSEITIG/EINGESCHRÄNKT INTERAKTIV/INTERAKTIV/SONSTIGE", "begründung": "Begründung für die Zuordnung", "interaktionselemente": ["Element1", "Element2"]}

Userprompt:

Analysiere die Interaktionsmöglichkeiten in diesem Webseiteninhalt und ordne ihn einer Kommunikationsart zu. KOMMUNIKATIONSART: Analysiere die Interaktionsmöglichkeiten der Website und ordne sie genau EINER Kategorie zu: A) "EINSEITIG" wenn ALLE diese Merkmale zutreffen: • Nutzer können nur Inhalte konsumieren, nicht beitragen • Keine Kommentarfunktion vorhanden • Keine Formulare außer ggf. Kontaktformular/Newsletter-Anmeldung • Primär Informationsvermittlung (Texte, Bilder, Videos) B) "EINGESCHRÄNKT INTERAKTIV" wenn EINES dieser Merkmale zutrifft: • Kommentarfunktion unter Artikeln/Beiträgen vorhanden • Bewertungs-/Feedback-System implementiert • Einfache Formulare zur Informationsabfrage • Digitale Arbeitsblätter/einfache Tests vorhanden C) "INTERAKTIV" wenn EINS dieser Merkmale zutreffen: • Zweiseitige Echtzeit-Kommunikation (Chat, Messenger) • User-Generated-Content als zentrales Element • Forum mit aktivem Austausch • Kollaborative Tools (gemeinsame Dokumente, Projekte) • Interaktive Anwendungen/Simulationen D) "SONSTIGE" nur wenn keine der obigen Kategorien eindeutig passt WICHTIG:

Prüfe META-Tags nach interaktiven Elementen, untersuche HTML auf Formular-Elemente und API-Einbindungen, und analysiere vorhandene UI-Elemente zur Nutzerinteraktion. Hier ist der Inhalt: {text} Antworte ausschließlich im JSON-Format.

5.4 Medien-Analysator (Textbasiert)

Systemprompt:

Du bist ein Analysesystem für Medieninhalte in Webseiten. Du antwortest ausschließlich im JSON-Format. Du analysierst die Medieninhalte und gibst deine Antwort in folgendem JSON-Format: {"medienformate": ["textbasiert", "Bilder", "Video"], "beschreibung": "Beschreibung der gefundenen Medien", "informationsgehalt": "hoch/mittel/niedrig", "besonderheiten": "Besonderheiten der Medien", "interaktive_elemente": true/false, "interaktive_beschreibung": "Beschreibung interaktiver Elemente (falls vorhanden)"}.

Userprompt:

Analysiere die Medienelemente im folgenden Webseiteninhalt. Erkannte Medien (automatisch): Bilder: {image_count} Videos: {video_count} Audio: {audio_count} Details zu den erkannten Medien: {media_info} Medienanalyse: Identifiziere basierend auf dem Inhalt, welche relevanten Medienelemente den Hauptinhalt unterstützen: – Wähle alle zutreffenden Medienformate aus: "textbasiert", "Video", "Audio", "Bilder", "Multimedia", "Sonstige" – "textbasiert" IMMER auswählen, wenn der Inhalt überwiegend aus Text besteht – "Video", "Audio", "Bilder", "Multimedia", "Sonstige" nur wenn sie wesentlich zum Inhalt beitragen – Beurteile Informationsbilder gezielt nach Typ: Infografiken, Diagramme, Charts, Datenvisualisierungen, thematische Illustrationen – Ignoriere explizit: Portraitfotos bei Interviews, dekorative Bilder ohne Informationsgehalt, Logos, Icons, Banner, Navigations-elemente – Bewerte, wie stark die Medien den Textinhalt unterstützen (wesentlich/ergänzend/minimal) – Enthalten die Medien einzigartige Informationen, die im Text nicht vorkommen? – Analysiere bei Bildern den Dateinamen und Alt-Text nach Hinweisen wie "chart", "graph", "infographic", etc. – Gibt es interaktive Elemente wie Datentabellen, klickbare Visualisierungen, Checklisten oder interaktive Karten? Hier ist der Inhalt: {text} Antworte ausschließlich im JSON-Format.

5.5 Bild-Analysator (Visuell)

Die visuelle Bildanalyse erfolgt in einem separaten Durchlauf. Jedes Bild einer Webseite wird einzeln an GPT-4o-mini Vision übergeben.

Systemprompt:

Du bist ein Analysesystem zur Bewertung des didaktischen Werts von Bildern im Kontext von Webseiten zur Katastrophenvorsorge. Du antwortest ausschließlich im JSON-Format. Deine Aufgabe ist es, jedes Bild hinsichtlich seines Lehr- und Informationswertes zu bewerten. Unterscheide zwischen: – Dekorativen Bildern (Stockfotos, Header, Stimmungsbilder, Logos, Icons) – Informativen Bildern (tragen sachliche Information bei) – Didaktischen Bildern (unterstützen aktiv das Lernen/Verstehen/Handeln) Antworte im folgenden JSON-Format: {"hasDidacticValue": true/false, "confidenceScore": 0.0–1.0, "contentDescription": "Kurze Beschreibung des Bildinhalts", "educationalElements": ["Element1", "Element2"], "visualType": "photo/diagram/chart/infographic/

```
illustration/screenshot/text/map/icon/other", "didacticFunctions": ["instruction", "explanation",  
"warning", "demonstration", "visualization", "overview", "example", "comparison"], "selfSuffici-  
ency": "self_explanatory/text_dependent/context_dependent", "cognitiveComplexity": "simple/  
moderate/complex", "evidence": ["Begründung 1", "Begründung 2"], "contextUsed": "none/file-  
name/alt_text/surrounding_text/caption"}
```

Userprompt:

Analysiere das folgende Bild hinsichtlich seines didaktischen Wertes im Kontext der Katastrophenvorsorge. Dateiname: {filename} Alt-Text: {alt_text} Bildunterschrift: {caption} Umgebender Text: {surrounding_text} Bewertungskriterien: 1. DIDAKTISCHER WERT (hasDidacticValue): TRUE wenn das Bild mindestens eines dieser Kriterien erfüllt: • Enthält Handlungsanweisungen oder Schrittfolgen • Visualisiert Prozesse, Abläufe oder Zusammenhänge • Zeigt Warnsymbole oder Sicherheitshinweise mit Informationsgehalt • Stellt Daten, Statistiken oder Vergleiche dar • Illustriert konkrete Techniken oder Methoden • Enthält beschriftete Diagramme oder Karten FALSE wenn: • Rein dekorativ (Stockfoto, Stimmungsbild) • Logo, Icon oder Navigationselement • Portraitfoto ohne informativen Kontext • Werbebanner oder Promotional Content 2. VISUAL TYPE: • photo: Fotografische Aufnahme • diagram: Schematische Darstellung, Flowchart • chart: Datenvisualisierung (Balken, Kreis, Linie) • infographic: Kombinierte Informationsgrafik • illustration: Gezeichnete/grafische Illustration • screenshot: Bildschirmaufnahme • text: Überwiegend Textinhalt als Bild • map: Karte oder Lageplan • icon: Symbol oder Icon • other: Sonstiges 3. DIDAKTISCHE FUNKTIONEN (wenn hasDidacticValue=true): • instruction: Anleitung zum Handeln • explanation: Erklärung von Konzepten • warning: Warnung oder Sicherheitshinweis • demonstration: Demonstration einer Technik • visualization: Visualisierung von Daten/Prozessen • overview: Überblick/Zusammenfassung • example: Konkretes Beispiel • comparison: Vergleichende Darstellung 4. SELBSTSTÄNDIGKEIT (selfSufficiency): • self_explanatory: Bild ist ohne Text verständlich • text_dependent: Benötigt Begleittext zum Verständnis • context_dependent: Benötigt spezifischen Kontext 5. KOGNITIVE KOMPLEXITÄT: • simple: Einfache, direkt erfassbare Information • moderate: Mehrere Informationsebenen • complex: Komplexe Zusammenhänge, erfordert Analyse Analysiere das Bild und antworte anschließend im JSON-Format.

Aggregationslogik:

Nach der Einzelbildanalyse werden die Ergebnisse wie folgt aggregiert: overallClassification: "didactic" wenn didacticPercentage $\geq 50\%$, "mixed" wenn $\geq 20\%$ und $< 50\%$, "decorative" wenn $< 20\%$. $\text{didacticPercentage} = (\text{didacticImages} / \text{analyzedImages}) * 100$. $\text{averageConfidence} = \text{sum}(\text{confidenceScores}) / \text{analyzedImages}$

5.6 Urheber-Analysator

Der Urheber-Analysator verwendet ein zweistufiges Verfahren:

1. Prüfung der MongoDB-Datenbank auf existierende Einträge für die Domain
2. Bei keinem Treffer: Web-Search via GPT-4o zur Recherche des Betreibers

Klassifikations-Prompt (nach Web-Search):

Basierend auf den folgenden Suchergebnissen, klassifiziere den Betreiber der Domain {domain} in GENAU EINE dieser Kategorien: – "staatlich" (Behörde, Ministerium, staatliche Einrichtung) – "nicht staatliche Hilfsorganisation" (NGO, Hilfswerk, gemeinnützige Organisation) – "sonstige Vereine" (Vereine, die keine Hilfsorganisationen sind) – "Organisationen" (internationale Organisationen, Verbände) – "Gemeinschaften" (informelle Gruppen, Communities, Netzwerke) – "Unternehmen" (kommerzielle Unternehmen, GmbH, AG, etc.) – "Privatperson" (von Einzelpersonen betriebene Webseiten) Suchergebnisse: {search_results} URL: {url} Antworte im JSON-Format: {"urheber": "KATEGORIE", "begründung": "Kurze Begründung für die Einordnung", "quellen": ["Quelle 1", "Quelle 2"]}

5.7 Herkunfts-Analysator

Der Herkunfts-Analysator ist regelbasiert (kein LLM) und bestimmt die Nationalität anhand der Top-Level-Domain:

TLD-Mapping: "de" → "Deutschland", "at" → "Österreich", "ch" → "Schweiz", andere → "unbekannt"

6. Entscheidungsregeln und Qualitätssicherung

6.1 Konservative Codierung (Grundregel)

Kategorien nur bei klarer Evidenz; Randthemen ignorieren; im Zweifel weniger Kategorien vergeben.

6.2 Dominanzkriterium (Instruktion vs. Information)

'Instruktion' nur, wenn Schrittfolge/Handlungssequenz den Hauptteil der Seite prägt; sonst 'Informationsressource'.

6.3 Evidenzmarker

Begründungen verweisen auf Textmarker und/oder DOM-/Metadaten (z. B. 'FAQ', 'Schritt 1–3', 'Kommentarbereich').

6.4 Umgang mit Unsicherheit

Bei fehlender Evidenz (dynamisch/verborgen) Kategorie 'unklar'/'sonstige'; Limitation dokumentieren.

6.5 Bildanalyse-Schwellenwerte

Ein Bild gilt als didaktisch wertvoll bei `confidenceScore >= 0.5` und `hasDidacticValue = true`. Die Gesamtklassifikation einer Webseite basiert auf dem Anteil didaktischer Bilder (siehe Aggregationslogik).

7. Kategoriensystem

Das Kategoriensystem zur inhaltlichen Klassifikation der Webseiten umfasst 13 Primärkategorien mit jeweils zugeordneten Unterkategorien. Mehrfachzuordnungen pro Webseite waren möglich; die Kategorien sind daher nicht als gegenseitig ausschliessend zu verstehen. Die Kategorien wurden für den Kategorien-Analysator als promptbasierte Klassifikationskategorien operationalisiert.

7.1 Primärkategorien

- Notfallvorsorge themenübergreifend
- Lebensmittelversorgung
- Unterkunft & Gebäudeschutz
- Energie & Infrastruktur
- Persönlicher Schutz & Gefahrenabwehr
- Information & Kommunikation
- Medizinische Versorgung
- Mobilität & Orientierung
- Psychologische Unterstützung
- Notfallplanung & Ausrüstung
- Naturkatastrophen
- Technologische Gefahren
- Krieg & Terror

7.2 Unterkategorien nach Primärkategorie

Primärkategorie	Unterkategorien
Notfallvorsorge themenübergreifend	Grundlagen der Notfallvorsorge, Behördliche Empfehlungen, Checklisten & Anleitungen, Präventive Maßnahmen
Lebensmittelversorgung	Bevorratung von Lebensmitteln & Wasser, Notfall-Lebensmittelzubereitung, Autarke Nahrungsmittelproduktion, Haltbarmachung & Konservierung, Wasseraufbereitung & -filtration
Unterkunft & Gebäudeschutz	Notunterkünfte errichten & finden, Gebäudeschutz vor Naturgefahren, Heizen & Kühlen im Notfall, Alternative Wohnformen, Hygiene & Sanitärversorgung
Energie & Infrastruktur	Blackout-Vorsorge, Energieversorgung im Notfall, Alternative Energiequellen, Energiespeicherung, Kritische Infrastruktur
Persönlicher Schutz & Gefahrenabwehr	Allgemeine Schutzmaßnahmen, Selbstverteidigung, Brandschutz, ABC-Gefahrenabwehr, Survival-Techniken, Schutz vor kriminellen Handlungen, Schutz vor terroristischen Handlungen, Schutz bei kriegerischen Auseinandersetzungen
Information & Kommunikation	Risiko- und Bedrohungsanalysen bei Krieg und Terror, Behördliche Warnsysteme & Lageberichte, Notfallkommunikation, Alternative Kommunikationsgeräte, Informationsquellen im Krisenfall, Digitale Sicherheit, Dokumentensicherung
Medizinische Versorgung	Erste Hilfe & Notfallmedizin, Medikamentenbevorratung, Medizinische Ausrüstung, Hygiene & Infektionsschutz, Chronische Erkrankungen im Krisenfall
Mobilität & Orientierung	Evakuierungsplanung, Navigation ohne technische Hilfsmittel, Fahrzeugvorbereitung für Notfälle, Fluchtrouten & Treffpunkte
Psychologische Unterstützung	Psychologische Selbsthilfe, Krisenintervention, Gemeinschaftliche Unterstützungssysteme, Umgang mit Trauma, Stressbewältigung
Notfallplanung & Ausrüstung	Notfallpläne entwickeln, Notfallrucksack & Bug-Out-Bag, Technische Hilfsmittel, Werkzeuge & Reparatur, Notfallübungen & Training, Notfallausrüstung für den Zivilschutz
Naturkatastrophen	Hochwasser & Starkregen, Sturm & Unwetter, Erdbeben, Waldbrände, Extremwetterereignisse, Vulkanausbrüche
Technologische Gefahren	Cybergefahren, Industrieunfälle, Infrastrukturausfall, Technische Störfälle, Gefahrstoffe
Krieg & Terror	Kriegerische Auseinandersetzungen, Hybride Kriegsführung, Terroranschläge

8. Datenbank-Schema (MongoDB/Mongoose)

Die folgenden TypeScript-Definitionen dokumentieren zentrale Datenstrukturen der Analysepipeline, die in MongoDB/Mongoose zur Speicherung der gescrapten Webseiten, Metadaten und Klassifikationsergebnisse verwendet wurden. Sie zeigen die für die Reproduzierbarkeit relevanten Felder, Enumerationen und Ergebnisstrukturen.

8.1 Enumerationen (Enums.ts)

Die Enumerationen definieren die zulässigen Werte für die Klassifikationsfelder.

```
enum Urheber {  STAATLICH = «staatlich»,  NICHT_STAATLICHE_HILFSORGANISATION = «nicht staatliche Hilfsorganisation»,  SONSTIGE_VEREINE = «sonstige Vereine»,  ORGANISATIONEN = «Organisationen»,  GEMEINSCHAFTEN = «Gemeinschaften»,  UNTERNEHMEN = «Unternehmen»,  PRIVATPERSON = «Privatperson» }

enum Format {  TEXTBASIIERT = «textbasiert»,  VIDEO = «Video»,  AUDIO = «Audio»,  BILDER = «Bilder»,  MULTIMEDIA = «Multimedia»,  SONSTIGE = «Sonstige» }

enum Zugriffsbeschränkung {  FREI = «frei»,  EINGESCHRÄNKT = «eingeschränkt»,  KOSTENPFLICHTIG = «kostenpflichtig» }

enum Kommunikationsart {  EINSEITIG = «einseitig»,  EINGESCHRÄNKT_INTERAKTIV = «eingeschränkt interaktiv»,  INTERAKTIV = «interaktiv»,  SONSTIGE = «Sonstige» }

enum DidaktikFormat {  SELBSTLERNEINHEIT = «Selbstlerneinheit»,  INSTRUKTION = «Instruktion»,  KOLLABORATION = «Kollaboration»,  SONSTIGE = «Sonstige» }
```

8.2 Metadata-Schema (Metadata.ts)

Das Metadata-Schema speichert die extrahierten Meta-Informationen einer Webseite.

```
interface IMetadata {  title: string;  description?: string;  language: string;  robots?: string;  ogTitle?: string;  ogDescription?: string;  ogUrl: string;  ogImage?: string;  ogLocale?: string;  ogLocaleAlternate?: string[];  ogSiteName?: string;  ogVideo?: string[];  modifiedTime?: Date;  publishedTime?: Date;  viewport?: string;  sourceURL: string;  // Required, unique  url?: string;  statusCode?: number; }
```

8.3 Internal-Schema (Internal.ts)

Das Internal-Schema verwaltet den Bearbeitungsstatus jedes Dokuments.

```
enum Status {  NEW = «new»,  WAIT = «waiting»,  CLOSED = «closed» }  interface IInternal {  status: Status;  notes: string; }
```

8.4 Hauptdokument-Schema (ScrapedWebpage.ts)

Das zentrale Schema speichert die gescrapte Webseite, Metadaten sowie manuelle und KI-gestützte Klassifikationsergebnisse.

```
interface IScrapedWebpage {  // Core Data  metadata: IMetadata;  html: string;  links: string[];  internal: IInternal;  scrape_id: string;  // Unique identifier  domain: string;  topLevelDomain: string;  subdomain: string;  markdown: string;  // Extracted content  inputURL?: string; }
```

```
Manual Classifications  nationalitaet?: string;  themen?: string[];
subThemen?: Map<string, string[]>;  urheber?: string;  medienformat?: string[];
zugriff?: string;  zugriffsBeschraenkung?: string;  kommunikationsart?:
string;  ressourcenArt?: string;  // AI Classifications (prefixed
with <ai>)  aithemen?: string[];  aisubThemen?: Map<string, string[]>;
ainationalitaet?: string;  aiurheber?: string;  aimedienformat?: string[];
aikommunikationsart?: string;  airessourcenArt?: string;  // Image Analysis
Results  imageAnalysis?: {  images: ImageAnalysisResult[];  summary:
ImageAnalysisSummary;  processingStatus: «completed» | «partial» | «failed»;
lastProcessedAt: Date;  };  // Timestamps (auto-generated)  createdAt:
Date;  updatedAt: Date; }
```

8.5 Bildanalyse-Schema

Das Bildanalyse-Schema speichert die Ergebnisse der visuellen Bildanalyse und deren Aggregation auf Webseitenebene.

```
interface ImageAnalysisResult {  filename: string;  md5Hash: string;
hasDidacticValue: boolean;  confidenceScore: number;  // 0.0 - 1.0
contentDescription: string;  educationalElements: string[];  visualType:
«photo» | «diagram» | «chart» | «infographic» | «illustration»
| «screenshot» | «text» | «map» | «icon» | «other»;
didacticFunctions: («instruction» | «explanation» | «warning» |
«demonstration» | «visualization» | «overview» | «example»
| «comparison»)[];  selfSufficiency: «self_explanatory» | «text_dependent» |
«context_dependent»;  cognitiveComplexity: «simple» | «moderate» | «complex»;
evidence: string[];  contextUsed: «none» | «filename» | «alt_text» |
«surrounding_text» | «caption»;  analyzedAt: Date;  isReused: boolean; }
interface ImageAnalysisSummary {  totalImages: number;  analyzedImages: number;
didacticImages: number;  didacticPercentage: number;  averageConfidence:
number;  overallClassification: «decorative» | «mixed» | «didactic»;
aggregatedAt: Date;  reusedCount: number; }
```

8.6 Datenbank-Indizes

Für optimierte Abfragen wurden folgende Indizes definiert:

- metadata.language (für Sprachfilterung)
- metadata.orgUrl
- metadata.sourceURL (unique)
- createdAt (descending, für chronologische Sortierung)