PROF. DR.-ING. DAVID ZELLHÖFER

DER DATENATLAS –
ANALYSE UND BEWERTUNG DER BEREITGESTELLTEN FUNKTIONEN DER INFORMATIONSSUCHE

PROFESSUR FÜR DIGITALE INNOVATION
IN DER ÖFFENTLICHEN VERWALTUNG –
HOCHSCHULE FÜR WIRTSCHAFT UND RECHT BERLIN,
FACHBEREICH 3: ALLGEMEINE VERWALTUNG

HERAUSGEBER: PROFESSUR FÜR DIGITALE INNOVATION IN DER ÖFFENTLICHEN VERWALTUNG – HOCHSCHULE FÜR WIRTSCHAFT UND RECHT BERLIN, FACHBEREICH 3: ALLGEMEINE VERWALTUNG

DOI: 10.5281/zenodo.17780441

Hinweis zur geschlechterneutralen Formulierung: Der Autor bemüht sich inklusive Formulierungen zu verwenden. Trotz aller Sorgfalt kann es jedoch im Einzelfall zur Nutzung männlicher Wortformen kommen. In diesem Fall beinhalten diese dann auch sämtliche Geschlechter.

Bearbeitungszeitraum: Mitte Juli-Ende August 2025 – Publikation: Berlin im November 2025

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Seit ich des Suchens müde ward,

Erlernte ich das Finden.

Friedrich Nietzsche,

Die fröhliche Wissenschaft; 1882

.

.

1
Executive Summary

Der Datenatlas wird als eine der Maßnahmen umgesetzt, die sich aus der Datenstrategie der letzen Bundesregierung ergeben. Die Implementierung erfolgt ausschließlich durch die Bundesdruckerei unter Federführung des Bundesministeriums der Finanzen (Bmf).

Beim Datenatlas handelt es sich um ein sogenanntes Metadaten-Portal, das eine Vielzahl von zusätzlichen Funktionen, wie die Verwaltung von Datenschemata oder einen explorativen Zugang zu den Datenbeständen der Bundesverwaltung bieten soll.

Die Software-Entwicklung wird dabei von einer Dateninventur in den Bundesbehörden begleitet.

Das Projekt setzte sich laut Pressemeldungen von Anfang an ambitionierte Ziele, die Implementierung kann jedoch nur weniger überzeugen. Diese erreicht die verkündeten Ziele bei weitem nicht.

Generell gesprochen verfehlt das Software-Entwicklungsprojekt den *Stand der Technik* in vielen Bereichen.

Funktionsumfang

Entgegen der eigenen Zielsetzung wird keine explorative Suche in Form des Browsings umgesetzt.

Die implementierte, gerichtete Schlagwort-Suche bietet starkes Verbesserungspotential und wird den Bedürfnissen der Mehrheit der Beschäftigten in der Bundesverwaltung nicht gerecht.

Der Verzicht auf das Browsing steht im krassen Widerspruch zu wissenschaftlichen Erkenntnissen auf dem Gebiet der *Informationsrecherche* und wird die Nutzbarkeit weiter einschränken.

Im Bereich der Anfragemöglichkeiten fällt der Datenatlas *hinter das Frühjahr 1986 zurück.* Er verfehlt damit erheblich die Ansprüche der Verwaltungsmodernisierung.

Hinzu kommen diverse Mängel im Bereich der *Usability*, die einerseits die Nutzung erschweren und andererseits dazu führen können, die Datenqualität im Datenatlas zu verringern.

Datenqualitätssicherung

Der DATENATLAS verfügt zum Begutachtungszeitpunkt (Juli 2025) über keinerlei wirksam implementierte Mechanismen der *Datenqualitätssicherung*.

Weder ist eine automatisierte Validierung von Datensätzen möglich noch werden Fehleingaben verhindert. Diese ließen sich mittels Methoden auf dem *Stand der Technik*, wie *kontrollierten Vokabularen* oder *Linked (Open) Data*, leicht vermeiden.

Ohne eine wirkungsvolle Datenqualitätssicherung rückt das Ziel, den DATENATLAS mit Anwendungen der *Künstlichen Intelligenz* zu verbinden, in weite Ferne.

Nachnutzbarkeit und Künstliche Intelligenz

Zum jetzigen Zeitpunkt zeigt sich der DATENATLAS wenig *interoperabel*. Das heißt, wenn seine Nutzung trotz der im Gutachten aufgezeigten Mängel verpflichtend würde, führte dies unmittelbar zu der Bildung eines weiteren, wenig interoperablen *Datensilos*.

Ferner ist davon auszugehen, dass die im DATENATLAS vorgehaltenen Datensätze kaum – wenn überhaupt – *maschinenlesbar* sind und sich dementsprechend schlecht für Anwendungen der *Künstlichen Intelligenz* oder für sonstige Nachnutzungsszenarien wie den *interoperablen Datenaustausch* eignen werden.

Eine für die BUNDESVERWALTUNG gewinnbringende Nutzungsperspektive des DATENATLAS im Bereich der *Künstlichen Intelligenz* ist aktuell nicht zu erkennen.

Digitale Souveränität

Die Digitale Souveränität der BUNDESVERWALTUNG wird durch den DATENATLAS gemindert. Durch den entstehenden *vendor lock-in* und die Bereitstellung einer *Closed-Source-Software* begibt sich die BUNDESVERWALTUNG in ein langfristiges Abhängigkeitsverhältnis zur BUNDESDRUCKEREI.

Der *vendor lock-in* wirkt sich ebenfalls auf die verwalteten Daten aus. Das kann hier im schlimmsten Fall sogar den *Totalverlust* dieser bedeuten, wenn nämlich keine standardisierte Datenhaltung implementiert wurde und der Dienstleister nicht länger für den Datenexport zur Verfügung steht.

Wirtschaftlichkeit

Auch wirtschaftlich gesehen ist die Entscheidung, in eine proprietäre Software zu investieren nicht nachvollziehbar.

So bieten typische *Open-Source-Lösungen* wesentlich mehr Funktionen und erreichen den *Stand der Technik*, wie am Open-Data-Portal des Bundes, GOVDATA, deutlich wird.

Entwicklungsperspektive

Das vorliegende Gutachten nennt diverse Verbesserungsmöglichkeiten und Umsetzungsmöglichkeiten, die Dienstleister für ihre Arbeit am DATENATLAS heranziehen könnten. Dass der aktuelle Dienstleister in der Lage ist, damit ein Recherche-Werkzeug auf dem *Stand der Technik* zu implementieren, scheint ausgeschlossen.

EINE WEITERENTWICKLUNG oder ein Refactoring des DATENATLAS würde diverse Kernfunktionen der Software betreffen und lässt sich – auch aufgrund der bereits existierenden und funktional besseren *Open-Source-Lösungen* – kaum wirtschaftlich darstellen.

LETZTENDLICH ist eine Neuentwicklung des DATENATLAS unumgänglich, da aufgrund seiner leicht zu entdeckenden Defizite erhebliche Zweifel an der Eignung des Dienstleisters für die Entwicklung von Software-Lösungen der *Informationsrecherche* bestehen.

AUFGRUND DER EKLATANTEN MÄNGEL ist das Software-Entwicklungsprojekt DATENATLAS mit *sofortiger Wirkung* zu stoppen, um nicht weitere Mittel in eine technisch und konzeptionell wenig überzeugende Lösung zu investieren, welche kaum den *Stand der Technik* erreicht.

Beim Vorliegen eines Werkvertrags ist die Abnahme entsprechend zu verweigern.

EINE INTEGRATION des DATENATLAS in den *Deutschland-Stack* verbietet sich sowohl aus Überlegungen, welche die *Digitale Souveränität* betreffen, als auch angesichts der existierenden Mängel.

Als konstruktives Ergebnis des Projekts bleibt eigentlich nur eine Dateninventur zu nennen, vorausgesetzt diese ist repräsentativ und umfassend durchgeführt worden.

DAS VORLIEGENDE GUTACHTEN wurde *pro bono* erstellt.

Der Autor möchte insbesondere den MASTODON-Nutzern Daniel Baránek (daelba@sciences.social) und Jan Ainali (ainali@social.coop) für die SPARQL-Unterstützung im Rahmen des illustrativen Beispiels in Anhang A.4 danken.

Über den Autor

Prof. Dr.-Ing. David Zellhöfer lehrt Digitale Innovation in der öffentlichen Verwaltung an der Hochschule für Wirtschaft und Recht Berlin und ist Lehrbeauftragter für das Modul „Digitale Informationsinfrastrukturen“ am Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin.

Neben der ganzheitlichen, menschzentrierten Digitalisierung öffentlicher Einrichtungen widmet er sich der agilen Transformation von Organisationen und dem Kompetenzerwerb für die

erfolgreiche digitale Transformation – insbesondere mit Hinblick auf Anwendungen der Künstlichen Intelligenz und den damit verbundenen Herausforderungen des Informations- und Datenmanagements. Durch seine Promotion an der Schnittstelle zwischen Medieninformatik und Informationswissenschaft greift er hierfür auf einen interdisziplinären Methodenkoffer zurück.

Nach Stationen in der Software-Entwicklung, der Unternehmensberatung und der Bundesverwaltung war Prof. Dr.-Ing. Zellhöfer als wissenschaftlicher Direktor in der Abteilung Informations- und Datenmanagement der Staatsbibliothek zu Berlin tätig und setzte dort seine Expertenkenntnisse im Bereich Datenmanagement und Datenkuration praktisch in hochskalierende, verteilte Informationssysteme um.

Im Jahr 2018 baute er in Vorbereitung und im Rahmen des BMBF-geförderten Projekts „QURATOR – Curation Technologies“ einen der ersten professionellen KI-Cluster im deutschsprachigen, wissenschaftlichen Bibliothekssektor auf.

Er blickt auf eine mehr als 25 Jahre andauernde Karriere im Feld der Informatik, Software-Entwicklung und des Informations- und Datenmanagements zurück.

Im Bereich der Forschung verfügt Prof. Dr.-Ing. Zellhöfer über Expertise im Bereich des Multimedia-Information-Retrievals und des User-centered Designs. Er ist Autor von mehr als 100 Veröffentlichungen und Vorträgen zu den Bereichen digitale Innovationskompetenzen, Open Data, Künstliche Intelligenz, Information Retrieval und Datenmanagement.

Er ist Mitglied der Fachgruppe „Informatik und Informationswissenschaft“ des IVVI (Institut für Verwaltungsforschung und Verwaltungsinnovation) und des d-cube (Institute for Data-Driven Digital Transformation) der HWR Berlin. Er engagiert sich in der Gesellschaft für Informatik e.V. u.a. in den Fachgruppen „Verwaltungsinformatik“, „Information Retrieval“ und „Informatik und Ethik“. Er ist Sprecher für den „Arbeitskreis Low Code / No Code in der öffentlichen Verwaltung“ des NEGZ · Kompetenznetzwerk Digitale Verwaltung und Mitglied weiterer Arbeitskreise wie „Kompetenzen & Lernen“ und „Design Thinking in der öffentlichen Verwaltung“.

Prof. Dr.-Ing. Zellhöfer berät primär Einrichtungen der Öffentlichen Verwaltung und aus dem zivilgesellschaftlichen Non-Profit-Sektor zu den oben genannten Themen.

Das vorliegende Gutachten wurde *pro bono* erstellt, da er den Aufbau von Datenkompetenzen, die nachhaltige Verwaltung und Kontextualisierung von Daten, z.B. in Form von *Linked (Open) Data*, und den Austausch dieser als essenziell für die Zukunftsfähigkeit

der Öffentlichen Verwaltung – nicht nur in Hinblick auf die bedarfsgetriebene Anwendung von Künstlicher Intelligenz – betrachtet.

Folgerichtig ist er Co-Autor des Vorschlags zur Novellierung der ÖV-Ausbildung[1] für die Innenministerkonferenz im Rahmen der Digitalen Transformation.

[1] Rektorenkonferenz d. HS f. d. öffentl. Dienst (2025)

2
Einleitung

Der Datenatlas wird als eine der Maßnahmen, die sich aus der Datenstrategie der letzen Bundesregierung[1] ergeben, umgesetzt. Die Implementierung erfolgt ausschließlich durch die Bundesdruckerei unter Federführung des Bundesministeriums der Finanzen (Bmf)[2]. Frei verfügbare Informationen zum Entwicklungsprozess sind außerhalb der Bundesverwaltung kaum oder nur schwer zugänglich und auch im Rahmen der journalistischen Arbeit des Autors nicht abschließend ermittelbar[3], so dass sich im Folgenden auf Pressemitteilungen und Strategiepapiere der zuständigen Bundesministerien verlassen werden muss. Diese Aussagen werden durch Berichte von Mitarbeitenden der Datenlabore und weiteren Personen, die im Rahmen des Projektverlaufs mit der Softwarelösung konfrontiert wurden, gestützt, welche auf Fachtagungen und Empfängen öffentlich getätigt wurden.

Die Anforderungen an den Datenatlas werden von der Bundesdruckerei mit Rückgriff auf die Datenstrategie wie folgt umrissen:

> Wir erstellen einen Datenatlas Bundesverwaltung, der Daten aller Ministerien und ihrer Geschäftsbereiche auf Metadatenebene zeigt. Damit schaffen wir Transparenz über den vorhandenen Datenbestand. [...] Der Datenatlas nutzt und ergänzt bestehende Verwaltungsdatenübersichten wie die Verwaltungsdaten-Informationsplattform (VIP) des Statistischen Bundesamtes, die Registerlandkarte des Bundesverwaltungsamtes oder das Metadatenportal GovData zu offenen Daten von Bund, Ländern und Kommunen. Für den Datenatlas sind in den Ministerien die Datenlabore zuständig.[4]

Die Entscheidung zum Aufbau eines internen Datenatlas für die Bundesverwaltung ist prinzipiell nachvollziehbar und sinnvoll. Ähnliche Initiativen finden sich in der Zivilgesellschaft[5] oder in der deutschen Verwaltung in Form des Datenportals GovData[6] für den Bereich *Open Data*. Generell können solche Lösungen durch ihren Charakter als One-Stop-Shop als nutzerorientiert betrachtet werden, wenn sie die zentralen Use Cases (s.u.) der Nutzenden implementieren. Die Bereitstellung einer solchen Anwendung ist zentral für Daten-getriebene Entscheidungen der Bundesverwaltung und die Implementierung von Anwendungen der *Künstlichen Intelligenz* (KI).

[1] Bundesministerium für Digitales und Verkehr, Bundesministerium für Wirtschaft und Klimaschutz, und Bundesministerium des Innern und für Heimat, Hrsg. *Fortschritt durch Datennutzung - Strategie für mehr und bessere Daten für neue, effektive und zukunftsweisende Datennutzung*. Die Bundesregierung, 2023. https://tinyurl.com/datenstrategiede

[2] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025

[3] David Zellhöfer. Der Begriff „Leuchtturm" ist in der Regel ein Warnsignal. *Der Tagesspiegel (23.05.2023)*, Seite B 24, 2023. https://tinyurl.com/kolumne-tsp. Letzter Abruf: 24.07.2025

[4] Das Erstellungsdatum der Website wurde aufgrund mangelnder bzw. widersprüchlicher Metadaten durch das verlinkte YouTube-Video auf den 08.05.2025 datiert.

Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025

[5] Datenatlas Zivilgesellschaft (betrieben durch die Bertelsmann Stiftung); https://datenatlas-zivilgesellschaft.de/; Letzter Abruf: 24.07.2025

[6] GovData ; https://www.govdata.de; Letzter Abruf: 21.07.2025

In einer weiteren Pressemitteilung aus dem Jahr 2022 beschreibt die Bundesdruckerei den Funktionsumfang anhand eines Mock-Ups (siehe Abb. 2.1) wie folgt:

> „Im Datenatlas werden Meta-Informationen *[sic!]* zu Datenbeständen der öffentlichen Verwaltung übersichtlich dargestellt und graphisch erkundbar gemacht. KI-Methoden unterstützen dabei, Informationen zu extrahieren, Verbindungen zwischen Elementen aufzuzeigen und die Nutzenden schnell zum gewünschten Suchergebnis zu führen.“[7]

[7] Bundesdruckerei. *Erstes Vollständiges Datenmodell Der Bundesverwaltung - Pressemitteilung*, 2022. https://tinyurl.com/bdr-pm3. Letzter Abruf: 21.07.2025

Aus diesen Eigendarstellungen wird deutlich, dass es sich beim Datenatlas um ein sogenanntes Metadaten-Portal[8] handelt, welches eine Vielzahl von zusätzlichen Funktionen, wie die Verwaltung von Datenschemata[9] oder das Browsing[10] in Metadaten-Beständen, bieten soll. Hierbei handelt es sich um Funktionen, die dem Stand der Technik derartiger IT-Systeme entsprechen.

[8] Siehe Abschnitt 2.2.

[9] Siehe Abschnitt 3.6.

[10] Siehe Abschnitt 3.5.

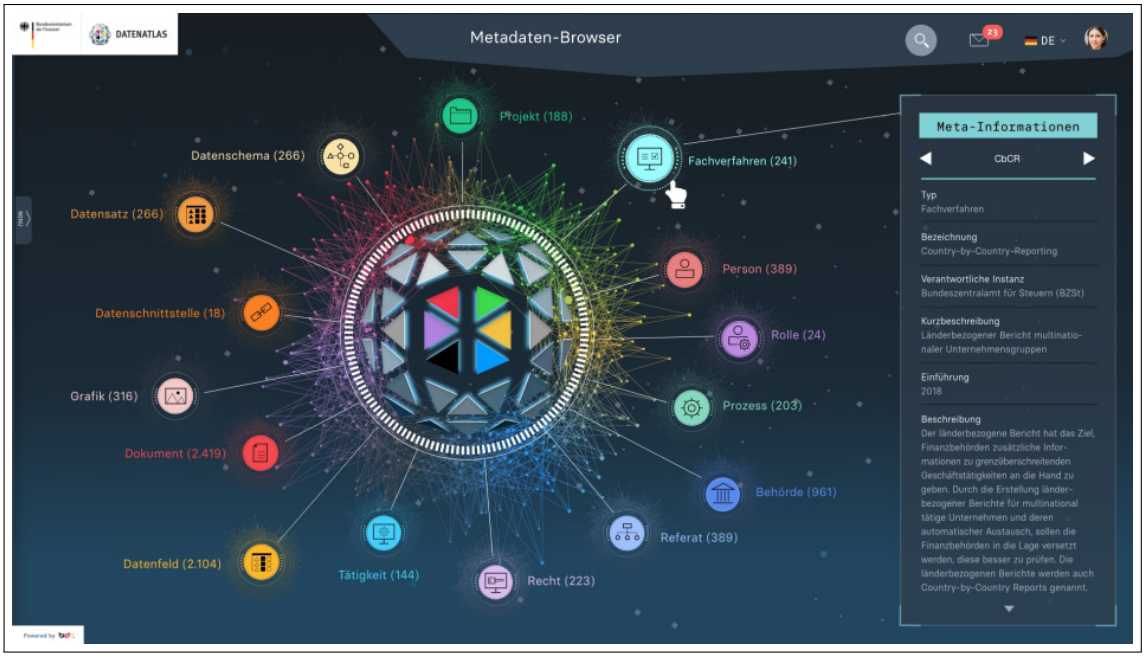


Abbildung 2.1: Visualisierung des Datenatlas aus der Pressemitteilung der Bundesdruckerei vom 26.10.2022

Der Begriff des Metadaten-Portals entzieht sich aufgrund seiner Unschärfe einer zielführenden Begutachtung.

Stattdessen lohnt sich die Betrachtung *zweier unterschiedlicher Minimal-Use-Cases*[11], die der Datenatlas unterstützen muss, um den Grundbedarf der Bundesverwaltung decken zu können:

[11] *Anwendungsfälle*; d.h. eine Zusammenfassung von Szenarien, die es Nutzenden ermöglicht ein spezifisches Ziel zu erreichen.

Datenatlas – Minimale Use Cases der Bundesverwaltung

1. die *Recherche von Metadaten*, welche vom Großteil der Beschäftigten der Bundesverwaltung genutzt wird, und
2. die *Erfassung von Metadaten*, welche durch spezialisiertes Personal erfolgen dürfte.

Hinzu kommen weitere Anforderungen, die sich vor allem auf die Speicherung der Daten und das Retrieval[12] dieser auswirken. Auf diese Aspekte wird in Kapitel 3 näher eingegangen.

[12] Der Abruf von Daten aus einer Datenbank mittels einer Anfrage.

Im Projektverlauf wurde eine Dateninventur in Bundesministerien und ihrer nachgeordneten Behörden durch die Bundesdruckerei [13] durchgeführt.

Dieses Vorgehen ist begrüßenswert, da offensichtlich ist, dass die deutsche Bundesverwaltung aktuell über keinen umfassenden Überblick über existierende Datenbestände oder deren Datenqualität (siehe Abschnitt Datenqualität und -semantik) verfügt.

Die Durchführung einer Dateninventur flankiert deshalb die seit langem notwendige Etablierung der Datenlabore (siehe unten).

Es ist jedoch unbekannt, in welchem Umfang und welcher Tiefe diese Dateninventur stattgefunden hat und inwiefern ein Erhebungsgrad bestimmt werden kann.

Die Projektumsetzung in den einzelnen Verwaltungseinheiten erfolgt dreistufig in Form der Initialisierung, der Datenerhebung und dem finalen Einsatz der Anwendung[14].

[13] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025

[14] Ebenda.

Inwiefern die Datenlabore als zentraler Anforderungsgeber mit ihrer fachlichen Expertise in die menschzentrierte Entwicklung[15] des Datenatlas einbezogen wurden, ist aufgrund der Aussagen einzelner Vertreter der Datenlabore kaum zu bestimmen. Hier ergibt sich ein diverses Meinungsbild[16].

Es ist jedoch naheliegend, dass die Datenlabore als Kompetenzzentren für die Daten-getriebene Verwaltungsmodernisierung einerseits einen wichtigen Kreis an Nutzenden als auch einen zentralen Stakeholder für das Projekt Datenatlas darstellen.

Laut der Pressemitteilung der Bundesdruckerei [17] wird ein „nutzerzentrierter Datenatlas“ konzipiert. Der Datenatlas soll dabei einen „Meilenstein auf dem Weg zur datengetriebenen Verwaltung“ und die „Basis für künftige Datenanalysen und KI-Anwendungen wie Maschinelles Lernen“[18] darstellen.

[15] Siehe Abschnitt 3.4.

[16] Siehe Abschnitt 5.3.

[17] Bundesdruckerei. *Erstes Vollständiges Datenmodell Der Bundesverwaltung - Pressemitteilung*, 2022. https://tinyurl.com/bdr-pm3. Letzter Abruf: 21.07.2025

[18] Siehe Abschnitt 5.6.

Nach Erkenntnissen des Autors wird aktuell diskutiert, den Datenatlas auch in anderen Einrichtungen einzuführen, deren mitunter spezifische Anforderungen im bisherigen Projektverlauf nicht betrachtet wurden.

Nach dem Kenntnisstand des Autors vom 01.07.2025 soll der Betrieb des Datenatlas an das Bundesministerium für Digitales und Staatsmodernisierung (Bmds) übergehen.

Es ist notwendig, die Qualität des Datenatlas zum aktuellen Stand (Mitte Juli 2025) im Rahmen eines Gutachtens zu prüfen. Dabei soll der Funktionsumfang mit Blick auf den aktuellen Forschungsstand der Informatik und Informationswissenschaften sowie auf den *Stand der Technik* beurteilt werden.

2.1 *Aufbau des Gutachtens*

Lesehinweise zum Gutachten

Das vorliegende Gutachten ist in seiner thematischen Tiefe und seinem Argumentationsniveau ungefähr auf dem Bachelor-Niveau von Informatik- oder Informationswissenschaften-Studierenden angesiedelt und sollte deshalb auch für die angenommene Zielgruppe in der Bundesverwaltung gut verständlich sein.

Für weniger technisch versierte oder zeitlich ausgelastete Leserinnen und Leser bietet sich zumindest die Konsultation der folgenden Abschnitte an, um der grundlegenden Argumentation des Autors folgen zu können.

Dabei wird empfohlen, mit dem folgenden Abschnitt

- Grundbegriffe des Informationsmanagements zu beginnen. Gefolgt von:
- Kapitel 3: Abschnitt Informationsrecherche,
- Kapitel 3: Abschnitt Datenqualität und -semantik,
- Kapitel 3: Abschnitt Regulatorische Grundlagen und
- Kapitel 4: Exemplarische User Journeys und Einordnung. Danach sollte zuerst das
- Kapitel 6: Fazit und Ausblick gelesen werden, da es konkrete Handlungsempfehlungen beinhaltet, die eine direkte Auswirkung auf die Umsetzung der *Desiderata* haben, um dann mit
- Kapitel 5: Bewertung und Desiderata abzuschließen.

Zum einfacheren Auffinden der Schlüsselbegriffe findet sich ein umfassender Index am Ende des Gutachtens.

Im folgenden Abschnitt werden die zentralen Grundbegriffe des Informationsmanagements im Kontext des DATENATLAS definiert.

DER STAND DER TECHNIK in einigen für den DATENATLAS relevanten Feldern wie der Informationsrecherche oder der Datenqualität und -semantik wird in Kapitel 3 thematisiert. Regulatorische Grundlagen werden ebenfalls benannt.

Der Abschnitt Information-Retrieval-Modelle stellt die wissenschaftlichen Erkenntnisse in diesem Forschungsfeld seit den 1970er-Jahren dar. Abschnitt 3.4 präsentiert wissenschaftliche Erkenntnisse und praktische Maßnahmen, welche umgesetzt werden müssen, um Anwender bestmöglich bei der Informationsrecherche zu unterstützen.

Das Kapitel benennt außerdem die Limitationen des Gutachtens.

KAPITEL 4 zeigt die Leistungsfähigkeit des DATENATLAS anhand exemplarischer User Journeys zweier Personas auf, die sich an den minimalen Use Cases der Bundesverwaltung (siehe S. 20) orientieren. Die aus den User Journeys gewonnenen Erkenntnisse werden begleitend diskutiert und eingeordnet, um die Verständlichkeit des Gutachtens zu erhöhen.

DIE BEWERTUNG UND DIE ABLEITUNG VON DESIDERATA bezüg-

lich des Funktionsumfanges des DATENATLAS erfolgt in Kapitel 5. Das Gutachten schließt mit einem Fazit und Ausblick, das auch konkrete Handlungsempfehlungen enthält.

2.2 *Grundbegriffe des Informationsmanagements*

Auch wenn Begriffe wie „Daten“, „Metadaten“ oder „Datenbanken“ mittlerweile häufig im Alltag Verwendung finden, ist es notwendig diese und weitere Begriffe korrekt zu definieren, um diese im Verlauf des Gutachtens präzise voneinander abgrenzen zu können. Der Aufbau und die Definition der zentralen Grundbegriffe orientiert sich, wenn nicht anders angeben, an Zellhöfer (2015).

Daten

Als Daten im Sinne der Informatik versteht man im Allgemeinen „zum Zweck der Verarbeitung zusammengefasste Zeichen, die aufgrund bekannter oder unterstellter Abmachungen Informationen (d.h. Angaben über Sachverhalte und Vorgänge) darstellen.“[19] Aus dieser Definition wird deutlich, dass Daten alleine noch keinen pragmatischen „Wert“ – also einen Bezug zum konkreten *Informationsbedürfnis*[20] – für Nutzende haben, wie Abb. 2.2 zeigt.

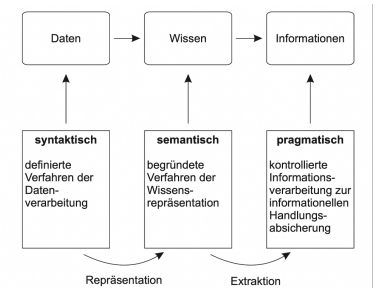


Abbildung 2.2: Abgrenzung der Begriffe Daten, Wissen und Information; Abb. 1.6 (Henrich, 2008)

[19] Lexikon-Redaktion des Gabler Verlages, Hrsg. *Gabler Kompakt-Lexikon Wirtschaft*. Gabler Kompakt-Lexikon Wirtschaft. Springer Gabler, 11. Auflage, 2013. ISBN 978-3-658-00008-0

[20] Siehe Abschnitt 2.2.

Metadaten

Folgt man dem Kompetenzzentrum Open Data (2023) so sind digital erfasste Metadaten

> strukturierte Daten, die Informationen über andere Informationsressourcen enthalten. Ein Merkmal von Metadaten ist, dass sie maschinell lesbar und auswertbar sind. Die beschreibenden Metadaten liefern die nötigen Informationen, um den Inhalt des Datensatzes darzustellen, ohne den Datensatz selbst öffnen zu müssen. Metadatenstandards vereinheitlichen die auszufüllenden Felder, um ein einheitliches Qualitätsniveau zu erreichen.[21]

Diese Definition stellt klar, dass Metadaten neben einem klaren Verwendungszweck über weitere wichtige Eigenschaften wie die *Maschinenlesbarkeit* und eine bestimmte *Datenqualität* verfügen müssen. Abschnitt 3.6 widmet sich letztgenanntem Thema detailliert. Die Nutzung von Metadaten-Standards ist essenziell, um die Interoperabilität von Systemen ermöglichen zu können.

[21] Kompetenzzentrum Open Data. *Leitfaden Metadaten - Version 2.0*, Bundesverwaltungsamt, 2023. https://tinyurl.com/leitfaden-md. Letzter Abruf: 24.07.25

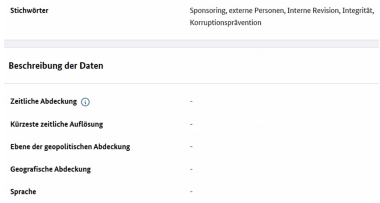


Abbildung 2.3: Auszug aus den Metadaten eines Beispieldatensatzes des DATENATLAS

DIE BEGRIFFE DATEN UND METADATEN sind nicht immer trennscharf abzugrenzen, da der Begriff Metadaten immer vom Standpunkt des Betrachters bzw. des Einsatzgebiets abhängt. Im Falle eines Metadaten-Portals handelt es sich bei den jeweiligen Feldbezeichnern um die Metadaten und bei deren konkreter Ausprägung (dem Inhalt) um die eigentlichen Daten. Dieser Festlegung soll im Rahmen dieser Arbeit gefolgt werden.

Abb. 2.3 zeigt einen beispielhaften Datensatz aus dem DATENATLAS. Hier handelt es sich bei den Datenfeldern „Stichwörter“ oder

„Sprache“ um Metadaten-Felder, die im Falle der Stichwörter mit den folgenden Daten befüllt sind: „Sponsoring, externe Personen, [...]“.

Datenbanken und Datenbankmanagement-Systeme

Unter einer Datenbank (DB) versteht man gemeinhin eine „Sammlung von Daten, die einen Ausschnitt der realen Welt beschreiben“[22]. Ein Softwaresystem, welches den Nutzenden die Erstellung, Pflege und Recherche von Datenbanken ermöglicht, wird als *Datenbankmanagement-System* (DBMS) bezeichnet.

[22] Thomas Kudraß und Thomas Brinkhoff, Hrsg. *Taschenbuch Datenbanken*. Fachbuchverl. Leipzig im Carl-Hanser-Verlag, 2007. ISBN 978-3-446-40944-6

Beginnend in den 1970er-Jahren schlugen Codd (1970) und Date (1982) relationale Datenbankmanagement-Systeme vor, die einen Paradigmenwechsel beim Thema Datenbanken herbeiführten und bis heute bestimmend sind (siehe Abb. 2.4). Dabei handelt es sich vereinfacht gesprochen um Systeme, die Daten in tabellarischer Form organisieren.

Abbildung 2.4 verdeutlicht die Marktdominanz des relationalen Paradigmas: so finden sich sieben bekannte relationale *Datenbankmanagement-Systeme* und ein *Information-Retrieval-System* (Platz 9; siehe unten) unter den ersten zehn Plätzen.

Rang Jul 2025	Rang Jun 2025	Rang Jul 2024	DBMS	Datenbankmodell
1.	1.	1.	Oracle	Relational, Multi-Model
2.	2.	2.	MySQL	Relational, Multi-Model
3.	3.	3.	Microsoft SQL Server	Relational, Multi-Model
4.	4.	4.	PostgreSQL	Relational, Multi-Model
5.	5.	5.	MongoDB	Document, Multi-Model
6.	6.	↑ 7.	Snowflake	Relational
7.	7.	↓ 6.	Redis	Key-value, Multi-Model
8.	8.	↑ 9.	IBM Db2	Relational, Multi-Model
9.	9.	↓ 8.	Elasticsearch	Multi-Model
10.	10.	10.	SQLite	Relational

Abbildung 2.4: DB-Engine-Ranking; https://db-engines.com/de/ranking; Letzter Abruf: 24.07.2025

Die direkte Interaktion mit relationalen *Datenbankmanagement-Systemen* erfolgt in der Regel mittels hochstrukturierter, künstlicher Anfragesprachen wie SQL[23].

[23] *Structured Query Language*; strukturierte Anfragesprache, ein ISO/IEC-Standard zur Interaktion mit relationalen DBMS.

Bei der Anfrage-Auswertung nutzen relationale *Datenbankmanagement-Systeme* die Boolesche Logik[24]. Infolgedessen können relationale *Datenbankmanagement-Systeme* die Anfrage genauer Ergebnisse in Bezug auf eine eindeutig definierte Anfrage garantieren – andernfalls würden sie als fehlerhaft gelten.

[24] Das heißt, Datenbankeinträge können entweder zu einer Anfrage passen („wahr“ sein) oder nicht.

Die Anfrageauswertung geschieht auf Grundlage Boolescher Operatoren und verschiedener Mengen-Operatoren. Als Ergebnis wird eine *ungeordnete Multimenge* zurückgegeben, die einzelne Elemente (Datenbankeinträge) mehrfach enthalten kann.

Information-Retrieval-Systeme

Aus historischer Sicht beschäftigt sich das *Information Retrieval* mit dem Auffinden von Informationen, die implizit in einer Menge von (zumeist) natürlichsprachigen Dokumenten[25] enthalten und für die vom Benutzer formulierte Suchanfrage relevant sind. Als solche können die erwarteten Ergebnisse ungenau sein, da die *Relevanz* der Dokumente nicht immer abschließend bestimmt werden kann, wie in diesem Abschnitt dargelegt wird.

Dies unterscheidet das Feld des *Information Retrieval* wesentlich von relationalen Datenbanken oder anderen Techniken zur Datenabfrage, die auf Daten mit klar definierter Struktur und Semantik basieren[26].

Information-Retrieval-Prozess Information Retrieval kann als die folgende Abbildung definiert werden[27]:

$$IR : (U, IN, Q, D) \rightarrow \mathcal{R} \tag{2.1}$$

Dabei steht U für einen Benutzer (*user*), der ein bestimmtes *Informationsbedürfnis* (IN, *information need*) hat, das vom *Information-Retrieval-System* (IRS) erfüllt werden soll.

Um mit dem *Information-Retrieval-System* zu interagieren, formuliert der Benutzer eine Anfrage (Q; *query*), die das *Informationsbedürfnis* vollständig oder teilweise widerspiegelt – wobei Letzteres der Regelfall ist und das sogenannte *Anfrageformulierungsproblem*[28] beschreibt.

Als Antwort auf Q, die anschließend mit einer Sammlung von Dokumenten D abgeglichen wird, welche die Datenbank des IRS bildet, erhält der Benutzer eine Menge von relevanten Dokumenten ($\mathcal{R}$; *relevant documents*) durch das sogenannte *Matching*.

Die Zusammensetzung der Ergebnismenge im Rahmen des *Matchings* wird durch eine *Ranking-Funktion* berechnet, welche die Einschätzung des Systems darüber darstellt, ob Dokumente in Bezug auf die Benutzeranfrage relevant sind – also die „am besten passenden" in der Datenbank enthaltenen Dokumente zur gestellten Anfrage.

Als Dokument sind hier nicht zwangsläufig Textdokumente oder ähnliches zu verstehen. Im Kontext des DATENATLAS handelt es sich bei den einzelnen Metadatensätzen um die verwalteten Dokumente.

ABBILDUNG 2.5 ILLUSTRIERT diesen Prozess, der ebenfalls analog auf *Datenbankmanagement-Systeme* übertragen werden kann. Aus der Abbildung wird deutlich, dass ein *Information-Retrieval-System* nicht zwangsläufig aus einer Komponente besteht. Aus technischer Sicht muss es nicht monolithisch aufgebaut sein. Das heißt, dass es in der Praxis häufig vorkommt, dass die Dokumenten- oder Metadatenspeicherung beispielsweise in einem relationalen *Datenbankmanagement-System* erfolgt, während die Suchfunktionali-

[25] Bruce W. Croft, Donald Metzler, und Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Pearson, Boston, Mass., international edition Auflage, 2009

Wenn nicht anders angegeben, verwenden wir den Begriff „Information" als Synonym für (relevantes) Wissen oder Fakten – im Gegensatz zur mathematischen Definition, wie sie in der Informationstheorie von Shannon (1948) eingeführt wurde.

[26] Ricardo Baeza-Yates und Berthier Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search*. Pearson Addison-Wesley [u.a.], Harlow, 2. Auflage, 2011

[27] Sándor Dominich. *The Modern Algebra of Information Retrieval*. Springer-11645 / Dig. Serial]. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2008

[28] Siehe Abschnitt 3.5.

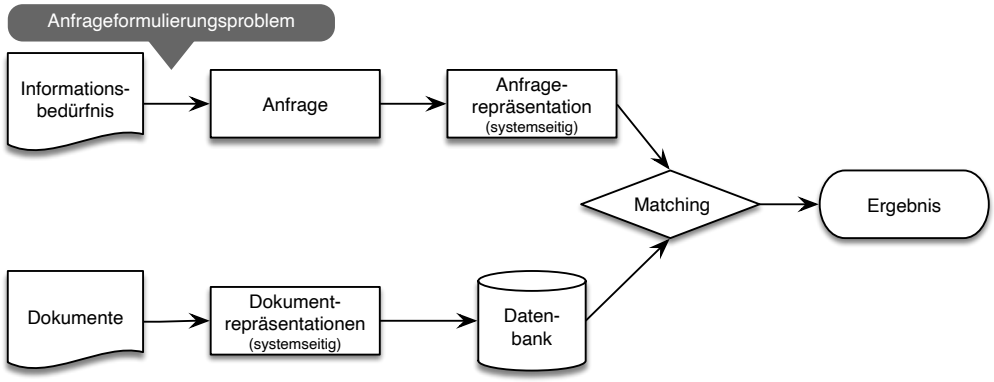


Abbildung 2.5: Ablauf des Information Retrievals

tät durch ein *Information-Retrieval-System* umgesetzt wird, welches über seinen *Suchindex* das *Matching* durchführt.

Beim *Information Retrieval* in einem relationalen *Datenbankmanagement-System* liegen die Dokumente in Form von Tabellen vor, wobei das Matching aufgrund der verwendeten Auswertelogik eine ungeordnete Multimenge[29] ergibt.

[29] Siehe Abschnitt 2.2.

Die direkte Interaktion mit *Information-Retrieval-Systemen* erfolgt in der Regel mittels natürlicher Sprache und ergänzenden, optionalen Steuerbefehlen[30]. Typischerweise werden natürlichsprachige Begriffe intern mittels des sogenannten *Stemmings* auf Stammformen reduziert oder es werden automatisch Synonyme ergänzt, um die Ergebnisqualität zu erhöhen. Aus Platzgründen werden diese Techniken hier jedoch ausgespart, da sie für das generelle Verständnis nicht nötig sind.

[30] Siehe Abschnitt 3.2.

Moderne Information-Retrieval-Systeme geben i.d.R. eine *Totalordnung der Dokumente* (eine Ergebnisliste) zurück, die nach *Relevanz* geordnet ist.

Typische Vertreter solcher Information-Retrieval-Systeme sind große Websuchmaschinen, wie Bing oder Google, lokal betreibbare Suchmaschinen wie Elasticsearch oder Solr, die beide die gleiche technische Basis teilen[31], beziehungsweise moderne Bibliothekskataloge, die in Abschnitt 3.2 als Beispiel für den *Stand der Technik* im Bereich der *Informationsrecherche* dienen.

[31] D.h. Apache Lucene, siehe Abschnitt 3.3.

3
Stand der Technik

In diesem Kapitel werden einige Aspekte des Stands der Technik mit Bezug zum Datenatlas und die Limitationen des vorliegenden Gutachtens vorgestellt.

Abschnitt 3.2 illustriert dabei die Entwicklung hin zum Stand der Technik im Bereich der *Informationsrecherche* am Beispiel der Literaturrecherche in Bibliotheken, die einem großen Kreis an Leserinnen und Lesern bekannt sein sollte und eine gewisse Parallelität zum Datenatlas aufweist.

Grundlage für die erfolgreiche *Informationsrecherche* stellen passend zu wählende Information-Retrieval-Modelle dar, die einen erheblichen Einfluss auf deren Funktionalität haben und deshalb im Folgeabschnitt kurz vorgestellt werden.

Menschzentriertes Information Retrieval – also zu implementierende Untertstützungsmöglichkeiten, um eine erfolgreiche Recherche durch Nutzende sicherzustellen – wird in Abschnitt 3.4 und 3.5 adressiert.

Die Sicherstellung der Datenqualität stellt neben der Wahl geeigneter Information-Retrieval-Modelle einen wichtigen Aspekt für die erfolgreiche *Informationsrecherche* dar und wird in Abschnitt 3.6 thematisiert.

Das Kapitel schließt mit einer Skizzierung der relevanten, regulatorischen Grundlagen (siehe Abschnitt 3.7), welche die Bundesverwaltung und damit den Datenatlas betreffen.

3.1 Limitationen des Gutachtens

Diese Diskussion zum *Stand der Technik* ist nicht abschließend. Aufgrund der beschränkten Recherchemöglichkeiten werden wesentliche Gebiete wie die verpflichtende Barrierefreiheit für Anwendungen der Bundesverwaltung[1], die IT-Sicherheit und weitere *explizit* ausgespart, auch wenn sich dadurch Risiken für den Betrieb und die Weiterentwicklung des Datenatlas ergeben.

Da der Autor keinen juristischen Hintergrund hat, können rechtliche Aspekte nur sehr oberflächlich diskutiert werden und sollten unter Hinzuziehung rechtswissenschaftlicher Expertise geprüft werden.

[1] BITV 2.0. *Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung)*, 2023

3.2 *Informationsrecherche*

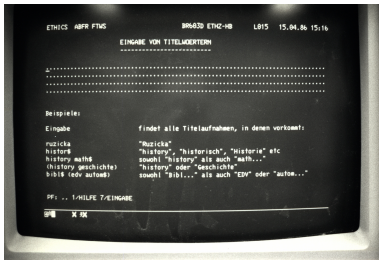


Abbildung 3.1: Titelwortabfrage im historischen digitalen Katalog der ETH Zürich (15.04.1986) CC BY NC SA ETH Zürich DOI: 10.3932/ethz-a-000014637

Das Feld der *Informationsrecherche* ist ohne Zweifel eine Domäne der Bibliotheken, die über Jahrhunderte ein großes Maß an Expertise aufbauen konnten. Beginnend mit einfachen Bestandsverzeichnissen über verschiedene Bibliothekskatalogformen, die bereits verschiedene *Informationssuchstrategien* (siehe Abschnitt 3.5) , wie die systematische oder die Suche nach Schlagworten, unterstützen, wurden die Recherchemöglichkeiten stetig weiter ausgebaut.

Spätestens Ende der 1980er-Jahre wurden Bibliotheken einem enormen Digitalisierungsdruck ausgesetzt[2], die zur Einführung der sogenannten OPACs[3] führten. Sich stetig wandelnde Nutzungsanforderungen (siehe Abschnitt 3.4) sowie der wissenschaftliche Fortschritt (siehe Abschnitte 3.3 und 3.5) trugen zudem zu einer kontinuierlichen Weiterentwicklung des angebotenen Funktionsumfangs bei.

Ein historisches Beispiel für einen OPAC der 1980er-Jahre findet sich in Abb. 3.1.

[2] Siehe Zellhöfer (2022).

[3] *Online Public Access Catalogue*; ein Online-Katalog zur *Informationsrecherche*.

Diese kontinuierliche Fortentwicklung erfolgt seit Jahrzehnten global, da Bibliotheken z.B. aufgrund der Fernleihe Medien international verteilen und deshalb (Meta-)Daten *interoperabel* und *standardisiert* erfassen können müssen. Gleichzeitig sind diese Systeme dafür ausgelegt, viele Millionen an Datensätzen und/oder Dokumenten zu verwalten und den Nutzenden schnellstmöglich bereitzustellen.

Die zu diesem Zweck erstellten Software-Systeme sind zum großen Teil, zumindest aber im Bereich ihrer Schnittstellen, ebenfalls seit Jahrzehnten als *Open Source*[4] verfügbar und werden aufgrund der wachsenden Datenmengen und der dynamischen Nutzungsanforderungen kontinuierlich weiterentwickelt (siehe Abschnitt 5.2).

[4] *Open Source*; Software mit verfügbarem Quellcode und permissiver Lizenz, siehe Abschnitt 3.7.

Bereits seit Mitte der 1990er-Jahre lässt sich der Betrieb in Bibliotheken nicht mehr ohne Software-seitige Unterstützung realisieren und wird mittlerweile als entsprechend basal betrachtet, wie das folgende Zitat gut belegt:

> Because the library has become software, it is no longer viable for our services to exist separately from our software. Our best opportunities for intervention, for reference, and for instruction are within our software. Our users succeed every day in information seeking online, as they shop, [...]. In none of these cases do our users expect that services vital to the endeavor exist outside of the software in front of them. Rather, the software is the service.[5]

[5] Eine Übersetzung des Zitats von Hanson (2015) findet sich in Anhang A.1.

Recherchemöglichkeiten an Bibliotheken

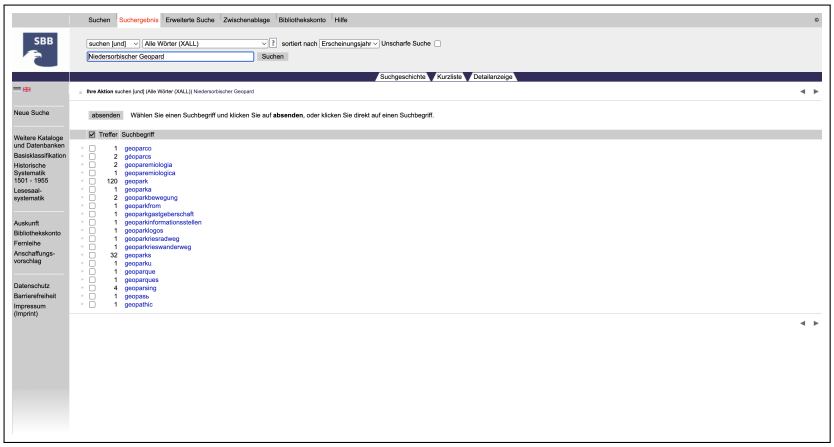
Ein klassischer OPAC in Form des mittlerweile „StaBiKat classic“ genannten Systems ist in Abb. 3.2 zu sehen.

Der im Jahr 2001 an der STAATSBIBLIOTHEK ZU BERLIN eingeführte OPAC hat seine technischen Wurzeln im Jahr 1998[6].

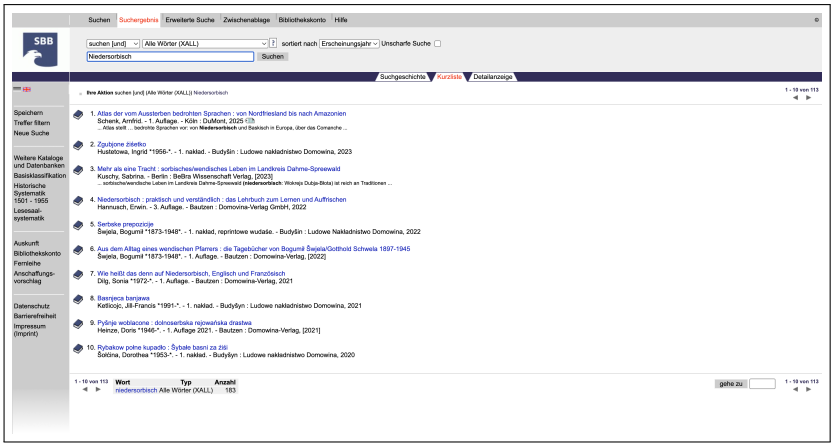
DIE LESENDEN seien darauf hingewiesen, dass es in der folgenden Argumentation *ausschließlich* um die Funktionen der Anwendungen und nicht deren teils altmodisch wirkende grafische Gestaltung der Nutzerschnittstelle geht. Das grafische Erscheinungsbild ist deshalb zu vernachlässigen, da es dem teils hohen Alter der vorgestellten Systeme geschuldet ist.

[6] Das zugrundeliegende OPAC-System, OCLC PICA opc4, ist seit 1998 in Betrieb, wie dem Quellcode der Website entnommen werden kann.

Staatsbibliothek zu Berlin. *SBB StaBiKat - Online-Hilfe*, 2001. https://t1p.de/k2xhc. Letzter Abruf: 22.07.2025



(a) Trefferliste zum Suchbegriff „Niedersorbischer Geopard“ *[sic!]*



(b) Trefferliste zum Suchbegriff „Niedersorbisch“

Abbildung 3.2: Typische Funktionen von OPAC-Systemen; https://lbssbb.gbv.de; Letzter Abruf: 01.08.2025

Die Recherche im OPAC erfolgt (mindestens) anhand eines Suchbegriffs. Als Suchbegriff wurde in diesem Fall „Niedersorbischer Geopard“ *[sic!]* eingegeben. In diesem Beispiel ist deutlich zu erkennen, dass die Suchanfrage einen Rechtschreibfehler sowie ein unwahrscheinliches Verbreitungsgebiet von Geparden beinhaltet. Erwartungsgemäß präsentiert der OPAC eine leere Ergebnisliste (Abb. 3.2 (a)), die jedoch sinnvollerweise mit verwandten Suchbe-

griffen und den dort zu erwartenden Treffern versehen wurden, um eine erfolgreiche *Informationsrecherche* zu begünstigen.

Das Ergebnis einer erfolgreichen Suche mit dem Suchbegriff „Niedersorbisch“ ist in Abb. 3.2 (b) dargestellt. Wie im oberen Bereich der Webseite zu sehen ist, werden weitere Funktionen, wie die Sortierung nach verschiedenen Kriterien, u.a. dem Erscheinungsjahr oder der *Relevanz*, angeboten. Des weiteren werden die folgenden Möglichkeiten zur Formulierung einer Suchanfrage bereitgestellt, die sich mittlerweile als *Mindestmaß der Anfrage-Formulierungsmöglichkeiten*, etabliert haben:

Mindestmaß der Anfrage-Formulierungsmöglichkeiten

1. Suchterme (Schlagwort-basierte Suche)
2. Boolesche Operatoren (`UND, ODER, NICHT`)
3. Wildcard-Operatoren
4. Proximity-Operatoren
5. Annähernde/unscharfe bzw. Fuzzy Suche[7]
6. In-/Exklusion von Begriffen
7. Definition von Suchbegriffen für einzelne Metadaten-Felder
8. Phrasensuche

[7] Eine Suchmöglichkeit, welche es ermöglicht, nicht nur nach exakten Begriffen sondern, anhand von Ähnlichkeit zu diesen zu recherchieren.

Für die konkrete Erläuterung dieser Begriffe sei exemplarisch auf die Dokumentation der Staatsbibliothek zu Berlin (2001) verwiesen.

Da sich die Nutzungserwartung bei der *Informationsrecherche* u.a. durch die Verbreitung von Websuchmaschinen weiterentwickelt hat, sind oder wurden solche OPAC-Systeme seit den frühen 2010er-Jahren größtenteils durch sogenannte *Discovery-Systeme* verdrängt, auch wenn OPAC-Systeme spezifische Stärken z.B. bei der sogenannten *Known-Item-Search*[8] haben.

[8] *Known-Item-Search*; die zielgerichtete Suche nach bereits bekannten Objekten, siehe Abschnitt 3.5.

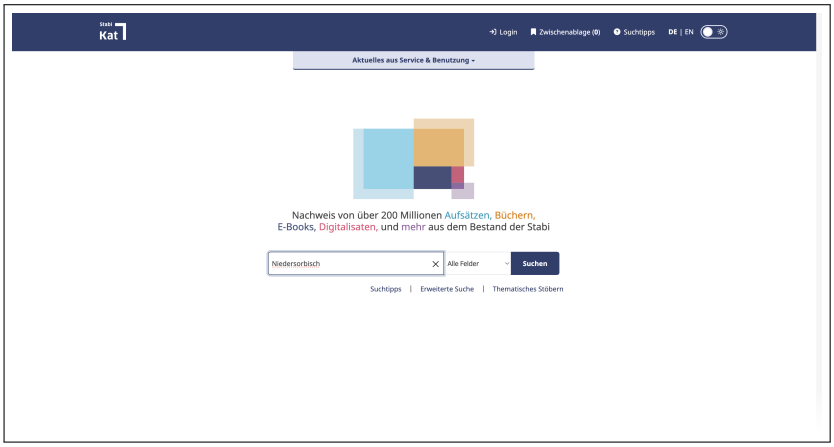
Zum direkten Vergleich ist das *Discovery-System* „StabiKat“ in Abb. 3.3 abgebildet. Das System wurde 2023 eingeführt und hat seine technischen Wurzeln im Jahr 2010[9].

Es bietet einen den modernen Nutzungserwartungen entsprechenden Suchterm-gestützten Sucheinstieg (Abb. 3.3 (a)) und eine nach *Relevanz* geordnete Trefferliste (Abb. 3.3 (b)), welche den Zugang zu den Inhalten durch die Einbindung sogenannter *Facetten*[10] erleichtert. *Facetten* unterstützen Nutzende indem sie den Inhalt der Ergebnisliste, ähnlich wie Filter, weiter strukturieren und zu erwartende Treffer in den facettierten Kategorien, z.B. durch Einordnung des Medientyps, direkt visualisieren. Damit unterstützt das System auch grundlegende *Browsing-Funktionen*[11].

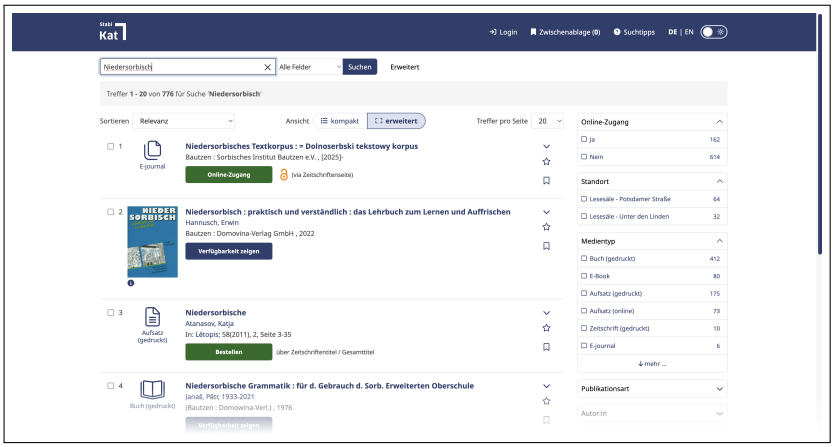
[9] Es basiert auf VuFind, welches seitdem kontinuierlich weiterentwickelt wird; `https://vufind.org/vufind/`; Letzter Abruf: 21.07.2025.

[10] Siehe Abschnitt 3.5.

[11] Siehe Abschnitt 3.5.



(a) Sucheinstieg mit dem Suchbegriff „Niedersorbisch“



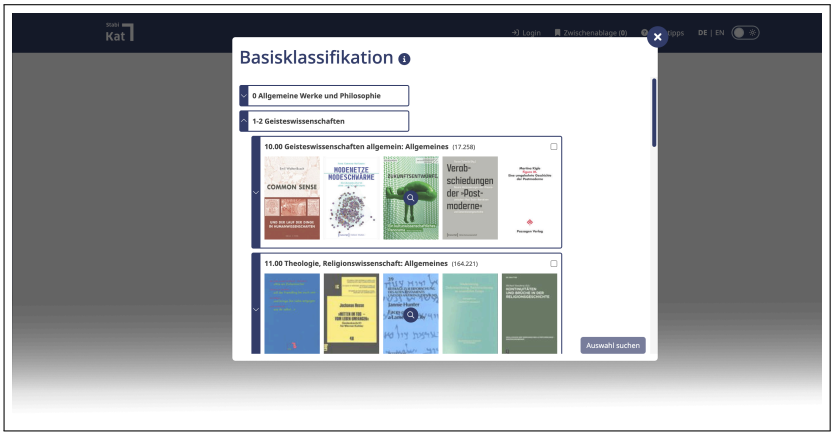
(b) Trefferliste zum Suchbegriff „Niedersorbisch“

Abbildung 3.3: Typische Funktionen von *Discovery-Systemen*; `https://stabikat.de`; Letzter Abruf: 01.08.2025

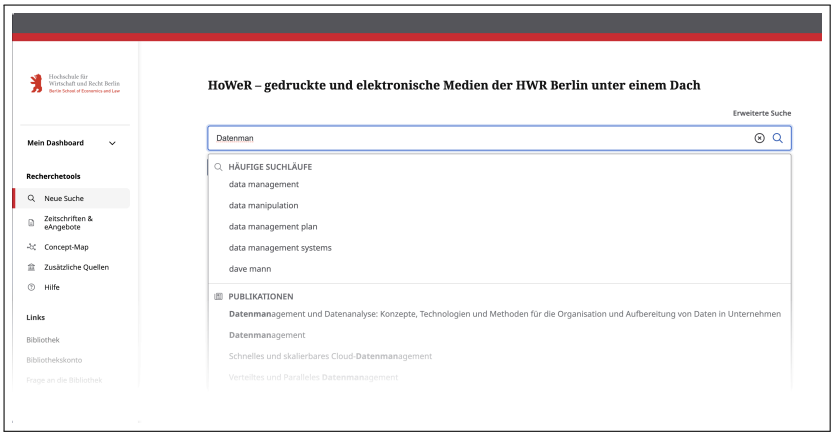
Weitere typische Funktionen auf dem *Stand der Technik* von *Discovery-Systemen*, wie der Browsing-basierte Einstieg oder die Autovervollständigung von Suchbegriffen zur Ermöglichung *explorativer* Suchansätze[12], sind in Abb. 3.4 dargestellt.

Discovery-Systeme unterstützen i.d.R. zusätzlich die typischen Anfrage-Formulierungsmöglichkeiten eines OPACs.

[12] Siehe Abschnitt 3.5.



(a) Browsing-Einstieg mittels der Basisklassifikation von Publikationen



(b) Autovervollständigung von Suchbegriffen anhand von häufigen Suchverläufen bzw. vorhandenen Publikationen

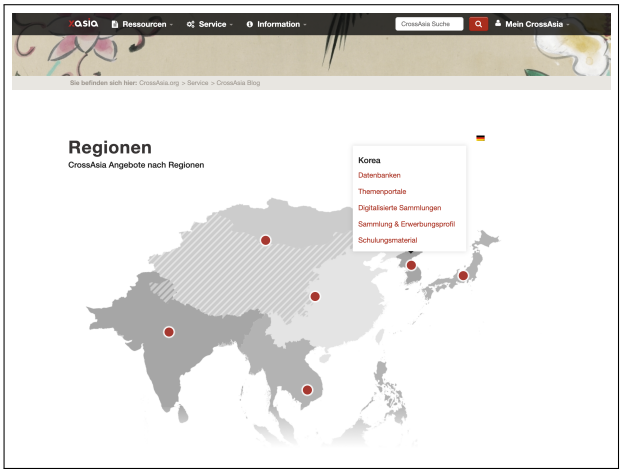
Abbildung 3.4: Weitere typische Funktionen von *Discovery-Systemen*; (a) `https://stabikat.de`; Letzter Abruf: 01.08.2025, (b) `https://hower.hwr-berlin.de`; Letzter Abruf: 01.08.2025

Repository-Systeme bilden die dritte Klasse der Systeme zur *Informationsrecherche*. Diese Systeme werden auch als *Digital-Asset-Management-Systeme (DAM)* bezeichnet. Sie verwalten Sammlungen an digitalen Assets sowie deren Metadaten und stellen diese Nutzenden i.d.R. webbasiert bereit. Hierbei umfasst der Kreis an Nutzenden sowohl Endanwender, die recherchieren, als auch Datenerfasser, welche Datensätze bzw. Dokumente erstellen oder modifizieren. Das heißt, dass *Repository-Systeme* üblicherweise mindestens die Use Cases von OPACs und *Discovery-Systemen* unterstützen und diese teilweise stark erweitern[13].

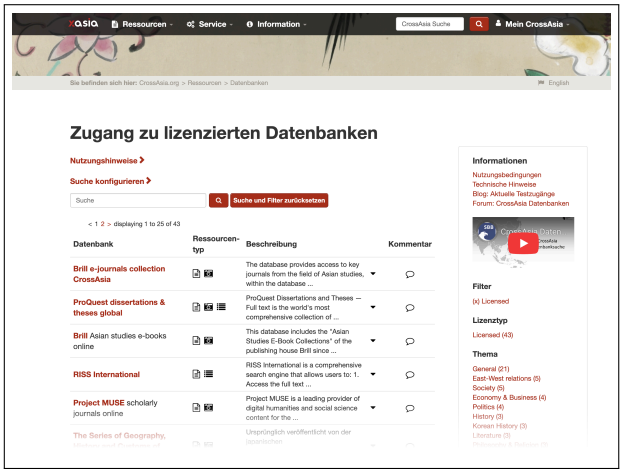
[13] Siehe Abschnitt 5.2.

Die zugrundeliegenden Kataloge eines OPACs bzw. *Discovery-Systems* werden hingegen mit separaten Werkzeugen editiert bzw. über externe Schnittstellen gespeist.

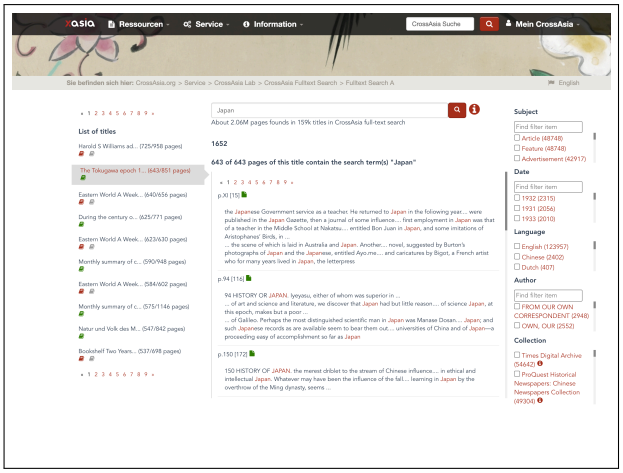
Abbildung 3.5 stellt ein solches *Repository-System* mit einer kleinen Auswahl seiner Funktionsangebote dar. Ein typisches Feature, das als *Stand der Technik* gelten kann, ist die Hervorhebung von Suchtreffern im Rahmen der Volltextsuche (siehe Abb. 3.5 (c)).



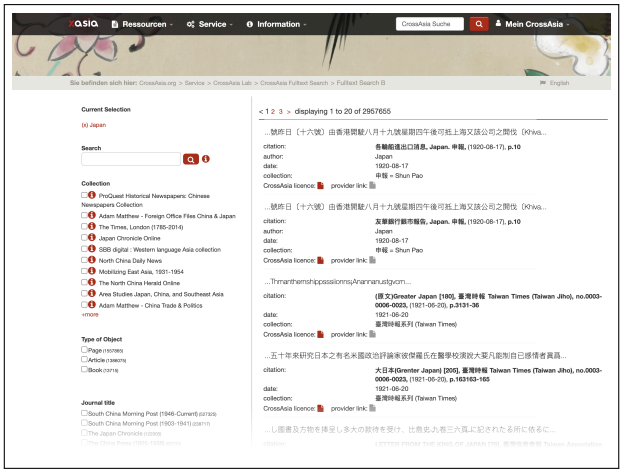
(a) Browsing-Einstieg mittels geographischer Regionen



(b) Einstieg zu Datenbank-Recherche inkl. Suche, Filtern und Facetten



(c) Volltext-Suche inkl. visueller Hervorhebung von Treffern (rot)



(d) Explorative Volltext-Suche

Abbildung 3.5: Typische Funktionen von *Repository-Systemen* am Beispiel von CrossAsia; `https://crossasia.org`; Letzter Abruf: 01.08.2025

Technisch gesehen handelt es sich bei den oben vorgestellten Systemen zur *Informationsrecherche* um *Information-Retrieval-Systeme* (siehe Abschnitt 2.2), die auf unterschiedlichen theoretischen Modellen basieren, welche wiederum Einfluss auf den bereitgestellten Funktionsumfang im Rahmen der *Informationsrecherche* haben. Diese als *Information-Retrieval-Modelle* bezeichneten Modelle werden im folgenden Abschnitt 3.3 kurz vorgestellt.

Der Datenatlas wird in Abschnitt 5.1 technisch eingeordnet.

3.3 *Information-Retrieval-Modelle*

In Abschnitt 2.2 wurde der generelle Ablauf des Information-Retrieval-Prozesses eingeführt, welcher in Abb. 3.6 abgebildet ist.

Vereinfacht gesprochen bestimmt das *Information-Retrieval-Modell* darüber, wie die Dokumentenrepräsentation im System ausgestaltet ist und auf welche Art und Weise das *Matching* mit der bereitgetellten Anfrage realisiert wird, um ein Ergebnis zu erhalten.

Das IR-Modell entscheidet aufgrund seiner mathematischen Grundlage darüber, welche Eigenschaften die Ergebnismenge hat,

z.B. ob diese geordnet, beispielsweise in Form einer nach Relevanz geordneten Liste, ist oder nicht.

Das gewählte IR-Modell beeinflusst die Leistungsfähigkeit des IR-Prozesses maßgeblich, wie unten ausgeführt wird.

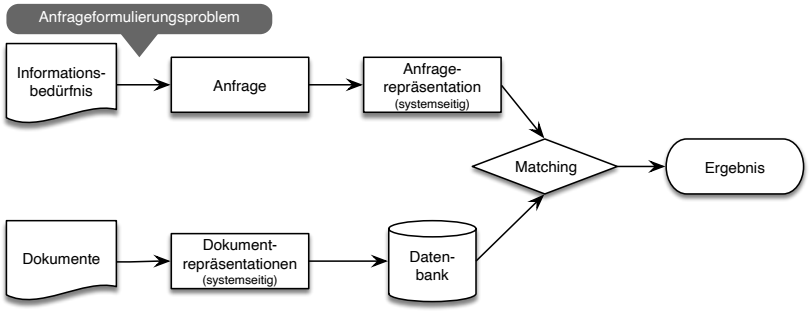


Abbildung 3.6: Ablauf des Information Retrievals

Die Geschichte des *Information Retrievals* lässt sich bis in die Mitte der 1950er-Jahre zurückverfolgen. In diesem Zeitraum wurden eine Vielzahl an *Information-Retrieval-Modellen* vorgeschlagen, die hier nicht alle diskutiert werden können.

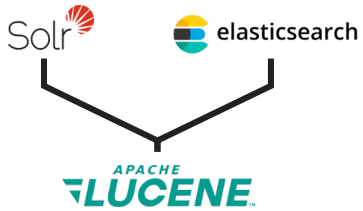


Abbildung 3.7: Beziehung zwischen Apache Lucene und verbreiteten Suchmaschinen

Folglich ist eine Begrenzung der Betrachtungstiefe im Rahmen dieses Gutachtens sinnvoll.

Die nachfolgende Darstellung beschränkt sich auf diejenigen IR-Modelle, welche durch Apache Lucene[14] unterstützt werden bzw. dort standardmäßig voreingestellt sind.

Apache Lucene ist eine *Open Source*-Bibliothek für die Volltextindizierung und -suche – in anderen Worten ein *Information-Retrieval-System*, welches die Basis für weitverbreitete Suchmaschinen wie Elasticsearch oder Solr bildet (siehe Abb. 3.7).

Apache Lucene kombiniert beim Matching das Boolesche Modell mit dem Vektorraum-Modell[15]. Beide IR-Modelle werden in diesem Abschnitt detailliert vorgestellt.

Die Diskussion dieser IR-Modelle orientiert sich, wenn nicht anders angegeben, an Zellhöfer (2015). Dort finden sich auch die mathematisch korrekten und vollständigen Beschreibungen der Modelle, welche hier nur sehr oberflächlich diskutiert werden, um den Lesefluss nicht zu stören.

Weitere Aspekte des Information Retrievals wie die Aufbereitung natürlicher Sprache, z.B. mittels Stemming, oder die Indizierung von Dokumenten, werden aus Verständlichkeits- und Platzgründen im Rahmen des Gutachtens nicht erläutert.

Für weitere Details zum Forschungsfeld *Information Retrieval* und für die Beschreibung der probabilistischen IR-Modelle, welche z.B. in Form der *Okapi BM25*-Rankingfunktion ebenfalls durch Apache Lucene unterstützt werden, sei auf Robertson und Spärck Jones (1976); Robertson (1977); van Rijsbergen (1979, 1986); Dominich (2008); Croft et al. (2009); Baeza-Yates und Ribeiro-Neto (2011) oder andere einschlägige Lehrbücher wie Henrich (2008) verwiesen.

[14] https://lucene.apache.org; Letzter Abruf: 21.07.2025

[15] https://lucene.apache.org/core/10_2_2/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html; Letzter Abruf: 27.07.2025

Aus Platzgründen wird auf die Präsentation von *Okapi BM25* verzichtet, da die Retrieval-Ergebnisse mit denen des Vektorraum-Modells weitestgehend vergleichbar, wenn auch mathematisch anders begründet, sind.

Information-Retrieval-Modelle auf Basis der Mengenlehre

Boolesches Information Retrieval Das Boolesche Retrieval-Modell (BRM) ist das älteste Retrieval-Modell und mathematisch relativ einfach aufgebaut. Es basiert auf der Mengenlehre und der Booleschen Algebra.

Die Kernidee des BRM besteht darin, Dokumentrepräsentationen als eine Menge von Elementen des Indexvokabulars[16] zu verstehen, wobei deren Häufigkeit im Original-Dokument keine Rolle spielen. Das Indexvokabular beinhaltet alle Wörter, die das *Information-Retrieval-System* zur Recherche anbietet.

Die *Anfrage* besteht ebenfalls aus Elementen des Indexvokabulars, die mit Booleschen Junktoren (`UND, ODER, NICHT`) verbunden werden können. Diese Anfrage dient als eine Art „Filter", welcher im Rahmen des Matchings die (ungeordnete) Ergebnismenge ergibt. Grundlage dafür bildet ein sogenanntes *Exact Matching*, welches für jede Dokumentenrepräsentation bestimmt, ob ein Dokument zur Ergebnismenge gehört oder nicht.

Durch die Verwendung der Junktoren bzw. *nicht exakt übereinstimmender* Schlagworte zwischen der Anfrage und den Dokumentrepräsentationen ist es möglich, dass eine leere Ergebnismenge generiert wird, da das BRM keine teilweise relevanten Dokumente zurückliefert, wie es beim *Best Matching* (s.u.) der Fall wäre.

Ebenso ist es möglich, dass zu kleine oder zu große Ergebnismengen erzeugt werden, wenn beispielsweise Nutzende die Junktoren mathematisch nicht korrekt anwenden[17].

Das relationale Datenbank-Modell[18] teilt das gleiche mathematische Fundament mit dem BRM – mitsamt seiner benannten Schwächen.

Tatsächlich wird *Datenbank-Retrieval* von einigen Autoren, z.B. van Rijsbergen (1986) oder Nottelmann und Fuhr (2003), als Spezialisierung des *Information Retrievals* gesehen. Das Gutachten schließt sich dieser Interpretation an.

Erweitertes Boolesches Retrieval Um die Strenge des dem BRM zugrundeliegenden Exact-Match-Prinzip zu überwinden, wurden verschiedene Erweiterungen für dieses Modell diskutiert. Eine Reihe von Autoren hat Erweiterte Boolesche Retrieval-Modelle vorgeschlagen, z. B. Waller und Kraft (1979); Salton et al. (1983); Yager (1988); Fox et al. (1992) oder Lee (1994).

Diese Modelle haben gemeinsam, dass sie versuchen, eine Rangordnung der gefundenen Dokumente bereitzustellen, um auch nur teilweise relevante Dokumente mit in die Ergebnismenge aufneh-

[16] *Indexvokabular*; Das Indexvokabular beinhaltet alle Terme, welche das *Information-Retrieval-System* verarbeiten kann. Es sollte zumindest die wesentlichen, semantisch relevanten Begriffe des gesamten Dokumentenkorpus beinhalten. Aus Effizienzgründen ist das Indexvokabular immer begrenzt, u.a. um Anfragen möglichst schnell bearbeiten zu können.

[17] Dagobert Soergel. *Organizing Information: Principles of Data Base and Retrieval Systems*. Academic Press Professional, Inc., San Diego, CA, USA, 1985

[18] Siehe Abschnitt 2.2.

Im Folgenden wird *Information Retrieval* als Generalisierung des Datenbank-Retrievals betrachtet.

men zu können. Außerdem bieten sie i.d.R. Möglichkeiten zur subjektiven Gewichtung einzelner Anfrageteile sowie „abgeschwächte“ Versionen der Booleschen Operatoren.

Grob gesagt besteht die Kernidee der „abgeschwächten“ Booleschen Operatoren darin, diese Verknüpfungen mit einem Parameter auszustatten, der ihre logische Charakteristik beeinflusst. Das heißt, dass ein Parameter den Grad der Konjunktivität oder Disjunktivität eines Operators steuert, der zwei Anfrageterme verbindet.

Oft erlauben Erweiterte Boolesche Ansätze auch eine Gewichtung der einzelnen Anfragebegriffe, um deren Wichtigkeit für Nutzende auszudrücken und so auch Dokumente in die Ergebnismenge aufzunehmen, die nur einigen Teilen der Anfrage genügen und somit die oben genannten Nachteile des BRM überwinden.

Vergleichbare Gewichtungsansätze finden sich auch im Bereich der Datenbanken, z. B. bei Fagin und Wimmers (2000).

Fuzzy-Logik-basiertes Retrieval Das Fuzzy-Retrieval-Modell (FRM) basiert auf dem mathematischen Konzept der *Fuzzy-Mengen*[19] und der damit assoziierten Logik[20].

Entgegen klassischen Mengen erlauben Fuzzy-Mengen die Zugehörigkeit von Elementen mittels Zugehörigkeitsfunktionen auszudrücken, welche die graduelle Zugehörigkeit eines Elements zu einer Fuzzy-Menge angibt. Daher können Elemente u. U. auch mehreren Fuzzy-Mengen angehören.

Mittels Fuzzy-Mengen und Fuzzy-Logik können Konzepte wie Ungewissheit oder Relevanz mathematisch abgebildet werden.

Zur Operation auf Fuzzy-Mengen bzw. der Formulierung von Ausdrücken der Fuzzy-Logik sind diverse Äquivalente zu den Mengenoperatoren/Booleschen Operatoren `AND`, `OR` und `NOT` verfügbar[21]. Anfragen im FRM können folglich – analog zum BRM – mittels dieser Operatoren formuliert werden.

[19] *Fuzzy*; unscharf.

Lotfi A. Zadeh. Fuzzy Sets. *Information and Control*, 3(8):338–353, January 1965. DOI: 10.1016/S0019-9958(65)90241-X

[20] A. Lotfi Zadeh. Fuzzy Logic. *IEEE Computer*, 21(4):83–93, 1988

[21] Zadeh (1965, 1988); Zimmermann (1996)

Im Gegensatz zum BRM und dessen Erweiterungen ermöglicht es das FRM, die graduelle Relevanz von Dokumenten bezüglich einzelner Elemente des Indexvokabulars zu modellieren.

So wird es beispielsweise möglich, den Begriff „Metadaten“ dem Indexterm „Daten“ zuzuordnen. Im Ergebnis würden so bei der Anfrage „Daten“ auch Dokumente, die nur den Term „Metadaten“ enthalten zurückgeliefert – wenngleich auch weiter unten im nach Relevanz geordneten Ranking des FRM.

Existieren keine Dokumente in der Datenbank, die den Begriff „Daten“ enthalten, so würde ein auf dem FRM basierendes *Information-Retrieval-System* die Dokumente, die „Metadaten“ beinhalten, zurückliefern.

Diese Art der Ergebnisgenerierung wird als *Best Matching* bezeichnet.

Das Vektorraum-Modell

Das Vektorraum-Modell (VRM) erhält seinen Namen dadurch, dass sowohl die Dokumentenrepräsentationen als auch die Anfrage als Vektoren eines Vektorraums modelliert werden. Die Dimension des Vektorraums ergibt sich aus der Anzahl der Terme im Indexvokabular.

Die einzelnen Dimensionen der Vektoren der Anfrage und der Dokumentrepräsentationen geben die Gewichtung (beziehungsweise deren „Wichtigkeit“) bezüglich der Terme des Indexvokabulars an und berücksichtigen dabei im Gegensatz zu den zuvor diskutierten Ansätzen auch Worthäufigkeiten und Dokument- bzw. Anfragelängen[22].

Zumeist wird zur Berechnung der Termgewichte die $tf * idf$-Formel[23] genutzt, welche auch in Apache Lucene Verwendung findet[24].

Das Matching geschieht auf Grundlage einer algebraischen Ranking-Funktion, wie der Kosinus-Ähnlichkeit, $tf * idf$-kompatiblen Ähnlichkeitsmaßen oder *Okapi BM25*.

Alle diese Ranking-Funktionen eint, dass sie für jede Kombination aus Anfrage und Dokumentrepräsentationen einen Wert zwischen 0 und 1 berechnen, wobei 1 für die vollständige *Relevanz* eines Dokuments bezüglich der Anfrage steht.

Auf dem VRM basierende *Information-Retrieval-Systeme*, wie z.B. Apache Lucene, führen deshalb stets ein *Best Matching* auf Grundlage aller Dokumente mit der Anfrage durch. Die nach den Ergebnissen der Ranking-Funktion geordnete Ergebnismenge bildet eine nach Relevanz sortierte Liste.

Die Relevanzsortierung ist i.d.R. gut für Nutzende nachvollziehbar, da sie z.B. auf Grundlage der einzelnen Termgewichte erklärt werden kann[25].

Das VRM bildet die Basis vieler moderner *Information-Retrieval-Systeme*, weil es Dokumente in eine mathematische Form bringt, die Berechnungen und Vergleiche erlaubt.

Aufgrund seiner sauberen mathematischen Grundlage kann das VRM in jedem Szenario verwendet werden, das sich mit Vektoren darstellen lässt, wie auch Schmitt (2006) anmerkt.

Die Nutzung des VRM begünstigt die spätere Nachnutzung der Datenmodellierung im Bereich der *Künstlichen Intelligenz*[26], welche ebenfalls zu großen Teilen auf Vektorrepräsentationen basiert.

Abschliessend kann festgehalten werden, dass die Verwendung des VRM das Mindestmaß des zu erreichenden Stands der Technik im Bereich des Information Retrievals darstellt.

[22] Somit kann kompensiert werden, dass in langen Dokumenten auch die Wahrscheinlichkeit häufiger auftretender Indexterme ansteigt. Ansonsten würden lange Dokumente stets die Ergebnis-Rankings dominieren.

[23] Gerard Salton und Christopher Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988. http://dx.doi.org/10.1016/0306-4573(88)90021-0

[24] https://lucene.apache.org/core/10_2_2/core/org/apache/lucene/search/similarities/ClassicSimilarity.html; Letzter Abruf: 18.08.2025

[25] Apache Lucene bietet hier beispielsweise entsprechende Funktionen.

[26] Siehe Abschnitt 5.6.

3.4 Grundsätze der menschzentrierten Gestaltung von IT-Systemen

Historisch gesehen rückten die Bedürfnisse der Nutzenden mit der zunehmenden Leistungsfähigkeit von IT-Systemen bei deren Gestaltung immer mehr in den Fokus der Aufmerksamkeit, da ressourcenbedingte Limitationen Interaktionsmöglichkeiten zunehmend weniger beschränken. Das weite Feld der *Mensch-System-Interaktion* wird aktiv seit den frühen 1980er-Jahren in verschiedenen Disziplinen, u.a. der Informatik, der Psychologie oder der Informationswissenschaften, beforscht und kann demnach im Folgenden nur kurz umrissen werden.

Die Einhaltung der vorgestellten Prinzipien wird für die Bundesverwaltung in den einschlägigen Rechtsnormen gefordert, welche in Abschnitt 3.7 weiter beschrieben werden.

Die folgende Definition der *menschzentrierten Qualität* ist Geis und Tesch (2019) entnommen und lehnt sich an die Normenfamilie DIN EN ISO 9241 zum Thema menschzentrierte Gestaltung an:

> Das Ausmaß, in dem ein interaktives System Anforderungen bezüglich
>
> - Gebrauchstauglichkeit (Usability),
> - Benutzererlebnis (User Experience),
> - Barrierefreiheit (Accessibility) und
> - Vermeidung von Schäden durch die Benutzung (Avoidance of harm from use)
>
> erfüllt.

Die Diskussion der beiden letzten Anforderungen liegt, wie in Abschnitt 3.1 erwähnt, außerhalb des Erfassungsbereichs dieses Gutachtens und stellt keine Abwertung dieser Aspekte dar.

Zum besseren Verständnis werden im Folgenden zentrale Grundbegriffe kurz umrissen und kontextualisiert.

Usability

Der Begriff *Usability* bezeichnet das „Ausmaß, in dem ein *interaktives System* von bestimmten *Benutzern* benutzt werden kann, um in einem bestimmten *Nutzungskontext* bestimmte *Ziele effektiv, effizient* und *zufriedenstellend* zu erreichen."[27]

User Experience

Unter *User Experience* versteht man die „Wahrnehmungen und Reaktionen eines Benutzers, die sich aus der Nutzung und/oder der erwarteten Nutzung eines interaktiven Systems ergeben".[28] *User Experience* lässt sich von *Usability*, wie in Abb. 3.8 dargestellt, abgrenzen.

[27] Thomas Geis und Guido Tesch. *Basiswissen Usability und User Experience: Aus- und Weiterbildung zum UXQB® Certified Professional for Usability and User Experience (CPUX) - Foundation Level (CPUX-F)*. dpunkt.verlag, Heidelberg, 1. Auflage, 2019. ISBN 978-3-86490-599-5

[28] Ebenda.

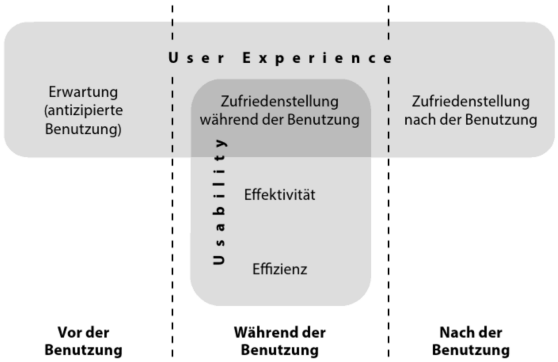


Abbildung 3.8: Unterschied zwischen *Usability* und *User Experience* (Geis und Tesch (2019), Abb. 2-2)

Menschzentrierter Gestaltungsprozess

Zur Erreichung der Ziele der *menschzentrierten Gestaltung* wird ein iterativ-inkrementelles (siehe Abb. 3.9) Vorgehensmodell gewählt, welches von Abb. 3.10 illustriert wird. Aus der Abbildung wird deutlich, dass das Verständnis des konkreten Nutzungskontexts essentiell ist, um konkrete Nutzungsanforderungen ableiten zu können, aus denen nutzbare Lösungsansätze entwickelt werden können. Da es sich dabei um Hypothesen handelt, müssen diese stetig evaluiert werden (s.u.).

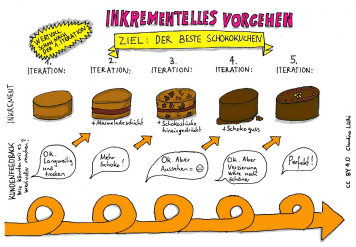


Abbildung 3.9: Iterativ-inkrementelles Vorgehen; CC BY Claudius Lüthi

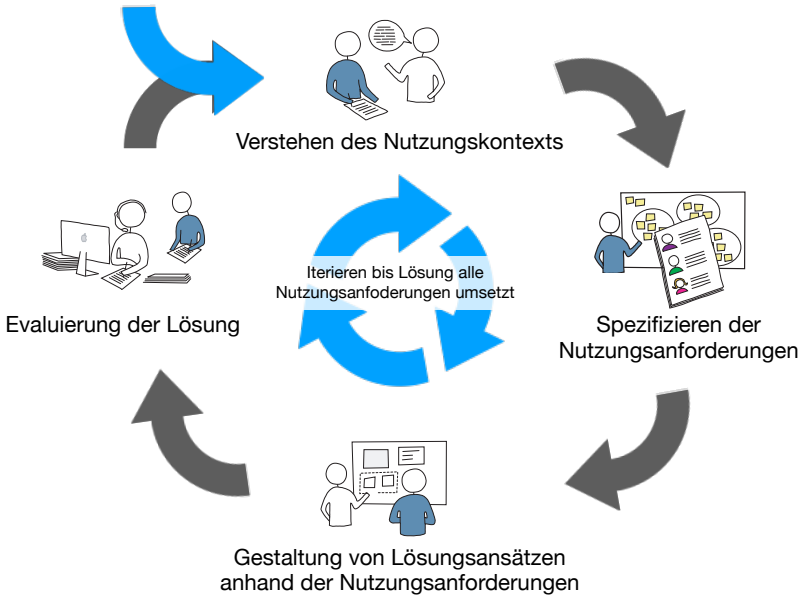


Abbildung 3.10: Menschzentrierter Entwicklungsprozess nach DIN EN ISO 9241-210 (2020)

Der menschzentrierte Gestaltungsprozess wird in DIN EN ISO 9241-210 (2020) beschrieben.

Da das Verständnis des Nutzungskontexts und die Spezifikation von Nutzungsanforderungen mit den Bedarfen der Zielgruppe übereinstimmen müssen, diese im Entwicklungsprozess jedoch u.U. nicht dauerhaft zur Verfügung steht, werden üblicherweise Methoden, wie die der *Personas*, in den menschzentrierten Gestaltungsprozess integriert, welche dazu dienen, die Nutzenden bestmöglich zu repräsentieren.

Personas

Mit den sogenannten *Personas*[29] steht eine bewährte, niedrigschwellig einsetzbare Methode zur Modellierung von Nutzenden und ihrer Nutzungskontexte zur Verfügung. Während es auf den ersten Blick einleuchtend erscheinen mag, ein System für eine diverse Nutzungsgruppe derart zu gestalten, dass möglichst viele und vielfältige Funktionalitäten implementiert werden, ist es vielmehr richtig, bei der menschzentrierten Gestaltung auf die spezifischen Anforderungen dieser zu achten[30].

[29] Alan Cooper. *The Inmates Are Running the Asylum*. Macmillan Publishing Co., Inc., Indianapolis, IN, USA, 1999

[30] Alan Cooper, Robert Reimann, und Dave Cronin. *About Face 3: The Essentials of Interaction Design*. Wiley, Indianapolis, Ind., 2007

Hierzu ist es notwendig, sich Klarheit über die Ausgestaltung der Zielgruppe zu verschaffen. Eine Möglichkeit, archetypische Nutzende in den Entwicklungsprozess zu integrieren, stellen dabei *Personas* dar. Diese basieren im besten Fall auf den Charakteristika einer Gruppe real-existierender Nutzender und fassen diese in einer leicht verständlichen Form zusammen. Während der Entwicklung dienen Personas dazu, Design-Entscheidungen mit den spezifischen Bedarfen der jeweiligen Persona abzugleichen und zu bewerten. In Abbildung 3.11 ist eine beispielhafte Persona zu sehen, welche *ausschließlich* den Autor repräsentiert.

Je nach Einsatzzweck werden *Personas* mit weiteren Eigenschaften versehen, welche es dem Entwicklungsteam ermöglichen sollen, sich möglichst empathisch in deren Rolle und Gefühlswelt während der *Mensch-System-Interaktion* hineinversetzen zu können.

Um Dienste menschzentriert gestalten oder entsprechend bewerten zu können, werden *Personas* häufig – wie auch in diesem Gutachten – mit der User-Journey-Methode[31] verbunden, um deren emotionale Erlebnisse während der *Mensch-System-Interaktion* mit einem Dienst, wie Frustration oder Erleichterung, und damit ihre individuelle *User Experience*, zu erfassen.

Die Methode ist seit Jahrzehnten so verbreitet und erfolgversprechend, so dass sie mittlerweile auch außerhalb der IT, z.B. bei der Erstellung von Fachliteratur[32] Verwendung findet.

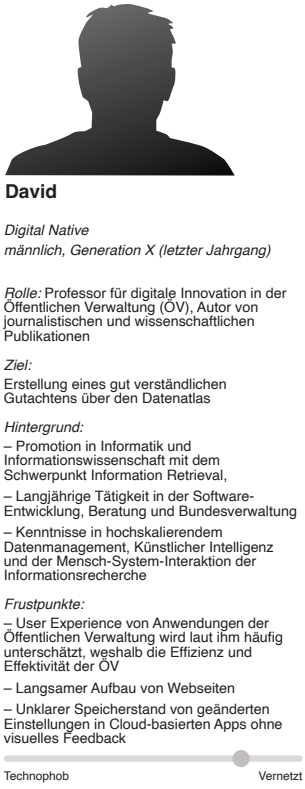


Abbildung 3.11: Persona am Beispiel des Autors

Golden Rules of Interface Design

Bei der konkreten Gestaltung von *User Interfaces (UI)*[33] finden häufig die sogenannten *Golden Rules* Shneidermans Anwendung, welche als Heuristik experimentell ermittelt und über einen Zeitraum von mehr als zwei Jahrzehnten verfeinert und präzisiert wurden (Shneiderman und Plaisant, 2005).

Der wesentliche Vorteil der acht Regeln besteht darin, dass sie sich auf fast alle Arten von interaktiven Systemen, d.h. nicht nur IT-Systemen, anwenden lassen und sich *unmittelbar positiv* auf die *Usability* dieser auswirken:

1. *Konsistenz* Ein gut nutzbares System lässt sich konsistent bedienen, d.h. seine Interaktionsgrammatik[34] und Terminologie folgt gleichbleibenden Mustern.

[31] Siehe Kapitel 4.

[32] `https://it-in-bibliotheken.de/mitarbeit.html`; Letzter Abruf: 21.07.2025

[33] *Nutzerschnittstelle*; d.h. die grafische Schnittstelle, z.B. ein Webseiten-Interface, über das die *Mensch-System-Interaktion* stattfindet.

[34] D.h., dass beispielsweise Konventionen der Bedienplattform wie die Platzierung von Abschluss-/Okay- oder Abbruch-Buttons gleich bleibt Dies erstreckt sich auch auf die Konsistenz von Interaktonssequenzen, d.h. dass vergleichbare Handlungsfolgen in ähnlichen Situationen durch Nutzende befolgt werden müssen.

2. *Universelle Bedienbarkeit* Ein Interface sollte verschiedene Nutzergruppen und Nutzungsstrategien unterstützen, z.B. die Mausbedienung für Laien bzw. Shortcuts[35] für Experten.

3. *Informatives Feedback* Das System informiert proaktiv über den aktuellen Zustand und die zu erwartenden Konsequenzen aus den durch den Nutzer ausgeführten Aktionen.

4. *Abgeschlossene Aktionen* Einzelne oder Sequenzen von Nutzeraktionen sind deutlich abgeschlossen, d.h. dass das System darüber informiert, wenn eine Aktion beendet ist.

5. *Fehlervermeidung* Das System sollte mögliche Fehlbedienungen proaktiv verhindern, z.B. durch Ausgrauen irrelevanter Menu-Einträge, und dafür Sorge tragen, dass Nutzende während der *Mensch-System-Interaktion* nicht in eine gefährliche Situation kommen können.

6. *Einfache Umkehrbarkeit von Aktionen* Ausgeführte Aktionen sollten jederzeit rückgängig gemacht werden können, z.B. weil Nutzende einen Fehler gemacht haben oder weil die Handlung nicht zum erwarteten Ergebnis geführt hat. Dies ermöglicht explorativ es Lernen.

7. *Geringe Belastung des Arbeitsgedächtnisses* Anwenderinnen und Anwender sollten das System ohne Rückgriff auf das Arbeitsgedächtnisses verwenden können, d.h. Informationen, die relevant für eine Arbeitsaufgabe sind[36], sollten sichtbar bzw. im Zugriff sein. Der Systemstatus muss jederzeit erkennbar sein.

8. *Kontrollierbarkeit* Nutzende haben die Kontrolle über das System, d.h. sie interagieren proaktiv und nicht reaktiv – in anderen Worten: Sie initiieren Aktionen, die das System ausführen soll (Vermeidung von Akausalität[37]).

Diese Regeln schlagen sich auch in den einschlägigen Normen der DIN EN ISO 9241-Familie nieder, welche die Basis für Software-Ergonomie und Nutzerfreundlichkeit darstellen und deren Einhaltung u.a. für die Bundesverwaltung verbindlich ist (siehe Abschnitt 3.7). Insbesondere einschlägig ist hier die DIN EN ISO 9241-210 (2020), welche eine Tätigkeitsanleitung zur menschzentrierten Gestaltung enthält.

Evaluierung

Da im Entwicklungsprozess fortwährend Annahmen über die Anforderungen der Zielgruppe getroffen werden, müssen Nutzende dauerhaft in die Evaluierung einbezogen werden.

Die Evaluierung erfolgt häufig unter der Verwendung eines sogenannten *Minimum Viable Products*[38] oder MVP.

Das MVP wird hierbei dazu herangezogen, frühzeitig und schnell Nutzendenfeedback zu generieren, um Fehlentwicklungen zu vermeiden, wie sie beispielsweise durch die Verwendung

[35] *Tastaturkurzbefehle;* wie Strg + C bzw. ⌘ + C zum Kopieren.

[36] Zum Beispiel sinnvolle Vorbelegungen mit Standardwerten, Erläuterungen von Fachbegriffen oder Abkürzungen etc.

[37] Brian R. Gaines. The Technology of Interaction—Dialogue Programming Rules. *International Journal of Man-Machine Studies*, 14(1):133–150, 1981. ISSN 0020-7373. DOI: 10.1016/S0020-7373(81)80037-5

[38] *Minimal brauchbares Produkt;* eine prototypische, i.d.R. funktionsfähige Iteration der Produktentwicklung.

klassischer Projektmanagment-Methoden ohne regelmäßige Einbeziehung des Nutzendenfeedbacks entstehen. Abbildung 3.12 illustriert diesen Effekt.

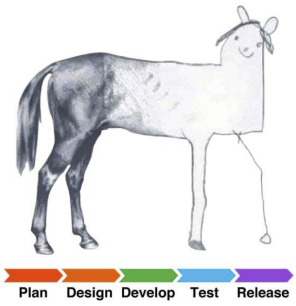


Abbildung 3.12: Risiko von Fehlentwicklungen bei Nutzung des Wasserfall-Projektmanagementmodells (Internet-Meme; o.A., o.D.)

Hierbei ist es wichtig zu unterstreichen, dass im Projektverlauf in der Regel nicht nur *ein* MVP entwickelt wird, sondern dass dieses fortwährend aufgrund von Feedback der Nutzenden weiterentwickelt wird, um ultimativ ein voll funktionstüchtiges Produkt zu erhalten.

Das MVP entspricht damit einem *Lösungsansatz* im menschzentrierten Gestaltungsprozess (siehe Abb. 3.10).

Es existiert eine Vielzahl an Methoden zur Durchführung von Evaluierungen, die hier nicht dargestellt werden kann. Stattdessen sei auf weit verbreitete Fachliteratur wie Preece et al. (2002), Shneiderman und Plaisant (2005) oder Albert und Tullis (2023) verwiesen.

Grob lassen sich die Evaluierungsansätze in *Beobachtungen* und *Inspektionen* unterscheiden, wie in Abb. 3.13 gezeigt wird.

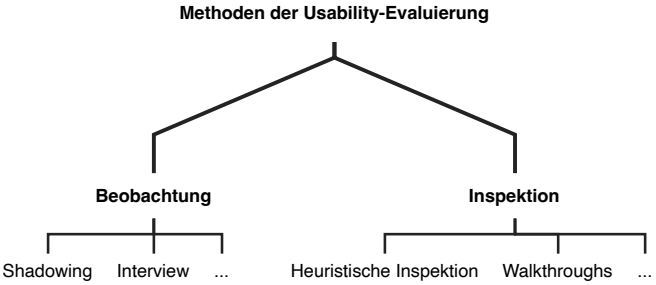


Abbildung 3.13: Überblick über Methoden der Usability-Evaluierung

Das Kosten- und Aufwand-Argument gegen die Durchführung sogenannter Usability- und Nutzendentests aus dem Beobachtungsbereich gilt seit spätestens den frühen 1990er-Jahren wissenschaftlich als widerlegt und wurde seitdem nur minimal auf die Einbeziehung diverser Nutzendengruppen angepasst[39].

Studien belegen, dass bereits Tests mit *fünf Personen* hinreichend aussagekräftig sind, um die *User Experience* einer *Mensch-System-Interaktion* durch die Behebung der entdeckten Usability-Probleme positiv zu beeinflussen. Beobachtungstests gelten anderen Testformen als überlegen, da sie tiefere Einblicke in konkrete *Mensch-System-Interaktionen* ermöglichen[40].

Die Evaluierung von IT-Systemen unter Einbeziehung Nutzender ist als *Stand der Technik* zu betrachten. DIN EN ISO 9241-220 (2020) beinhaltet beispielsweise eine detaillierte Beschreibung von Prozessen und Methoden, um die menschliche Nutzung interaktiver Systeme analysieren, gestalten und bewerten zu können.

Die in den Kapiteln 4 und 5 vorgestellten Erkenntnisse basieren auf Inspektionstests.

[39] Jakob Nielsen. *Why You Only Need to Test with 5 Users*, 2000. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/. Letzter Abruf: 27.07.2025

[40] Jakob Nielsen und Thomas K. Landauer. A Mathematical Model of the Finding of Usability Problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, Seiten 206–213. ACM, New York, NY, USA, January 1993. ISBN 0-89791-575-5. DOI: 10.1145/169059.169166

Servicestandard der Öffentlichen Verwaltung

Für die Öffentliche Verwaltung existiert der sogenannte *Servicestandard*, welcher als DIN SPEC 66336 (2025) vorliegt. Sein Vorläufer wurde bereits 2020 durch das Bundesministerium des Innern (Bmi) veröffentlicht[41] und umfasste Aspekte wie Nutzerzentrierung, die Nutzung eines iterativen Vorgehens, die Wichtigkeit interdisziplinärer Zusammenarbeit, offenes Arbeiten, die Forderung eines robusten Betriebs und die Wirkungsmessung.

[41] https://digitalservice.bund.de/projekte/servicestandard; Letzter Abruf: 27.07.205

Der Anwendungsbereich des aktualisierten *Servicestandards* wird wie folgt benannt:

> [*Der Servicestandard*] legt die Qualitätsanforderungen an informationstechnische Systeme der digitalen öffentlichen Verwaltung fest. Dieses Dokument ist anwendbar für informationstechnische Systeme, die für den übergreifenden informationstechnischen Zugang zu den Verwaltungsleistungen von Bund, Ländern und Kommunen genutzt werden.
>
> Dieses Dokument adressiert neu zu entwickelnde Onlineservices und -portale sowie Basisdienste (3.3). Dieses Dokument adressiert nicht bereits implementierte Onlineservices und -portale, es sei denn, der Onlineservice (3.16) wird einer grundlegenden Überarbeitung (3.9) unterzogen.[42]

[42] DIN SPEC 66336. *Qualitätsanforderungen für Onlineservices und -Portale der Öffentlichen Verwaltung (Servicestandard)*, 2025. https://servicestandard.gov.de/din-spec-66336/. Letzter Abruf: 28.07.2025

Auch wenn dieser Abschnitt suggeriert, dass der *Servicestandard* nur auf Dienste, die Verwaltungsleistungen anbieten, Anwendung findet, kann aus Art. 3.3 DIN SPEC 66336 (2025) durchaus geschlussfolgert werden, dass es sich beim Datenatlas um einen Dienst zur „Datenhaltung und -bereitstellung“ handelt, insbesondere dann, wenn dieser zur Erbringung von Verwaltungsleistungen genutzt wird, was durch die in Kapitel 2 vorgestellten *minimalen Use Cases der Bundesverwaltung* nicht ausgeschlossen werden kann.

> *Basisdienste Basiskomponenten* zentral für die digitale Verwaltung bereitgestellte Grundfunktionalitäten wie beispielsweise zur Bezahlung, zur Authentifizierung, zur Datenhaltung und -bereitstellung oder zur Realisierung der erforderlichen Datensicherheit[43]

[43] DIN SPEC 66336 (2025); 3.3

Der *Servicestandard* folgt dem oben geschilderten *Stand der Technik* und listet 13 Kriterien[44] für die Erstellung und Bereitstellung von menschzentrierter Software seitens der Öffentliche Verwaltung auf und erläutert diese gut verständlich:

[44] https://servicestandard.gov.de/#kriterien; Letzter Abruf: 27.07.2025

1. *Nutzende verstehen und Bedürfnisse erkennen* Schaffen Sie die Grundlage für einen Service, der Nutzenden wirklich hilft. Finden Sie heraus, wer Ihre Nutzenden sind und was sie in ihrer Situation brauchen. Schauen Sie nicht nur darauf, wie Nutzende den Service bedienen. Entscheidend ist, was sie erreichen wollen und was ihnen dabei hilft.

2. *Problem beschreiben und Ziele bestimmen* Beschreiben Sie, welches Problem der Service in Zukunft löst. Legen Sie klare Ziele fest, die mit dem Service erreicht werden sollen.

3. *Verantwortung übernehmen und Ressourcen sichern* Legen Sie organisatorische Strukturen fest und klären Sie, wer die Verantwortung für den Service trägt. Der Service muss verlässlich sein und fortlaufend verbessert werden. Dafür braucht es geklärte Zuständigkeiten und ausreichend Ressourcen.

4. *Lösungen entwickeln, testen, anpassen und Fachwissen einbinden* Bauen Sie den Service Schritt für Schritt auf. Beziehen Sie Experten und Expertinnen aus verschiedenen Fachbereichen ein. Passen Sie den Service regelmäßig an die Bedürfnisse der Nutzenden an.

5. *Bestehendes wiederverwenden und Neues gemeinsam gestalten* Nutzen Sie bestehende Services, bevor Sie einen neuen entwickeln. Entwickeln Sie neue Lösungen gemeinsam mit anderen Stellen der Verwaltung.

6. *Barrierefreie Nutzung sicherstellen und Teilhabe stärken* Entwickeln Sie einen Service, den alle nutzen können, egal, welche Fähigkeiten oder Kenntnisse Nutzende haben. Der Service muss verständlich, einfach zu bedienen und leicht zu finden sein.

7. *Offene Standards beachten und Schnittstellen bereitstellen* Stellen Sie sicher, dass der Service mit offenen Standards entwickelt wird. Setzen Sie Schnittstellen für den automatisierten Austausch von Daten ein.

8. *Datenschutz umsetzen und Risiken reduzieren* Planen Sie Datenschutz von Anfang an ein. Erkennen Sie die Risiken Ihrer Daten-Verarbeitung und setzen Sie geeignete technische und organisatorische Maßnahmen ein, um sie zu reduzieren.

9. *Sicherheit herstellen und Vertrauen schaffen* Sorgen Sie von Anfang an dafür, dass der Service sicher ist und auch bei außergewöhnlich hoher Belastung funktioniert. Stellen Sie sicher, dass es Unterstützung gibt, wenn Nutzende sie brauchen.

10. *Open Source nutzen und Code teilen* Veröffentlichen Sie den Quellcode, wenn Sie neue Services entwickeln. Bauen Sie auf bestehender, offener Software auf. Gemeinsam mit anderen machen Sie Software dadurch verfügbar, besser und sicherer.

11. *Verfügbarkeit sichern und Störungen beheben* Sorgen Sie dafür, dass der Service erreichbar ist, wenn Nutzende ihn brauchen. Planen Sie Maßnahmen bei Störungen oder einem Ausfall.

12. *Wirkung messen und auf Ergebnissen aufbauen* Sammeln Sie Feedback von Nutzenden und Daten anhand von Kennzahlen, die Sie am Anfang festgelegt haben. Mit diesen Informationen messen Sie die Wirkung. Nutzen Sie die Ergebnisse, um Erkenntnisse mit Beteiligten zu teilen und den Service weiter zu verbessern.

13. *Rechtliche Hürden erkennen und Regelungen verbessern* Achten Sie darauf, ob rechtliche Vorgaben die einfache Nutzung von

Services erschweren. Setzen Sie sich für die Änderungen ein, die den Service für Nutzende einfacher machen.

Aufgrund der bekannten Limitationen des Gutachtens finden die folgenden Kriterien des *Servicestandards* im weiteren Verlauf keine weitere Beachtung, da sie mit der in Kapitel 4 vorgestellten Methodik kaum seriös bewertet werden können:

- Kriterium 3 (Servicestandard der Öffentlichen Verwaltung)
- Kriterium 6 (Servicestandard der Öffentlichen Verwaltung)
- Kriterium 8 (Servicestandard der Öffentlichen Verwaltung)
- Kriterium 9 (Servicestandard der Öffentlichen Verwaltung)
- Kriterium 11 (Servicestandard der Öffentlichen Verwaltung)

3.5 *Informationssuchstrategien*

Im Rahmen der viele Jahrzehnte andauernden Forschung im Feld des interaktiven Information Retrievals wurden verschiedene *Informationssuchstrategien*[45] beschrieben.

Zur Unterstützung dieser ISS wurden unterschiedliche Interaktionsmöglichkeiten, wie gerichtete oder explorative Suchansätze, beschrieben und evaluiert[46].

Dieser Abschnitt stellt die wesentlichen Ansätze zur Interaktionsunterstützung während der *Informationsrecherche* kurz vor und ordnet diese ihren typischen Einsatzzwecken auf Grundlage der Erkenntnisse der *Mensch-System-Interaktion*-Forschung in diesem Feld zu.

Der Aufbau dieses Abschnitts orientiert sich, wenn nicht anders angeben, an Zellhöfer (2015).

[45] *Informationssuchstrategie*; auch als *information seeking strategy (ISS)* in der einschlägigen Literatur bekannt.

[46] J. N. Belkin, G. P. Marchetti, und C. Cool. BRAQUE: Design of an Interface to Support User Interaction in Information Retrieval. *Inf. Process. Manage.*, 29(3):325–344, 1993. http://dx.doi.org/10.1016/0306-4573(93)90059-M

Mensch-System-Interaktion bei der Informationsrecherche

Abschnitt 3.2 stellte verschiedene Möglichkeiten vor, wie die *Informationsrecherche* System-seitig ermöglicht werden kann.

Hierbei wurde der Schwerpunkt auf die Darstellung der historischen Entwicklung des stetig wachsenden Funktionsumfang gelegt, ohne tiefer zu ergründen, welche konkreten Use Cases mit den verschiedenen Funktionen bedient werden können.

Letztendlich wird die User Experience einer *Informationsrecherche* durch zwei wesentliche Faktoren beeinflusst: einerseits durch die technische „Mächtigkeit“ des *Information-Retrieval-Systems*, welche maßgeblich durch das gewählte Information-Retrieval-Modell[47] beeinflusst wird, und andererseits durch die konkretete Unterstützung der Nutzenden bei der Erfüllung ihrer Arbeitsaufgaben bzw. ihrer dabei ablaufenden kognitiven Prozesse.

[47] Siehe Abschnitt 3.3.

Informationsbedürfnis Um Nutzende bei der Suche unterstützen zu können, ist es notwendig, ihre Motivation, die von ihnen durchlaufenen kognitiven Veränderungsprozesse und ihr Streben nach der Befriedigung ihres *Informationsbedürfnisses* – d.h. ihr individuelles Ziel bei der *Informationsrecherche* – zu verstehen.

Der Begriff des *Informationsbedürfnisses* wurde bereits in Abschnitt 2.2 oberflächlich eingeführt. Aufgrund seiner Wichtigkeit für das Verständnis der Motivation und der Ziele Recherchierender soll er hier weiter präzisiert werden.

Das *Informationsbedürfnis* ist eng mit der *ASK*[48]*-Hypothese* verbunden: Nutzende entdecken eine Anomalie ihres Wissenszustands im Rahmen einer beliebigen Situation, woraus sich das *Informationsbedürfnis* ergibt, um die festgestellte Wissenslücke schließen zu können.

[48] *Anomalous State of Knowledge*; Anomalie des Wissenszustands.

Das *Informationsbedürfnis* selbst lässt sich nicht vom Vorwissen, dem Arbeitskontext, von individuellen Zielen etc. der Recherchierenden trennen. Im Rahmen der *Informationsrecherche* müssen diese Faktoren zusätzlich herangezogen werden, um eine zufriedenstellende Recherche zu ermöglichen. Herausfordernd dabei ist, dass Recherchierende diese Informationen nicht mitteilen, da sie aus ihrer Sicht offensichtlich sind. Im optimalen Fall muss das *Information-Retrieval-System* diese Informationen ergänzen bzw. Suchende entsprechend unterstützen.

Das folgende Beispiel soll diese abstrakte, informationswissenschaftliche Beschreibung verdeutlichen.

> Eine Mitarbeiterin der Bundesverwaltung möchte mehr über ein Förderprogramm erfahren. Der konkrete Name des Programms ist ihr entfallen (*Wahrnehmung des ASK*), sie kennt aber den Kreis der Antragsberechtigten und das ungefähre Themenfeld.
>
> Ihr Wissenszustand ist anomal, weil er unvollständig und lückenhaft ist.
>
> Aufgrund ihres festgestellten *Informationsbedürfnisses* nutzt sie den Datenatlas, um Daten zum konkreten Förderprogramm zu recherchieren.
>
> Dazu bietet der Datenatlas verschiedene Unterstützungsmöglichkeiten an, die weiter unten diskutiert werden.

Das wesentliche Ziel der Interaktion mit einem *Information-Retrieval-System* besteht folglich darin, das *Informationsbedürfnis* zu befriedigen[49].

[49] Man beachte die Nähe zum Begriff *Usability*; siehe Abschnitt 3.4.

Interaktives Information Retrieval Die Definition über das Individualziel der einzelnen Nutzenden verbindet das Forschungsfeld des interaktiven Information Retrievals mit den in Abschnitt 3.4 vorgestellten Ansätzen der menschzentrierten Gestaltung.

Durch die Einnahme dieser Perspektive rückt die kognitiv begründbare Interaktion mit *Information-Retrieval-Systemen* in den Mittelpunkt der Betrachtung, welche bestmöglich unterstützt werden muss, wie bereits Belkin (1996) anmerkt.

Zum Erkenntnisgewinn in diesem komplexen Spannungsfeld wird zumeist auf sozio-psychologische Methoden wie empirische Studien unter Einbeziehung realer Nutzender zurückgegriffen, um belastbare Erkenntnisse zu erhalten[50].

Da sich die Forschung im Bereich der IR-Interaktion bis zur Mitte der 1950er-Jahre zurückverfolgen lässt, sollen im Folgenden nur zwei zentrale Nutzenden-orientierte Modelle erläutert werden, die sowohl von Bedeutung für den DATENATLAS sind als auch das gesamte Forschungsfeld repräsentieren können.

Weiterführende Diskussionen finden sich u.a. bei Kelly (2009); Hearst (2009); Ingwersen (1992) oder Zellhöfer (2015).

[50] Peter Ingwersen. *Information Retrieval Interaction*. Taylor Graham, London, 1992. http://www.gbv.de/dms/hbz/toc/ht004327073.PDF

Berry Picking Model Das von Bates vorgestellte *berry picking model*[51] beschreibt u.a., dass das *Informationsbedürfnis* von Suchenden dynamisch ist und sich während der Interaktion mit einem *Information-Retrieval-System* stetig weiterentwickelt. Diese kontinuierliche Weiterentwicklung des *Informationsbedürfnisses* wird auf Lerneffekte zurückgeführt, da die Suchenden während der Recherche Kenntnisse sowohl über den zur Verfügung stehenden Dokumentenbestand als auch ihr *Informationsbedürfnis* auf- und ausbauen.

Diese Erkenntnisse sind durch mehrere Studien, u.a. die von Ellis (1989), bestätigt worden[52].

[51] „Beeren-Pflück-Modell"; J. Marcia Bates. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, 13(5):407–424, 1989

[52] A. Marti Hearst. *Search User Interfaces*. Cambridge Univ. Press, Cambridge, 2009

Information Search Process Model Kuhlthaus ISP[53]-Modell basiert auf einer groß angelegten Nutzenden-Studie mit 385 Teilnehmenden an 21 verschiedenen Bibliotheksstandorten.

Kuhlthaus Studie zeigt, dass Nutzende während der *Informationsrecherche* verschiedene Phasen durchleben, z.B. eine explorative Phase, in der sie mehr darüber herausfinden wollen, wie sie ihre aktuelle Rechercheaufgabe lösen können. Diese Phasen sind mit verschiedenen Emotionen verbunden, wie beispielsweise Unsicherheit oder Verwirrung während der Exploration, da die Suchenden noch nicht in der Lage sind ihr *Informationsbedürfnis* hinreichend präzise spezifizieren zu können.

Die Studie stellt ferner fest, dass Suchende unterschiedliche *Informationssuchstrategien* nutzen, um Phasen-spezifische Herausforderungen zu überwinden.

Die beobachtete Transition zwischen verschiedenen *Informationssuchstrategien* wird u.a. auch durch Belkin (1993); Ellis und Haugan (1997) oder Reiterer et al. (2000) beschrieben.

[53] *Information Search Process*; Informationssuchprozess; C. C. Kuhlthau. Inside the Search Process: Information Seeking from the User's Perspective. *Journal of the American Society for Information Science*, 42(5):361–371, 1991

Auswirkungen auf die Mensch-System-Schnittstelle Auf Grundlage der oben genannten Studien und Modelle wurden bereits frühzeitig *Information-Retrieval-Systeme* realisiert, die verschiedene *Informationssuchstrategien* unterstützen, wie z.B. von Belkin et al. (1993) oder Marchionini et al. (2000).

Interessanterweise wird bereits Ende der 1980er-Jahre ohne die Grundlage der oben präsentierten empirisch begründeten Erkenntnisse I³R implementiert[54], welches sowohl gerichtete Such- als auch

[54] Croft und Thompson (1987)

Browsing-Funktionen anbietet.

Diese unterschiedlichen *Informationssuchstrategien* werden in den folgenden zwei Abschnitten näher vorgestellt.

Gerichtete Suche und Anfrage-basiertes Information Retrieval

Die *Gerichtete Suche* kann als die wohl älteste *Informationssuchstrategie* betrachtet werden und ist noch immer die vorherrschende Strategie, welche sowohl durch *Datenbankmanagement-Systeme*[55] oder *Information-Retrieval-Systeme*[56] unterstützt wird.

[55] Siehe Abschnitt 2.2.

[56] Siehe Abschnitt 2.2.

Bei der *Gerichteten Suche* formulieren Nutzende eine Anfrage (*Query*), die ihr *Informationsbedürfnis* widerspiegelt. Das entsprechende System berechnet daraufhin eine passende Antwort (*Response*), welche entweder akzeptiert oder mittels einer neuen Anfrage modifiziert wird.

Dieser Prozess wird auch als *Query-Response-Zyklus* bezeichnet und ist in Abb. 3.14 illustriert.

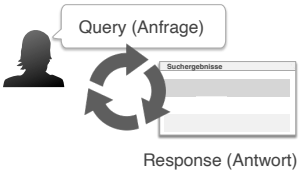


Abbildung 3.14: Query-Response-Zyklus

Da der Prozess iterativ angelegt ist, birgt er das Risko, dass Nutzende die Interaktion mit dem System aus Gründen der Ermüdung oder der Unzufriedenheit mit den Ergebnissen vorzeitig, d.h. ohne dass ihr *Informationsbedürfnis* befriedigt ist, abbrechen. Dies wirkt sich wiederum negativ auf die *User Experience* aus.

Eine umfassende Diskussion des *Query-Response-Zyklus* und dessen Überwindung findet sich bei White und Roth (2009).

Mittlerweile wurden eine Vielzahl an Ansätzen der *Gerichteten Suche* vorgeschlagen, von denen im Folgenden nur die beiden vorgestellt werden, welche in der Praxis weite Verbreitung erfahren haben: *Query by Language* und *Query by Example*.

Query by Language Der intuitivste Weg, mit einem *Information-Retrieval-System*[57] zu interagieren, ist die Nutzung natürlicher Sprache. Bilal (2000) zeigt auf Grundlage einer Studie, dass Menschen, die keinerlei Erfahrung im Umgang mit *Information-Retrieval-Systemen* haben, diesen typischerweise Fragen in natürlicher Sprache stellen.

[57] Der Argumenation aus Abschnitt 3.3 folgend subsumieren wir *Datenbankmanagement-Systeme* unter den *Information-Retrieval-Systemen*.

Nichtsdestotrotz unterstützen die meisten *Information-Retrieval-Systeme* nur eine Teilmenge der natürlichen Sprache zur Anfrageformulierung: Schlagwörter.

Um die Antworterzeugung von *Information-Retrieval-Systemen* besser steuern zu können, unterstützen diese zumeist seit langem Boolesche und weitere Operatoren, wie bereits in Abschnitt 3.2 angesprochen wurde.

Einen Sonderfall stellen *relationale Datenbankmanagement-Systeme* dar, welche sich wie bereits in Abschnitt 2.2 dargelegt, auf künstliche Anfragesprachen wie SQL stützen. Diese Art der Anfrageformulierung eignet sich gut für Expertinnen und Experten, wirkt jedoch auf andere Personenkreise aufgrund der mit ihnen verbundenen Lernkurve abschreckend.

Das Anfrageformulierungsproblem wurde bereits in Abschnitt 2.2 kurz eingeführt[58]. Das Problem besteht darin, dass Nutzende ihr *Informationsbedürfnis* in Form einer natürlichsprachigen oder künstlichen Sprache in Form einer Anfrage an das *Information-Retrieval-System* übermitteln müssen.

[58] Siehe Abbildung 2.5.

Das heißt, dass bei der Anfrageformulierung angenommen wird, dass Nutzende ihr *Informationsbedürfnis* hinreichend präzise spezifizieren können, um zufriedenstellende Ergebnisse zu erhalten.

Diese implizite Annahme gilt aufgrund der oben vorgestellten Studienlage als widerlegt.

Tatsächlich spiegelt diese Annahme nur einen Extremfall des kognitiven Zustands Nutzender, der in Abb. 3.16 (1) dargestellt ist, wider. Nutzende in dieser Fallgruppe verfügen über eine hohe Gewissheit über ihr Suchziel und können die Relevanz der Treffer gut einordnen. Dies ist beispielsweise der Fall bei der sogenannten *Known-Item-Search*, bei der Nutzende Dokumente suchen, über deren Existenz sie bereits wissen.

Zur Kompensation des Anfrageformulierungsproblems werden u.a. *kontrollierte Vokabulare* eingesetzt, für deren korrekte Verwendung Nutzende jedoch im Vorfeld der Datenerfassung und -recherche geschult werden müssen.

Die nötigen Überlegungen mit Bezug zum Anfrageformulierungsproblem aus dem Bereich der *Datenqualitätssicherung* werden in Abschnitt 3.6 thematisiert.

Query by Example Zur Adressierung des Anfrageformulierungsproblems, vor allem für Laien-Nutzende im Bereich relationaler Datenbanken, schlägt Zloof (1975) das *Query by Example (*QBE*)*-Paradigma vor. Hierbei werden Anfragen anstelle einer künstlichen Anfrage in SQL mittels eines Tabellengerüstes formuliert, welchem die Ergebnisse genügen müssen.

Heutzutage wird dieser Umsatz von einer Vielzahl visueller *Datenbankmanagement-Systeme* implementiert. Den Leserinnen und Lesern ist diese Art der Anfrageformulierung sicherlich aus *Microsoft Access* bekannt.

Vergleichbare Lösungen existieren für die inhaltsbasierte Bildsuche, welche besonders unter dem Anfrageformulierungsproblem[59] leidet, z.B. in Form des *Query by Image*-Ansatzes von Flickner et al. (1995). Derartige Lösungen sind mittlerweile bei großen Suchmaschinen wie Bing oder Google Standard.

[59] Dies ist auf den polythetischen Charakter von Bildern zurückzuführen; oder in anderen Worten: „Ein Bild sagt mehr als tausend Worte“.

Exploratorative Suche

Aufgrund der vorgestellten Studienlage wurde deutlich, dass nicht davon ausgegangen werden kann, dass Nutzende immer in der Lage sind, eine Anfrage an ein *Information-Retrieval-System* formulieren zu können, was wiederum die Voraussetzung für den Interaktionsbeginn einer *Gerichteten Suche* ist.

Obwohl diese Studienergebnisse seit langem vorliegen, setzen

viele Endanwender-Systeme zur *Informationsrecherche* zumindest initial ausschließlich auf diese *Informationssuchstrategie*, wie auch White und Roth (2009) kritisieren.

Explorative Suchansätze setzen keine initale Anfrageformulierung voraus, sondern bieten Nutzenden Zugang zu den Inhalten einer Datenbank, indem sie z.B. ähnliche Inhalte gruppieren und damit Nutzenden einen Einblick in die Art und den Umfang der recherchierbaren Datenbestände geben.

Damit bietet sich der Einsatz explorativer Ansätze, von denen im Folgenden zwei vorgestellt werden, insbesondere dann an, wenn anzunehmen ist, dass die Zielgruppe des Systems entweder (aus diversen Gründen) Probleme bei der Präzisierung ihres *Informationsbedürfnisses* hat oder sich vor allem mit dem Bestand der Datenbank vertraut machen will[60].

[60] Für fachlich versierte Lesende bietet sich ein alternativer Erklärungszugang an. Während die *Gerichtete Suche* daraufhin optimiert wird, ein hohes Maß an *Precision*, d.h. eine Minimierung an irrelevanten Dokumenten unter den Ergebnissen, zu erreichen, lautet die Zielstellung explorativer Suchstrategien, ein hohes Maß an *Recall* zu erreichen, damit Nutzende möglichst viele potenziell relevante Dokumente als Systemantwort enthalten.

Browsing Grob gesprochen bezeichnet *Browsing* die Navigation innerhalb eines Informationsraums anhand – nicht notwendigerweise sichtbarer – Verbindungen zwischen verschiedenen Dokumenten. Die dem Browsing zugrundeliegende Struktur ist häufig nur schwach bis gar nicht für Nutzende erkennbar bzw. teilweise zu komplex für eine Visualisierung.

Der Begriff *Browsing* ist den Lesenden aus der täglichen Navigation im *World Wide Web* bekannt, in dem die Bewegung entlang von Hyperlinks erfolgt.

In Abschnitt 3.2 wurden weitere Browsing-Möglichkeiten auf Grundlage der Basisklassifikation von Publikationen, häufig genutzten Suchanfragen[61] oder anhand geografischer Kriterien[62] präsentiert.

[61] Siehe Abbildung 3.4.

[62] Siehe Abbildung 3.5.

Andere Browsing-Ansätze, z.B. auf Grundlage zeitlicher oder räumlicher Nähe von Datensätzen usw., sind realisierbar, um nur einige Beispiele zu nennen.

Für einen weitergehenden Überblick sei auf Hearst (2009) oder Zhang (2008) verwiesen.

Facettierte Navigation Im Gegensatz zum *Browsing*, welches eine i.d.R. schwach strukturierte Exploration des Informationsraums ermöglicht, basiert die *Facettierte Navigation* bzw. das facettierte Browsing auf Ordnungskriterien, den sogenannte *Facetten*, um Nutzende bei der Exploration zu unterstützen.

Facetten basieren auf Eigenschaften, die Teilmengen der in der Datenbank enthaltenen Dokumente voneinander abgrenzen und sich deshalb zur Nutzung als Filter der Datenbank eignen. Normalerweise werden orthogonale Facetten, die unabhängige Aspekte der Dokumente beschreiben, gewählt, die sich optimalerweise hierarchisch gliedern lassen[63].

[63] A. Marti Hearst. *Search User Interfaces*. Cambridge Univ. Press, Cambridge, 2009

Facetten werden zumeist manuell ausgewählt, um deren Verständlichkeit und Nützlichkeit für die adressierte Zielgruppe zu gewährleisten[64]. Eine Erstellung von Facetten mittels Verfahren der

[64] White und Roth (2009)

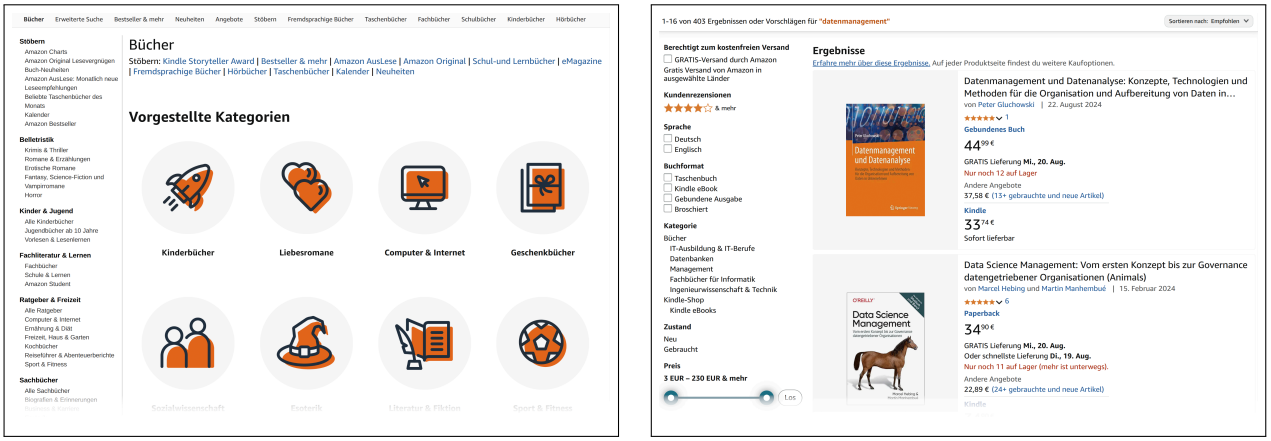
Künstlichen Intelligenz, z.B. mittels Clustering, ist möglich, birgt jedoch Risiken bzgl. deren Verständlichkeit für Nutzende und wirkt sich damit negativ auf deren *Usability* aus.

Ein anderer Faktor, der sich negativ auf die *Usability* von Facetten auswirkt, sind zu viele bzw. zu tief untergliedernde Facetten[65,66], welche die Übersichtlichkeit mindern.

[65] A. Marti Hearst. *Search User Interfaces*. Cambridge Univ. Press, Cambridge, 2009

[66] Siehe dazu auch Abschnitt 3.4, insb. Kriterium 7 (Golden Rules of Interface Design).

Beispielhafte Facetten im Kontext des Datenatlas wären die obersten Bundesbehörden mit ihren nachgeordneten Behörden oder das Erstellungsdatum eines Datensatzes, welches sich nach Jahren, Quartalen oder Monaten weiter untergliedern lässt.



(a) Browsing-Einstieg anhand Literatur-Kategorien

(b) Facettierung (links) nach der Schlagwortsuche „datenmanagement“

Abbildung 3.15: Ausschnitte Browsing- und Facetten-Ansichten bei Amazon; `https://www.amazon.de`; Letzter Abruf: 16.08.2025

Abbildung 3.15 stellt den Browsing-Einstieg (a) für Bücher des Online-Händlers Amazon mit der Facettierung innerhalb der Trefferliste (b) gegenüber. Dieses Angebot an *Informationssuchstrategien* ist nicht nur bei Online-Händlern mittlerweile weit verbreitet, wenngleich Facetten wesentlich häufiger Verwendung finden. Im Rahmen der professionellen *Informationsrecherche*[67] sind sowohl *Browsing* als auch *Facettierte Navigation* als *Stand der Technik* anzusehen.

[67] Siehe Abschnitt 3.2.

Ein typisches Anwendungsbeispiel von Facetten ist in Abb. 3.15 (b) zu sehen. Hier wird die *Facettierte Navigation* mit einer *Gerichteten Suche* kombiniert. In diesem Anwendungsfall dient die Schlagwort-basierte Suche als erster „Filter“ auf dem gesamten Datenbestand, der dann weiter mittels Facetten exploriert und eingegrenzt werden kann, um relevante Dokumente entsprechend dem *Informationsbedürfnis* zu erhalten und zeitgleich mehr über den Inhalt der Datenbank zu erfahren.

Einsatzgebiete gerichteter und explorativer Suchansätze

Abbildung 3.16 stellt vier typische Extremfälle von *Informationsbedürfnissen* Nutzender nach Ingwersen (1996) einander gegenüber und visualisiert, welche ISS die jeweilige *Informationsrecherche* unter Einbeziehung der Ausprägung des aktuellen *Informationsbedürfnisses* bestmöglich unterstützt.

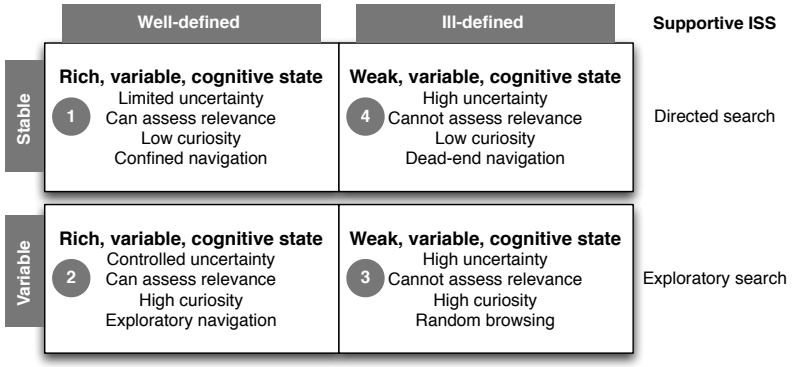


Abbildung 3.16: Matrix der vier intrinsischen Extremfälle an Informationsbedürfnis-Ausprägungen nach Ingwersen (1996)

Hierbei wird deutlich, dass der Einsatz *gerichteter Informationssuchstrategien* bedingt, dass Recherchierende über ein stabiles *Informationsbedürfnis* verfügen *müssen*, um überhaupt Anfragen formulieren zu können.

Dem gegenüber stehen variable *Informationsbedürfnisse*, die von Neugier, Unwissenheit und einem gewissen Maß an Unsicherheit geprägt sind und bestmöglich durch *explorative Informationssuchstrategien* unterstützt werden.

Es ist folglich essenziell, sich bereits vor der Implementierung eines *Information-Retrieval-Systems* mit den *typischen Suchszenarien* der Zielgruppe auseinanderzusetzen bzw. diese zu erheben. Dies steht in Einklang mit dem menschzentrierten Gestaltungsprozess, der in Abschnitt 3.4 vorgestellt wurde.

Auch Kriterium 1 (Servicestandard der Öffentlichen Verwaltung) unterstreicht die Wichtigkeit des Verständnisses von Nutzenden und ihren Bedürfnissen, um ein System zu entwickeln, dass den „Nutzenden wirklich hilft".

Existieren keine Kenntnisse über die Zielgruppe, ist es aufgrund der durch empirische Studien regelmäßig reproduzierten Erkenntnisse zwingend, verschiedene *Informationssuchstrategien* anzubieten, um ein hilfreiches Werkzeug zur *Informationsrecherche* bereitstellen zu können.

Die Auswirkungen der Nichtbeachtung der Studienlage sollen anhand eines fiktiven Beispiels für den Datenatlas illustriert werden.

> Würde der Datenatlas beispielsweise ausschließlich die *Gerichtete Suche* implementieren, hieße dies, dass vor allem Nutzende mit einer

genauen Kenntnis ihres Rechercheziels effektiv mit dem Werkzeug arbeiten könnten.

Im Extremfall bedeutet dies, dass einzig der *Known-Item-Search*-Use-Case effektiv durch den DATENATLAS unterstützt würde.

Hierbei wäre zu prüfen, ob sich dies noch mit den eingangs skizzierten *minimalen Use Cases der Bundesverwaltung*[68] deckt oder ob in diesem Fall nicht davon auszugehen ist, dass Verwaltungsmitarbeitende, welche bereits Kenntnis über einen konkreten Datensatz haben, diesen nicht schon auf ihrem eigenen Abteilungslaufwerk vorliegen haben, was den Nutzen des DATENATLAS *grundsätzlich* in Frage stellen würde.

[68] Siehe Kapitel 2.

Der empirischen Evidenz folgend, ist es in jedem Fall angeraten, zumindest *Facettierte Navigation* zu nutzen, um eine gute *User Experience* der *Informationsrecherche* sicherstellen zu können. Der Vorteil der facettierten Suche gegenüber anderen explorativen *Informationssuchstrategien*, wie dem *Browsing*, liegt darin begründet, dass Nutzende auf diesem Weg Ordnungskriterien des Informationsraums kennenlernen. Diese Kenntnisse können bei nachfolgenden Suchen dazu genutzt werden, um *Gerichtete Suchen* durchführen zu können.

Zur konkreten Ausgestaltung bzw. der Unterstützung gerichteter Suchen durch die *Facettierte Navigation* existieren eine Vielzahl an Arbeiten wie die von Russell-Rose und Tate (2013); Morville und Callender (2010); Hearst (2009); White und Roth (2009) oder Taylor (2006).

3.6 Datenqualität und -semantik

Eine wesentliche Herausforderung beim *Datenmanagement* ist die Sicherstellung einer für die konkreten Einsatzzwecke der Anwendung ausreichende *Datenqualität*. Diesem Aspekt widmet sich der nächste Abschnitt.

Der darauf folgende Abschnitt adressiert die semantische Anreicherung bzw. Kontextualisierung von *Daten* und die Bedeutung von *Linked (Open) Data* bzw. dem *Semantic Web*.

Dieser Teil des Gutachtens schließt mit einer ganzheitlichen Betrachtung des Themas *Datenqualität*.

Datenqualität

Die IT-gestützte Datenverarbeitung folgt grundsätzlich dem *EVA-Prinzip*[69], d.h. die Ausgabe hängt maßgeblich von der Eingabe ab. Dieses Prinzip wird in der Informatik passenderweise auch als *„garbage in – garbage out"* bezeichnet und drückt aus, dass bei einer mangelhaften *Datenqualität* der Eingabedaten auch nur eine mangelhafte Ausgabe erzeugt wird, welche dann wiederum als Grundlage einer Daten-getriebenen (Fehl-)Entscheidung dient. Natürlich hat diese Aussage auch für Anwendungen der *Künstlichen Intelligenz* (siehe Abschnitt 5.6) Bestand.

[69] *EVA-Prinzip*; Eingabe – Verarbeitung – Ausgabe.

Das Gebiet der Datenqualität wird deshalb folgerichtig seit langer Zeit beforscht. Ausgehend von frühen Studien bezüglich der Bedeutung von Datenqualität, wie z.B. der von Wang und Strong (1996), hat die Dama[70] verschiedene Dimensionen von Datenqualität ausgemacht[71].

Auch im nationalen Kontext existieren vergleichbare Arbeiten, welche teils überlappende, teils gleiche Datenqualitätsdimensionen benennen[72].

Abbildung 3.17 stellt die aktuell durch die Dama ermittelten Einflussdimensionen auf die *Datenqualität* dar. Hell gefärbte Zellen stellen die allgemeine Kombination von Datenqualitätsdimensionen und Datenkonzepten dar. Die 12 dunkel gefärbten Zellen zeigen die am häufigsten verwendeten Kombinationen von Datenqualitätsdimensionen und Datenkonzepten auf.

Die Anzahl der von Organisationen genutzten Datenqualitätsdimensionen kann je nach Organisation variieren und liegt typischerweise im Bereich von 5 bis 20.

Häufig werden die folgenden sechs *grundlegenden* Dimensionen bewertet, um ein Lagebild der in einer Organisation vorherrschenden *Datenqualität* zu ermitteln:

1. *Genauigkeit (Accuracy)*, d.h., wie nah die Daten am tatsächlichen oder anerkannten Wert liegen;
2. *Vollständigkeit (Completeness)*, d.h., ob alle erforderlichen Daten vorhanden sind;
3. *Konsistenz (Consistency)*, d.h., ob die Daten in sich stimmig sind und mit anderen verwandten Daten übereinstimmen;
4. *Eindeutigkeit (Uniqueness/Uniqueness/Deduplication)*, d.h., ob doppelte Datensätze vorhanden sind;
5. *Aktualität (Timeliness)*, d.h., wie aktuell die Daten sind;
6. *Gültigkeit (Validity)*, d.h., ob die Daten den definierten Regeln und Formaten entsprechen.

Das Thema *Datenqualität* wird auch auf europäischer Ebene betrachtet. So listet der Open Data Support (2013) der Europäischen Kommission weitere Dimensionen mit Hinblick auf die Wiederverwendbarkeit von *Open Data* auf. Auch das Kompetenzzentrum Open Data (2023) weist auf die Bedeutung einer hohen *Datenqualität* von Metadaten hin und stellt Vorteile wie die bessere Auffindbarkeit (z.B. durch eine *Gerichtete Suche*) oder die Nutzbarkeit einzelner Datensätze heraus.

Die Gültigkeit von Daten lässt sich einfach mithilfe der sogenannten *Schema-Validierung* prüfen. Unter einem *Schema* (oder Datenschema) versteht man innerhalb der Informatik die formale Beschreibung einer Datenstruktur, die i.d.R. maschinell auswertbar ist, um z.B. die *Gültigkeit* oder weitere Merkmale von Daten automatisiert überprüfen zu können.

[70] *Data Management Association*.

[71] https://dama-nl.org/dimensions-of-data-quality-en/; Letzter Abruf: 24.07.2025

[72] Lina Bruns, Benjamin Dittwald, und Fritz Meiners. *Leitfaden für Qualitativ Hochwertige Daten und Metadaten*, 2019. https://www.fokus.fraunhofer.de/content/dam/fokus/dokumente/dps/flyer/NQDM_Leitfaden_2019.pdf. Letzter Abruf: 01.08.2025

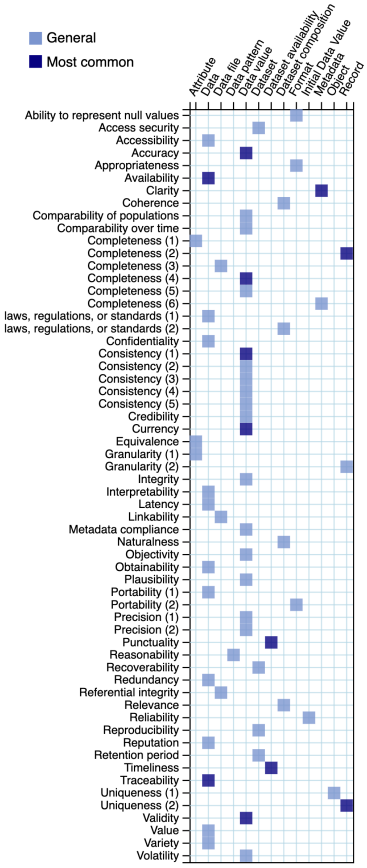


Abbildung 3.17: Dimensionen der Datenqualität; CC BY DAMA https://kamal-ahmed.github.io/DQ-Dimensions.github.io/; Letzter Abruf: 24.07.2025

Der IT-Planungsrat hat für Verwaltungsdaten, die als *Open Data* publiziert werden, DCAT-AP als verbindliches *Metadaten-Modell* (oder *Schema*) beschlossen[73].

DCAT-AP ist frei im Internet[74] verfügbar. Das *Application Profile (AP)* als Teil von DCAT-AP wurde auf Initiative der Europäische Kommission hin entwickelt und erweitert DCAT für verschiedene Anwendungsszenarien.

Außerdem stellt DCAT-AP eine Reihe *kontrollierter Vokabulare*[75] bereit[76]. Die Verwendung *kontrollierter Vokabulare/Thesauri* dient einerseits dazu die *Informationsrecherche* – insbesondere die *Gerichtete Suche*[77] – zu vereinfachen und andererseits *Datenqualität* zumindest bezüglich der folgenden *Datenqualitätsdimensionen* sicherzustellen:

- „Genauigkeit“ (*Datenqualitätsdimension*) ,
- „Vollständigkeit“ (*Datenqualitätsdimension*) ,
- „Konsistenz“ (*Datenqualitätsdimension*) ,
- „Gültigkeit“ (*Datenqualitätsdimension*) .

Die Nutzung *kontrollierter Vokabulare* ist als *Stand der Technik* zu betrachten und aufgrund der oben genannten Vorteile weit verbreitet. Ihre Verwendung wird auch durch internationale Normen, wie ISO 25964-1 (2011), empfohlen.

Sie werden ebenfalls durch die Öffentliche Verwaltung, z.B. in Form der ESD-Standards[78], welche auch das *Integrated Public Sector Vocabulary*[79] enthalten, genutzt und definiert.

Eine automatisierte Validierung von DCAT-AP-Datensätzen ist mittels Turtle-Validation für RDF (siehe nächster Abschnitt „Datensemantik und -verfügbarkeit“) in Turtle[80]-Syntax möglich[81].

Die *Schema-Validierung* von XML[82]-Daten mittels Dokumenttypdefinitionen (DTD) ist ebenfalls üblich.

Vergleichbare Mechanismen sind auch in gängigen *Datenbankmanagement-Systemen* oder *Information-Retrieval-Systemen* wie Elasticsearch oder Solr implementiert.

Verwaltungsseitig ist die Bedeutung der Sicherstellung von ausreichender *Datenqualität* zur verlässlichen Herbeiführung Datengetriebener Entscheidungen seit langem bekannt. So stellt z.B. die Europäische Kommission einen Online-Dienst zur Prüfung von DCAT-AP bereit[84].

Die automatisierte Validierung und Plausibilitätsprüfung erfolgt ebenfalls beim Datenaustausch zwischen Landes- und Bundebehörden, um die *Datenqualität* zu sichern. Beispielhaft sei hier die AVV DatA (2010) genannt, welche in §5(2) konkrete Maßnahmen wie Verzeichnisse, Kodierkataloge (im Sinne von *kontrollierten Vokabularen*) und Prüfprogramme sowie deren Anwendung (§6) beschreibt.

Beim genannten Beispiel lassen sich diese Datenqualitätssicherungsmaßnahmen bereits bis ins Jahr 1998 zurückverfolgen[85].

[73] IT-Planungsrat. *Ergebnisprotokoll der 26. Sitzung des IT-Planungsrats (28.06.2018)*, 2018. https://t1p.de/95bke. Letzter Abruf: 20.07.2025

[74] https://www.dcat-ap.de/def/dcatde/3.0/spec/; Letzter Abruf: 21.07.2025

[75] *Kontrolliertes Vokabular*; ein solches beinhaltet eine Menge von Bezeichnern oder Wörtern, deren *Semantik* und *Syntax* klar definiert ist. *Thesauri* fallen ebenfalls in diese Kategorie.

[76] https://t1p.de/25hut; Letzter Abruf: 28.07.2025

[77] Siehe Abschnitt 3.5.

[78] https://interoperable-europe.ec.europa.eu/collection/esd-standards; Letzter Abruf: 28.07.2025

[79] https://t1p.de/kb13h; Letzter Abruf: 28.07.2025

[80] *Terse RDF Triple Language*; Knappe RDF-Tripel-Repräsentation, eine Kurzform für die Speicherung von RDF-Daten (s.u.).

[81] https://t1p.de/rlbcu; Letzter Abruf: 28.07.2025

[82] *Extensible Markup Language*; erweiterbare Auszeichnungssprache, ein W3C-Standard[83] für den plattformunabhängigen Austausch von menschen- als auch maschinenlesbaren Daten.

[83] *W3C*; das *World Wide Web Consortium* setzt die De-Facto-Standards des World Wide Webs, z.B. für die Gestaltung von Web-Seiten mittels HTML oder im Rahmen der Barrierefreiheit etc..

[84] https://www.itb.ec.europa.eu/shacl/dcat-ap.de/upload; Letzter Abruf: 27.07.2025

[85] Bundesrat (1998)

Abschliessend kann festgestellt werden, dass für die automatisierte Überprüfung von *Daten* die *Maschinenlesbarkeit* dieser und ihrer *Metadaten* essenziell, weitverbreitet und damit als *Stand der Technik* zu betrachten ist.

Kritisch mit Bezug auf die *Datenqualität* zu sehen ist, auch wenn §12a(8) EGovG (2024) nur auf *Open Data* Anwendung findet, dass dieser Paragraph Bundesbehörden davon freistellt, konkrete Maßnahmen zur Datenqualitätsicherung offener Daten zu ergreifen (siehe Abschnitte 3.7 und 5.4).

Datensemantik und -verfügbarkeit

Die bisherige Betrachtung von Daten erfolgte weitestgehend losgelöst von ihrem Kontext. Auch die bereits aus Abschnitt 2.2 bekannte Abb. 3.18 ordnete Daten genau der syntaktischen Ebene zu – mit der Konsequenz, dass Daten an sich keinerlei Bedeutung haben.

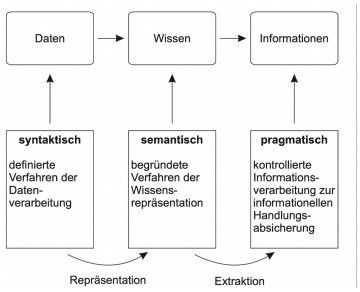


Abbildung 3.18: Abgrenzung der Begriffe Daten, Wissen und Information; Abb. 1.6 (Henrich, 2008)

Die Annahme, dass Daten ohne (semantischen) Kontext entstehen oder genutzt werden, lässt sich in der Realität jedoch nicht aufrechterhalten.

Weiter oben wurde am Beispiel der AVV DatA (2010) dargelegt, wie Länder- und Bundesbehörden seit langem Daten miteinander austauschen. Im dort geregelten Fall der Lebensmittelüberwachung erfassen die Länder Überwachungsdaten und übermitteln diese an das Bundesamt für Verbraucherschutz und Lebensmittelsicherheit (BVL), welches die Daten dann aggregiert bzw. kontextualisiert, um beispielsweise Distributionswege von Lebensmitteln über die Grenzen der Bundesländer hinweg nachvollziehen zu können.

Würden die zugrundeliegenden Datensätze ohne ihren Entstehungskontext betrachtet, wäre dies nicht ohne weiteres möglich.

Wie bereits erwähnt wurde, definiert die AVV DatA (2010) diverse Maßnahmen, um Daten einerseits maschinell validieren als auch bewerten zu können.

Die Bedeutung der *Maschinenlesbarkeit* von Daten besteht hier folglich in der Sicherstellung von *Datenqualität* als auch in der gesicherten *Ableitung von Wissen* (d.h. Bedeutung) aus den Meldedaten, das im Falle einer Lebensmittelkrise durch Menschen im BVL interpretiert werden würde.

Datensemantik – Semantic Web Die Idee mittels maschinellem logischen Schließen aus großen, vorliegenden Datenmengen Wissen mittels einer *Inferenz* abzuleiten, hat in der Informatik eine lange Tradition und reicht in Form von logischen Programmiersprachen wie *Prolog* bis in die Anfänge der 1970er-Jahre zurück[86].

Um die Jahrtausendwende erlebte diese Idee eine Renaissance in Form der evolutionären Weiterentwicklung des *World Wide Web*[87]: dem *Semantic Web*[88].

[86] Colmerauer und Roussel (1993)

[87] *World Wide Web*; das Web, welches aus multimedialen Hypertext-Dokumenten auf Grundlage von HTML (Hypertext Markup Language), HTTP (Hypertext Transfer Protocol) etc. entsteht.

[88] Berners-Lee et al. (2001)

Grob gesprochen grenzt sich das *Semantic Web* vom *World Wide Web* dadurch ab, dass es statt einfacher Hyperlinks zwischen WWW-Dokumenten für den Konsum durch Menschen, alle semantisch bedeutsamen Objekte mittels strukturierter *maschinenlesbarer* Auszeichnungen miteinander verknüpft werden. Das *W3C* stellt hier den Begriff „Web of data“ dem des „Web of documents“ gegenüber[89].

[89] https://www.w3.org/2001/sw/wiki/Main_Page; Letzter Abruf: 01.08.2025

Durch die *Maschinenlesbarkeit* wird es überhaupt erst möglich, dass Computer auf Grundlage der Daten des *Semantic Web* Aussagen automatisiert herleiten können.

So lässt sich beispielsweise über einen *eindeutigen, persistenten Identifier* (s.u.) des Autors dieses Gutachtens, seine *ORCID-iD*[90], auf dessen Dienststelle und deren Vorgängerorganisationen schließen, wie durch Abb. 3.19 illustriert wird.

[90] https://orcid.org/0000-0002-0403-457X; Letzter Abruf: 21.08.2025

ISNI:	0000 0000 9992 844X https://isni.org/isni/**000000009992844X**
Name:	Berlin (Germany : West) Fachhochschule für Verwaltung und Rechtspflege Berlin School of Economics and Law F.H.S.V.R. Fachhochschule für Verwaltung und Rechtspflege Berlin Fachhochschule fur Wirtschaft Berlin FHSVR FHVR FHVR Berlin FHW Berlin Hochschule für Wirtschaft und Recht Berlin Hochschule fur Wirtschaft und Recht Berlin - Campus Schoneberg HWR Berlin University of Applied Sciences Berlin
Location / Nationality:	Germany Germany Berlin Berlin

Abbildung 3.19: Beispieldatensatz *International Standard Name Identifier* (ISNI) der HWR Berlin; https://isni.org/isni/000000009992844X; Letzter Abruf: 01.08.2025

Zur Realisierung des *Semantic Web* wird auf eine Vielzahl von offenen Technologien gesetzt, welche teilweise älter sind, als die Idee des *Semantic Webs* selbst. Abbildung 3.20 bildet einige, aber bei weitem nicht alle, mit dem *Semantic Web* verbundenen Technologien ab. Typischerweise nutzen auch *Linked (Open) Data*-Anwendungen diese und weitere Technologien.

Dies ist beispielsweise bei GovData der Fall, wie in Abschnitt 5.4 detailliert dargestellt wird.

Semantic Web vs. Linked (Open) Data

Im Rahmen dieses Gutachtens können die Begriffe *Linked (Open) Data* und *Semantic Web* als mehr oder weniger synonym betrachtet werden.

Eine Technologie von zentraler Bedeutung ist der W3C-Standard RDF[91] für die Beschreibung von Ressourcen in Tripel-Form. RDF zielte ursprünglich ausschließlich auf die Beschreibung von Metadaten und ihren Relationen untereinander ab, ist aber nun zentraler Teil bei Anwendungen des *Semantic Web* wie z.B. *Linked (Open) Data*.

[91] *Resource Description Framework*; Beschreibungssystem für Ressourcen.

Oberflächlich gesagt definiert RDF ein Datenmodell, welches Aussagen über Ressourcen auf Grundlage sogenannter gerichteter Graphen erlaubt und die als *Tripel* bezeichnet werden. Die Tripel sind dabei immer gleich aufgebaut und bestehen aus einem *Sub-*

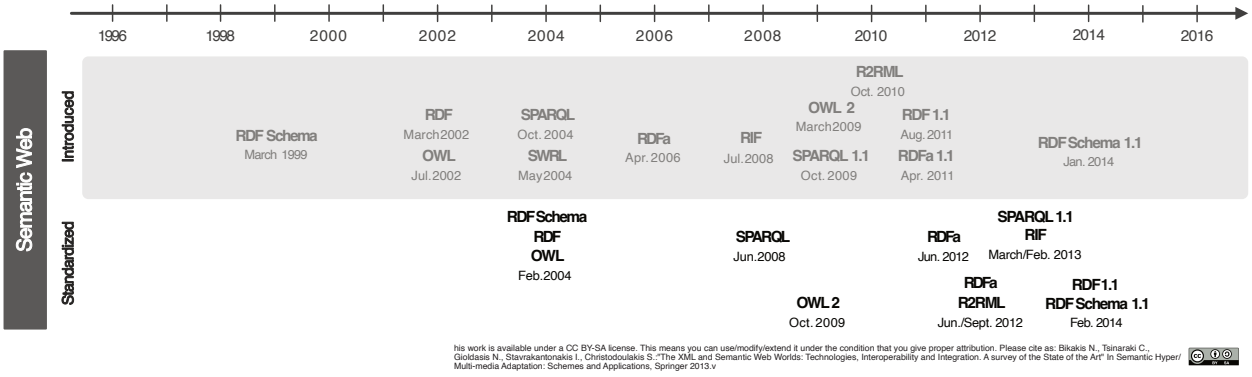


Abbildung 3.20: Zeitstrahl *Semantic Web*-Technologien

jekt, einem *Prädikat* und einem *Objekt*. Hierdurch ist eine formale Semantik gegeben, welche sich maschinell auswerten lässt.

Abbildung 3.21 stellt ein solches Tripel dar. Der untere Teil der Abbildung visualisiert die konkrete Aussage „David Zellhöfer ist Autor des vorliegenden Gutachtens“ als RDF-Tripel.

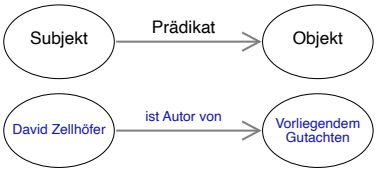


Abbildung 3.21: Aufbau eines RDF-Tripels und konkretes Beispiel (u.)

Die Ressource „David Zellhöfer“ könnte wiederum, dem oben vorgestellten Beispiel folgend, über ein Prädikat „arbeitet für“ mit der ISNI der HWR Berlin (siehe Abb. 3.19) verbunden werden, so dass sich ein Wissensgraph[92] – oder *Semantic Web* – bildet, welches z.B. Ontologien oder sonstige Beziehungen verschiedenster Ressourcen untereinander enthalten kann, die sich maschinell auswerten lassen.

Ein komplexeres Beispiel auf Grundlage der über das Bundesministerium des Innern verfügbaren Daten findet sich in Anhang A.6.

Die entstehende „Datensemantik“ besteht folglich aus maschinell auswertbaren Aussagen auf Grundlage dieses „Web of data“, wie: in welcher Stadt die Hochschule des Autors liegt oder welche sonstigen Beschäftigten für diese tätig sind.

Diese semantische Auswertbarkeit lässt sich nur schwer mit anderen Datenmanagementtechnologien realisieren.

Um solche Graphendaten effizient speichern und abrufen zu können, werden in der Regel spezialisierte *Triplestores*[93] oder Graphendatenbanken genutzt.

[92] *Knowledge Graph*; eine Graph-basierte Darstellung von semantischen Zusammenhängen, die maschinell auswertbar ist. Ein bekanntes Beispiel für einen *Knowledge Graph* ist Wikidata (`https://www.wikidata.org`).

[93] *Triplestore*; ein spezielles Datenbanksystem zur optimierten Verarbeitung von RDF-Tripeln.

Datenverfügbarkeit – Persistente Identifier Um Aussagen über Ressourcen treffen zu können, wurden bereits im einführenden Beispiel sogenannte persistente Identifier oder *Persistent Identifier (*PID*)* genutzt: die ORCID-iD des Autors und der ISNI seiner Hochschule.

Ein PID bezeichnet einen eindeutigen Namen, welcher der Identifikation einer Ressource oder eines Objekts, z.B. eines wissenschaftlichen Artikels im Internet oder einer Person, dient. Ein PID ist dauerhaft gültig und ermöglicht deshalb die permanente Identifikation und Auffindbarkeit eines Objekts und kann damit in langlebigen RDF-Aussagen genutzt werden.

In der Praxis verwendete PID folgen zumeist der URI-Syntax[94].

[94] *Uniform Resource Identifier*; ein eindeutiger Identifier, der auf abstrakte (z.B. eine Webseite) oder physische Ressourcen verweist und in einer definierten Syntax vorliegt (Berners-Lee, 1994).

Die Nutzung von PID ist nicht nur in der Wissenschaft in Form von DOI[95] üblich. So können ISNI (s.o.) dazu verwendet werden, schöpferisch tätige Personen oder Köperschaften eindeutig zu benennen. Bis zu einem gewissen Grad können auch ISBN-Nummern oder EAN-Codes, die von den Barcodes im Einzelhandel bekannt sind, als PID interpretiert werden.

Das PID Network Deutschland[96] listet weitere verschiedene Anwendungsfälle auf, wobei sich insbesondere der Use Case „Forschungsdaten“[97] auf den Einsatzzweck des Datenatlas übertragen lässt.

Neben der dauerhaften Adressierbarkeit mittels PID ist es natürlich notwendig, dass die identifizierten Ressourcen ebenfalls physisch oder digital verfügbar bleiben. Hierzu bieten die verschiedenen PID-Anbieter unterschiedliche Ansätze, die jedoch außerhalb des Fokus dieses Gutachtens liegen.

Bedeutung für die Öffentliche Verwaltung Es liegt auf der Hand, dass in der Öffentlichen Verwaltung eine Vielzahl an Daten anfällt, welche über ein hohes Maß an Binnenstruktur oder Kontext verfügen – seien es Daten auf Landes- bzw. Bundesebene oder in der Kameralistik, die sich u.a. in Haushaltskapitel und Haushaltstitel untergliedert.

Durch die Nutzung von *Linked (Open) Data*-Technologien lassen sich diese Arten von komplexen Datenzusammenhängen besser recherchieren und visualisieren, wie diverse Implementierungen und Proof of Concepts in der Öffentlichen Verwaltung zeigen.

So veröffentlichen die Länder Berlin und Schleswig-Holstein ihre Haushaltsdaten mittlerweile als *Linked (Open) Data*[98].

Eine andere Fallstudie zeigt die Nachnutzungspotenziale maschinenlesbarer Organigramme der Berliner Verwaltung – gerade im Hinblick auf KI-Anwendungsszenarien – auf[99].

Insgesamt 35 Fallstudien werden durch das *W3C* präsentiert, darunter neun mit direktem ÖV/eGovernment-Bezug, welche weitere Einsatzmöglichkeiten von Technologien aus dem Gebiet *Linked (Open) Data/Semantic Web* aufzeigen[100].

Die Zivilgesellschaft bietet mit der *DBpedia* (seit 2007)[101] und der Wikidata (seit 2012)[102] bereits umfangreiche *Knowledge Graphs* auf Grundlage der vorgestellten *Linked (Open) Data*-Technologien an.

Kommerzielle Suchmaschinen wie Google nutzen seit 2012 vergleichbare Ansätze, um die Qualität ihrer Ergebnisse zu erhöhen[103].

Beim bereits auf Seite 54 vorgestellten DCAT-AP, welches beim Austausch von *Open Data* der Öffentlichen Verwaltung verbindlich ist, handelt es sich ebenfalls um ein RDF-Vokabular[104]. GovData stellt außerdem ein URI-Konzept zur Adressierung von

[95] *Digital Object Identifier*; ein persistenter Identifier für i.d.R. Publikationen, der vor allem in der Wissenschaft verbreitet und akzeptiert ist, wie z.B. das Literaturverzeichnis dieses Gutachtens deutlich macht.

[96] Das PID-Kompetenzzentrum in Trägerschaft des Tib – Leibniz-Informationszentrum Technik und Naturwissenschaften.

[97] https://www.pid-network.de/pids/forschungsdaten; Letzter Abruf: 27.07.2025

[98] Siehe https://t1p.de/oo44i; Letzter Abruf: 28.07.2025.

[99] Lisa Stubert, Klemens Maget, Max Bruno Eckert, und Hans Hack. *Linked Open Data in der Praxis – Vernetzte Verwaltungsdaten am Beispiel der Berliner Organigramme*, 2025. https://t1p.de/p9n6b. Letzter Abruf: 27.07.2025

[100] https://www.w3.org/2001/sw/sweo/public/UseCases/; Letzter Abruf: 01.08.2025

[101] https://www.dbpedia.org

[102] https://www.wikidata.org

[103] https://blog.google/products/search/introducing-knowledge-graph-things-not/; Letzter Abruf: 01.08.2025

[104] https://www.govdata.de/informationen/metadatenschema; Letzter Abruf: 21.07.2025

Ressourcen bereit[105].

[105] https://www.dcat-ap.de/def/uriConcept/1.0.pdf; Letzter Abruf: 21.07.2025

Ganzheitliche Betrachtung von Datenqualität

Auch wenn viele der oben genannten Ansätze auf die Veröffentlichung von Daten im Sinne von *Open Data* abzielen und damit nicht mit dem Einsatzzweck des DATENATLAS übereinstimmen, so ist es wichtig anzumerken, dass Daten üblicherweise in einem Kontinuum existieren.

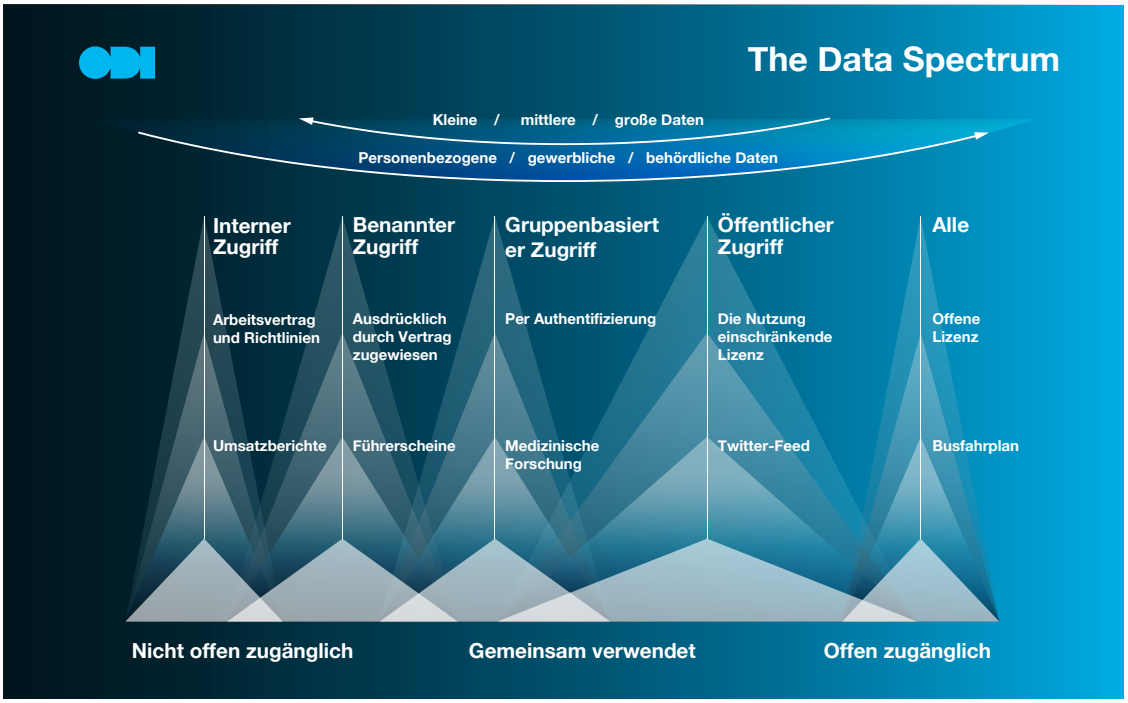


Abbildung 3.22: Datenspektrum nach dem ODI; CC BY Open Data Institute, https://theodi.org/insights/tools/the-data-spectrum/; Letzter Abruf: 01.08.2025

Abbildung 3.22 illustriert, dass auch verwaltungsinterne Daten u.U. einen gewissen Grad an „Offenheit“ besitzen, wenn z.B. ein gruppenbasierter Zugriff wie beim DATENATLAS angestrebt wird. Perspektivisch ist zudem nicht auszuschließen, dass aktuell noch verwaltungsinterne Daten zukünftig als *Open Data* veröffentlicht werden.

Technisch gesehen profitieren jedoch auch rein interne Daten durch die bereits genannten Vorteile, die sich aus der Nutzung von *Linked (Open) Data*-Technologien ergeben.

Gerade hierarchische Organisationen wie die ÖFFENTLICHE VERWALTUNG sind eigentlich dafür prädestiniert, *Ontologien*, *kontrollierte Vokabulare* und ähnliche Techniken einzusetzen, um ihre Daten besser zu kontextualisieren und maschinenlesbar aufzubereiten. Auch die Nutzung von PID sollte für öffentliche Einrichtungen keine Hürde darstellen, sondern vielmehr in ihrem eigenen Interesse liegen, um beispielsweise Vorgänge weiterverfolgen zu können, selbst wenn z.B. eine Zuständigkeit zwischen verschiedenen Ministerialbehörden wechselt.

Letztendlich ermöglichen solche Maßnahmen auf Basis der Datenmodellierung überhaupt erst eine Indikatoren-basierte Steue-

rung der Verwaltung und vereinfacht die Umsetzung von Berichtspflichten, da nicht erst einzelne Datensätze vereinheitlicht und manuell korrigiert werden müssen, wie es heute oft der Fall ist.

Inwiefern die Art der Datenbereitstellung und -modellierung, welche ohne Zweifel einen Einfluss auf die Nutzungsqualität von Daten hat, zu bewerten ist, zeigt Abb. 3.23 auf. Wenngleich dieses seit 2012 bestehende 5-Sterne-Bewertungsschema eigentlich dem Bereich *Open Data* entlehnt ist, so zeigt es doch deutlich, welche Qualitätsstufe auch interne Verwaltungsdaten erreichen sollten.

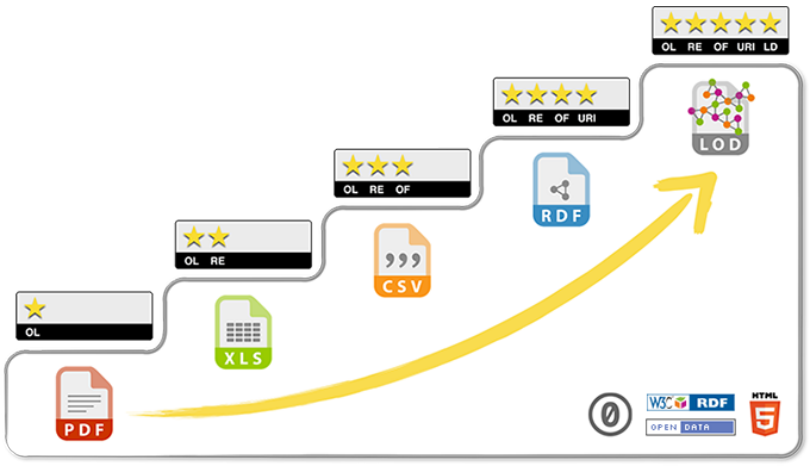


Abbildung 3.23: 5-Sterne-Modell für Offene Daten

Lässt man Stufe 1 außen vor, die sich nur auf die Veröffentlichung von Daten im Internet bezieht, bieten die anderen vier Stufen einen guten Ansatzpunkt, um die *Datenqualität* von verwaltungsinternen Daten ganzheitlich zu bewerten. Dies beinhaltet Aspekte wie die digitale Souveränität oder die Nachnutzbarkeit von Daten. Tabelle 3.1 fasst die Bedeutung der einzelnen Qualitätsstufen, die nicht zwangsläufig der Reihe nach durchlaufen müssen, kurz zusammen.

Stufe	Aufforderung an Datenerstellende
★	stelle deine Daten im Web unter einer offenen Lizenz bereit. Das Format ist dabei egal
★★	stelle Daten in einem strukturierten Format bereit (z. B. Excel anstelle eines eingescannten Bildes einer Tabelle)
★★★	verwende offene, nicht proprietäre Formate (z. B. CSV statt Excel)
★★★★	verwende URIs um Dinge zu bezeichnen, damit deine Daten verlinkt werden können
★★★★★	verlinke deine Daten mit anderen Daten um Kontexte herzustellen

Tabelle 3.1: 5-Sterne-Modell für Offene Daten – Stufenüberblick `https://5stardata.info/de/`; Letzter Abruf: 21.07.2025

Eine detaillierte Gegenüberstellung von Kosten und Nutzen der einzelnen Stufen findet sich unter `https://5stardata.info/de/`,

soll hier jedoch aus Platzgründen ausgespart werden.

3.7 Regulatorische Grundlagen

Wie in Abschnitt 3.1 erwähnt, werden in diesem Abschnitt verpflichtende Rechtsnormen für die BUNDESVERWALTUNG zur Barrierefreiheit (wie die BITV 2.0 (2023)) oder der IT-Sicherheit aus Platz- und Recherchegründen[106] ausgespart.

Der Fokus dieses Abschnittes liegt ausschließlich auf den Rechtsnormen, welche sich mit der Verwaltungsdigitalisierung bzw. dem E-Government sowie der Beschaffung und Bewertung von Leistungen in diesem Bereich befassen.

Die BUNDESVERWALTUNG muss generell die Grundsätze der Aktenführung beachten, auch wenn rein digital gearbeitet wird[107]. Dieser Aspekt wird in Abschnitt 5.7 vertieft.

[106] Aufgrund der kursorischen Sichtprüfung des DATENATLAS können hier außerdem keine belastbaren Aussagen getroffen werden, siehe Kapitel 4.

[107] Deutscher Bundestag - Wissenschaftliche Dienste. *Grundsätze der Aktenführung in der Bundesverwaltung*, 2023. https://tinyurl.com/veraktung. Letzter Abruf: 26.07.2025

Wirtschaftlichkeit

Grundsätzlich besteht in der BUNDESVERWALTUNG das *Wirtschaftlichkeitsgebot*, welches sich aus §7 BHO (2024) ergibt und sogenannte *„Wirtschaftlichkeitsuntersuchungen"* fordert, die im Alltagsgebrauch auch als *Wirtschaftlichkeitsbetrachtungen (WiBe)* bezeichnet werden.

Die VV-BHO (2025) präzisiert dabei in §7 (2.1) *Wirtschaftlichkeitsuntersuchungen als Planungsinstrument*, dass diese „mindestens Aussagen zu folgenden Teilaspekten enthalten:"

> Analyse der Ausgangslage und des Handlungsbedarfs,
> - Ziele, Prioritätsvorstellungen und mögliche Zielkonflikte,
> - relevante Lösungsmöglichkeiten und methodenabhängig die damit verbundenen Einnahmen und Ausgaben bzw. deren Nutzen und Kosten (einschl. Folgekosten), auch soweit sie nicht in Geld auszudrücken sind,
> - finanzielle Auswirkungen auf den Haushalt,
> - Eignung der einzelnen Lösungsmöglichkeiten zur Erreichung der Ziele unter Einbeziehung der rechtlichen, organisatorischen und personellen Rahmenbedingungen unter Berücksichtigung der Risiken und der Risikoverteilung,
> - Zeitplan für die Durchführung der finanzwirksamen Maßnahme,
> - Kriterien und Verfahren für Erfolgskontrollen (vgl. Nr. 2.2).

Bezüglich der Erfolgskontrolle nach Nr. 2.2 (s.o.) legt die VV-BHO (2025) den folgenden Mindestumfang fest:

> *Zielerreichungskontrolle* Mit der Zielerreichungskontrolle wird durch einen Vergleich der geplanten Ziele mit der tatsächlich erreichten Zielrealisierung (Soll-Ist-Vergleich) festgestellt, welcher Zielerreichungsgrad zum Zeitpunkt der Erfolgskontrolle gegeben ist. Sie bildet gleichzeitig den Ausgangspunkt von Überlegungen, ob die vorgegebenen Ziele nach wie vor Bestand haben.
>
> *Wirkungskontrolle* Im Wege der Wirkungskontrolle wird ermittelt, ob die finanzwirksame Maßnahme für die Zielerreichung geeignet und ursächlich war. Hierbei sind alle beabsichtigten und

unbeabsichtigten Auswirkungen der durchgeführten finanzwirksamen Maßnahme zu ermitteln.,

Wirtschaftlichkeitskontrolle Mit der Wirtschaftlichkeitskontrolle wird untersucht, ob der Vollzug der finanzwirksamen Maßnahme im Hinblick auf den Ressourcenverbrauch wirtschaftlich war (Vollzugswirtschaftlichkeit) und ob die finanzwirksame Maßnahme im Hinblick auf übergeordnete Zielsetzungen insgesamt wirtschaftlich war (Maßnahmenwirtschaftlichkeit (VV-BHO)).

Erfolgskontrollen sind auch durchzuführen, wenn die Dokumentation in der Planungsphase unzureichend war. In diesem Fall sind die benötigten Informationen nachträglich zu beschaffen.

Anforderungen an zu beschaffende Software

Der allgemeine Umgang mit digitalen Verwaltungsleistungen, deren Beschaffung und Ausgestaltung ist im OZGÄndG (2024) bzw. dessen Vorläufer OZG (2024) und im EGovG (2024) für Behörden des Bundes einschließlich der bundesunmittelbaren Körperschaften und weiterer geregelt.

Relevante Teilaspekte aus diesen Gesetzen werden im Folgenden kurz umrissen.

Bevorzugung von Open-Source-Software Das EGovG (2024) fordert für neu zu beschaffende Software-Systeme in §16a *Open Source*:

> Die Behörden des Bundes sollen offene Standards nutzen und bei neu anzuschaffender Software Open-Source-Software vorrangig vor solcher Software beschaffen, deren Quellcode nicht öffentlich zugänglich ist oder deren Lizenz die Verwendung, Weitergabe und Veränderung einschränkt.

Hierbei muss angemerkt werden, dass diese Anforderung erst nach der Vergabe des Datenatlas in das Gesetz aufgenommen wurde. Dies gilt auch für das OZGÄndG (2024).

Eine ähnliche Anforderung ergibt sich aus §4(3) *Elektronische Abwicklung von Verwaltungsverfahren; Verordnungsermächtigung* OZG (2024):

> Bei der Bereitstellung der IT-Komponenten im Sinne des Absatzes 1 sollen offene Standards und offene Schnittstellen verwendet werden und soll Open-Source-Software vorrangig vor solcher Software eingesetzt werden, deren Quellcode nicht öffentlich zugänglich ist oder deren Lizenz die Verwendung, Weitergabe und Veränderung einschränkt.

Open Data §12a *Offene Daten des Bundes* EGovG (2024) definiert den Umgang mit offenen Daten (*Open Data*). Allerdings ist dieser Paragraph nicht direkt auf den Datenatlas anwendbar, da er ausschließlich auf die Bereitstellung von Daten innerhalb der Bundesverwaltung abzielt.

Nichtsdestotrotz wird der Paragraph dadurch relevant, da er wesentliche Eigenschaften von offenen Daten benennt, die auch auf Daten zutreffen, die nur „verwaltungsintern offen“ sind – wie Daten die im Datenatlas[108].

[108] Siehe Abschnitt 3.6.

Kritisch zu sehen ist u.a. §12a(8) EGovG (2024), der die Bundesbehörden von der Ergreifung konkreter Maßnahmen zur Datenqualitätssicherung (siehe Abschnitt 3.6) der offenen Daten entpflichtet:

> Die Behörden des Bundes sind nicht verpflichtet, die bereitzustellenden Daten auf Richtigkeit, Vollständigkeit, Plausibilität oder in sonstiger Weise zu prüfen.

Nutzerfreundlichkeit und Barrierefreiheit Der Aspekt, ob digitale Angebote der Bundesverwaltung über ein gutes Maß an menschzentrierten Qualität (siehe Abschnitt 3.4) verfügen, wird in §16 *Nutzerfreundlichkeit und Barrierefreiheit* EGovG (2024) nur kurz erwähnt:

> Die Behörden des Bundes gestalten die elektronische Kommunikation und die elektronischen Dokumente nutzerfreundlich und barrierefrei. Für die barrierefreie Gestaltung gilt die Barrierefreie-Informationstechnik-Verordnung entsprechend.

Durch das OZGÄndG (2024) wurde §7 *Nutzerfreundlichkeit und Barrierefreiheit* OZG (2024) erheblich nachgeschärft:

> *(1)* Bund und Länder stellen durch geeignete Maßnahmen die Nutzerfreundlichkeit sowie eine einfache und intuitive Bedienbarkeit des übergreifenden Zugangs zu elektronischen Verwaltungsleistungen, einschließlich der für diesen Zugang relevanten IT-Komponenten, sicher. Nutzer sollen in die Entwicklung neuer elektronischer Angebote einbezogen werden.
>
> *(2)* Der übergreifende Zugang zu elektronischen Verwaltungsleistungen, einschließlich der für diesen Zugang relevanten IT-Komponenten, ist nach Maßgabe der Barrierefreie-Informationstechnik-Verordnung so zu gestalten, dass sie barrierefrei nutzbar sind.

Standardisierung Für den Anwendungsbereich des OZG (2024) werden in §6 *Standards; Verordnungsermächtigungen* weitere Vorgaben zu Architekturvorgaben, Qualitätsanforderungen und Interoperabilitätsstandards bis Ende 2026 angekündigt. Diese Standards werden elektronisch veröffentlicht werden[109].

[109] Siehe Abschnitt 5.5.

4
Exemplarische User Journeys und Einordnung

In diesem Kapitel wird der Funktionsumfang des Datenatlas anhand von User Journeys dargestellt und bewertet.

User Journeys[1] als Methode sind sowohl zur Erhebung von Ist-Zuständen als auch für die Gestaltung neuer Interaktionsabläufe geeignet[2]. Sie finden regelmäßig Verwendung im Bereich der Öffentlichen Verwaltung, u.a. durch die *DigitalService GmbH des Bundes*[3] oder das *CityLAB Berlin*[4], Berlins öffentlichem Innovationslabor.

Als Grundlage der User Journeys dienen die in der Einleitung beschriebenen, auf den Datenatlas bezogenen, *minimalen Use Cases der Bundesverwaltung*:

1. die *Recherche von Metadaten*, welche vom Großteil der Beschäftigten der Bundesverwaltung genutzt wird, und
2. die *Erfassung von Metadaten*, welche durch spezialisiertes Personal erfolgen dürfte.

Aus diesen Use Cases lassen sich zwei minimalistische *Personas* [5] konstruieren, die in Abb. 4.1 vergleichend gegenübergetellt sind.

Persona 1, *Uwe*, repräsentiert dabei einen Großteil der in der Bundesverwaltung Beschäftigten, welche ohne tiefergehende Recherche- und Datenkenntnisse den Datenatlas primär zum Recherchieren nutzen. Uwes User Journey im Rahmen der *Informationsrecherche* wird in Abschnitt 4.1 aufgezeigt.

Die andere Persona, *Elin*, verfügt als *Data Scientist* über erweiterte Recherche- und Datenkenntnisse und ein breiteres Aufgabengebiet als Uwe, weshalb in ihrem Fall zwei User Journeys vorgestellt werden, die typische Arbeitsaufgaben widerspiegeln: die Metadatenverwaltung und den Datenimport in den Datenatlas.

Die aus den User Journeys gewonnenen Erkenntnisse werden begleitend diskutiert und bewertend eingeordnet. Um die Lesbarkeit des Gutachtens zu erhöhen, sind die Bewertungen bewusst knapp gehalten, da aufgrund der limitierten Erhebungsmöglichkeit (siehe unten) nur ein Schlaglicht auf den Zustand des Datenatlas von Mitte Juli 2015 geworfen werden kann.

Die Bewertung erfolgt nach dem *Stand der Technik*, welcher in Kapitel 3 beschrieben wurde.

[1] *Nutzungsreise*; eine, i.d.R. visuelle, Abbildung der zentralen Berührungspunkte eines Akteurs mit einer Dienstleistung. Die User Journey umfasst dabei alle Phasen der User Experience, d.h., die Zeiten vor, während und nach der Nutzung. Überlicherweise werden User Journeys mit emotionalen Äußerungen der Nutzenden versehen, um gut funktionierende Prozess-Schritte von problematischen, den *painpoints*, abzugrenzen.

[2] Toni Steimle und Dieter Wallach. *Collaborative UX Design: Lean UX Und Design Thinking: Teambasierte Entwicklung Menschzentrierter Produkte*. dpunkt.verlag, 1. Auflage, 2018. ISBN 978-3-86490-532-2

[3] `https://t1p.de/dmbef`; Letzter Abruf: 01.08.2025

[4] Caroline Paulick-Thiel und Henrike Arlt. *Öffentliches Gestalten: Handbuch für innovatives Arbeiten in der Verwaltung*. Technologiestiftung Berlin, 1. Auflage, 2020. ISBN 978-3-00-065930-0. `https://t1p.de/b3ndy`

[5] Siehe Abschnitt 3.4.

Abbildung 4.1: Verwendete Personas im Vergleich

Als Bewertungsmethode der bereitgestellten Funktionen und der damit verbundenen *User Experience* wird eine *Heuristische Inspektion*[6] genutzt. Diese Evaluierungsmethode basiert auf einer Experteneinschätzung, die sich an den Golden Rules of Interface Design orientiert, welche in Abschnitt 3.4 eingeführt wurden. Für die Durchführung einer *Heuristischen Inspektion* muss keine voll funktionsfähige Software vorliegen. Der Vorteil dieser Inspektionsmethode besteht darin, dass sie jederzeit im Entwicklungsprozess, z.B. frühzeitig anhand von Mock-Ups oder im späteren Projektverlauf mittels Screenshots, wie im vorliegenden Fall, durchgeführt werden kann.

[6] Siehe Abschnitt 3.2.

Generell werden nicht immer alle Stärken und Schwächen des Datenatlas für jeden Arbeitsschritt wiedergegeben. Stattdessen werden Charakteristika herausgestellt, um den Nutzungseindruck der Anwendung kompakt darstellen zu können.

Aufgrund der Verwendung von Screenshots kann Kriterium 2 (Golden Rules of Interface Design), die *Universelle Bedienbarkeit*, nicht betrachtet werden.

Die Erhebung der für die *Heuristische Inspektion* nötigen Screenshots des User Interfaces des Datenatlas erfolgte Mitte Juli 2025 im Rahmen einer kursorischen Sichtprüfung im Umfang von ca. 30 Minuten am Produktivsystem des Datenatlas.

Insgesamt wurden 47 Screenshots analysiert, von denen im Folgenden die aussagekräftigsten in Form einzelner, zusammengehöriger Prozess-Schritte von User Journeys vorgestellt werden.

Die Datenqualität des Datenatlas wird, wenn möglich, be-

gleitend anhand der zur Einschätzung geeigneten *grundlegenden* sechs Dimensionen aus Abschnitt 3.6 bewertet. Kriterium 4 (Datenqualität), also die *Eindeutigkeit*, kann ohne Zugriff auf die Datenbank des DATENATLAS nicht begutachtet werden.

AUS DIESEN TEILASPEKTEN der Betrachtung ergibt sich ein Gesamtbild, welches Aussagen zum Servicestandard der Öffentlichen Verwaltung und zur Einhaltung der regulatorischen Grundlagen[7] durch den DATENATLAS ermöglicht. Diese eher generellen Einschätzungen werden in Kapitel 5 getroffen.

[7] Siehe Abschnitt 3.7.

4.1 *Informationsrecherche im Datenkatalog*

Uwes (siehe Abb. 4.1; links) Ziel besteht darin, im Rahmen seiner täglichen Arbeitsaufgaben, relevante Datensätze zum Thema „Förderprogramme zum Hochwasserschutz“ im DATENATLAS zu recherchieren.

Informationsrecherche I – Beginn der Recherche

Beginn der Recherche

Wichtiger Hinweis: Die Erhebung der einzelnen Prozess-Schritte erfolgte mittels eines Benutzer-Accounts, der auch zur Datenerfassung berechtigt ist. Weniger privilegierte Nutzende würden in diesem Fall eine kleinere Anzahl an Funktionsbereichen sehen.

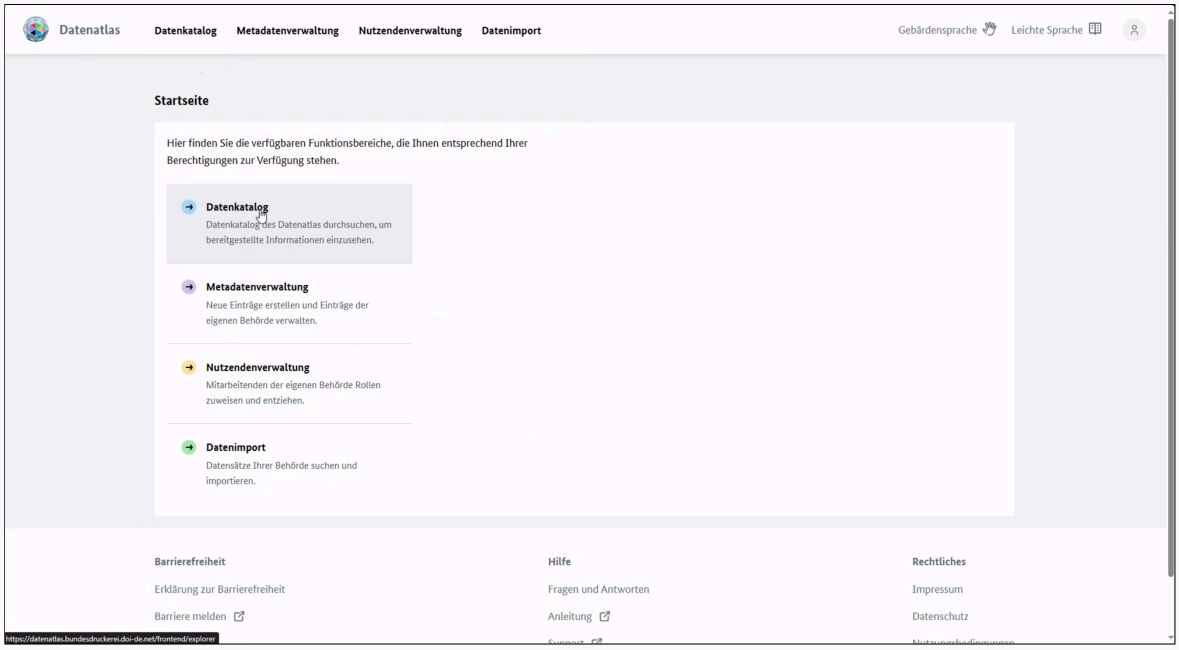


Abbildung 4.2: Einstiegsseite mit Auswahl der verschiedenen Funktionsbereiche des DATENATLAS

Uwe öffnet den Link zum DATENATLAS, um auf dessen Startseite zu gelangen, die in Abb. 4.2 zu sehen ist.

1. Stärkt „Informatives Feedback“ (*Usability*) – informativer Überblick, inkl. Hinweis auf be-

nötigte Berechtigungen; Verbesserungspotenzial: Ansprechpartner bei Problemen/Support ergänzen

2. Stärkt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – jedes Element ist erläutert
3. Stärkt „Kontrollierbarkeit“ (*Usability*) – Uwe initiiert den Einstieg in Datenkatalog
4. Stärkt „Fehlervermeidung“ (*Usability*) – man kommt nur an Systeme für die man berechtigt ist bzw. würden diese im Normalfall erst gar nicht sichtbar werden (s.o.)

Nach dem Klick auf die Auswahl „Datenkatalog“ gelangt Uwe auf die sogenannte „Metadatensuche“ (siehe Abb. 4.3).
Uwe ist überrascht. Hat er nicht gerade den „Datenkatalog“ gewählt? Er entdeckt oberhalb des Seitentitels „Metadatensuche“ jedoch schnell die Breadcrumb-Navigation („Datenatlas > Datenkatalog“), welche er von anderen Webseiten kennt, und findet sich wieder zurecht.
Trotzdem ist er sich sicher, dass er keine Suche gestartet hat und trotzdem 3.440 Suchergebnisse erhält.

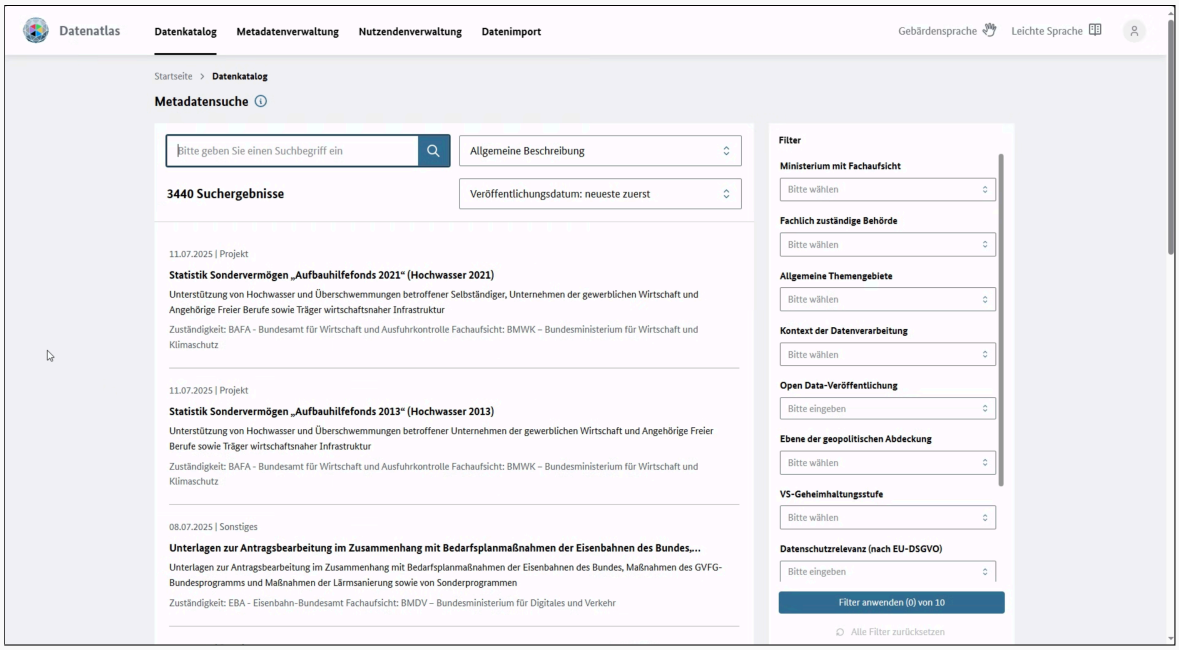


Abbildung 4.3: Einstiegsseite des Datenkatalogs

1. Entspricht dem *Stand der Technik*, weil Breadcrumb-Navigation als visuelles Entwurfsmuster genutzt wird
2. Verletzt den *Stand der Technik*, weil die Suche direkt ohne Sucheingabe gestartet wurde
3. Verletzt „Konsistenz“ (*Usability*) – Terminologie im UI wird nicht konsistent genutzt (Klick auf „Datenkatalog“ führt zur „Metadatensuche“)
4. Verletzt „Konsistenz“ (*Usability*) – üblicherweise erscheinen Suchergebnisse erst nach Eingabe eines Suchbegriffs
5. Verletzt „Informatives Feedback“ (*Usability*) – es werden direkt 3.440 Suchergebnisse angezeigt; das Suchfeld bleibt leer, so dass unklar ist, auf welcher Anfrage die Ergebnisse basieren

6. Verletzt „Abgeschlossene Aktionen“ (*Usability*) – es ist unklar, in welchem Zustand sich das System befindet: wird noch eine Handlung erwartet?
7. Verletzt „Fehlervermeidung“ (*Usability*) – da unsicher ist, wie das System in diesen Zustand kam, kann eine Fehlbedienung nicht ausgeschlossen werden
8. Verletzt „Einfache Umkehrbarkeit von Aktionen“ (*Usability*) – es ist nicht erkennbar, wie man zum definierten Zustand zurückkommt
9. Stärkt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – alle relevante Informationen, bis auf den Suchbegriff, sind im UI sichtbar
10. Verletzt „Kontrollierbarkeit“ (*Usability*) – Suche wurde direkt gestartet, ohne dass Uwe etwas getan hat

Informationsrecherche II – Schlagwort-basierte Suche

Schlagwort-basierte Suche

Auch wenn er es anders erwartet hatte, gibt Uwe den Suchbegriff „hochwasser“ in das oben platzierte Suchfeld ein und startet die Suche mit einem Klick auf die Lupe. Während er tippt, erscheinen keine Hinweise, die er zur Autovervollständigung nutzen kann.

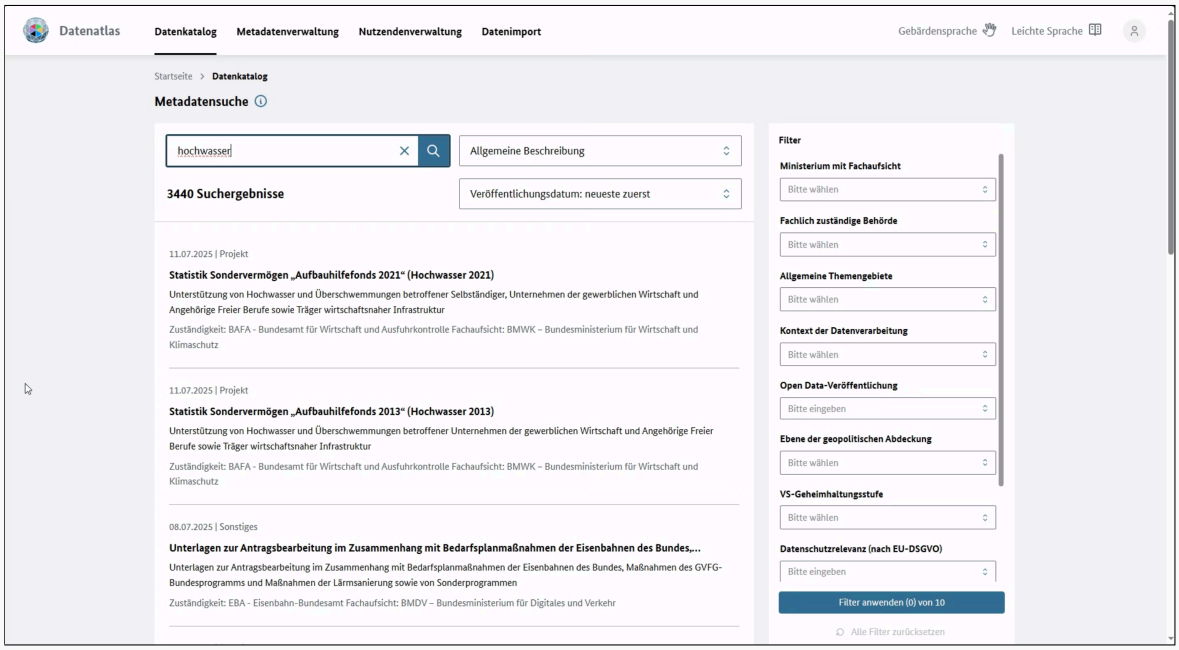


Abbildung 4.4: Datenkatalog; Eingabe Suchbegriff „hochwasser“

1. Stärkt „Konsistenz“ (*Usability*) – man kann wie gewohnt einen Suchbegriff eingeben
2. Verletzt „Konsistenz“ (*Usability*) – üblicherweise erscheinen Suchergebnisse erst nach Eingabe eines Suchbegriffs
3. Verletzt „Informatives Feedback“ (*Usability*) – direkte Anzeige, dass es 3.440 Suchergebnisse gibt, ohne dass Suche gestartet wurde
4. Verletzt den *Stand der Technik*, weil keine Autovervollständigung des Suchbegriffes o.ä. angebo-

ten wird

5. Verletzt den *Stand der Technik*, weil kein Browsing-Einstieg o.ä. verfügbar ist

Nach einem kurzen Moment erhält Uwe seine Suchergebnisse, wie in Abb. 4.5 dargestellt ist.

Da die Liste recht kurz ist, muss er sie nicht weiter verfeinern.
Die geringe Anzahl an Treffern zeigt Uwe deutlich, dass er noch einmal anders bei seiner Recherche ansetzen muss.

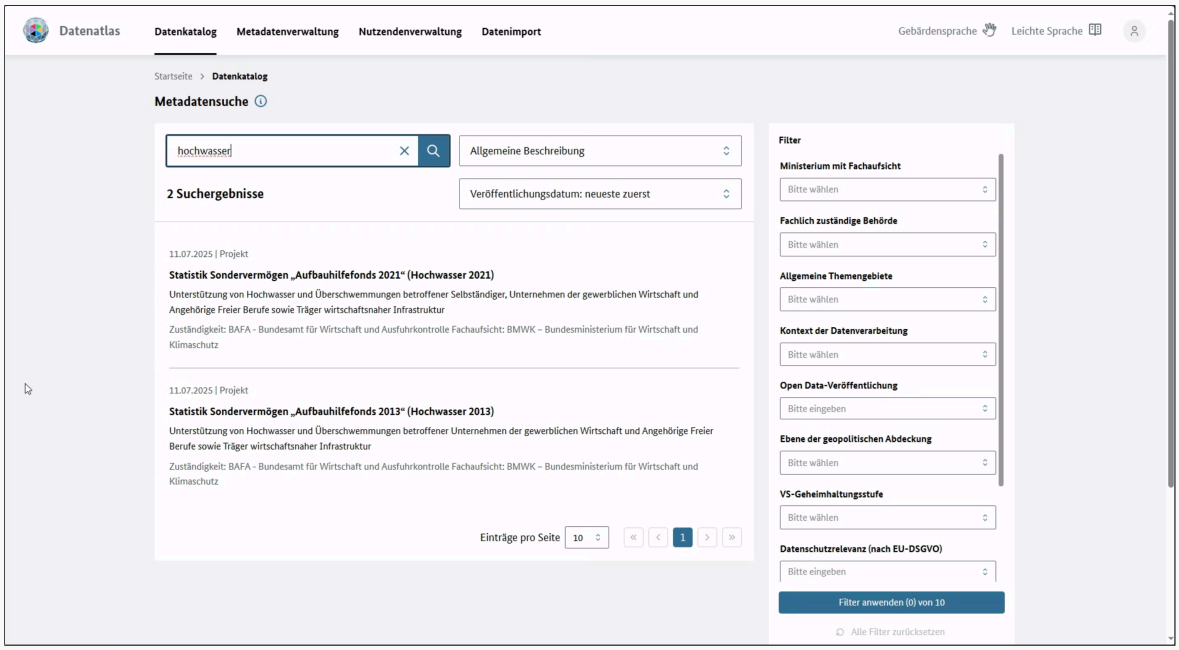


Abbildung 4.5: Trefferliste zum Suchbegriff „hochwasser“

1. Stärkt „Konsistenz“ (*Usability*) – die Ergebnisse werden erwartungskonform in Listensemantik präsentiert
2. Verletzt „Informatives Feedback“ (*Usability*) – keine Hervorhebung von Treffern in den Datensätzen; es ist nicht erkennbar, warum die Treffer relevant sein sollen
3. Stärkt „Fehlervermeidung“ (*Usability*) – ausgegraute Seitennavigation bei weniger als 10 Einträgen verhindert Fehlbedienungen
4. Verletzt den *Stand der Technik*, weil keine Facettierung zur Verfeinerung der Suche möglich ist
5. Verletzt den *Stand der Technik*, weil keine Hervorhebung von Treffern in den Datensätzen vorhanden ist

Informationsrecherche III – Schlagwort-basierte Suche: 2. Iteration

Schlagwort-basierte Suche: 2. Iteration

Leicht unzufrieden entschließt sich Uwe dafür, seine Anfrage zu modifizieren, um sein *Informationsbedürfnis* stillen zu können. Er ist sich sicher, dass es Förderprogramme zum Hochwasserschutz bzw. mit Bezug zum Oberthema Wasser gibt.

Nachdem es ihm nicht gelingt, Suchoperatoren zu verwenden, erinnert er sich an eine weitere Möglichkeit: Er weiß aus anderen Suchmaschinen, dass er Wildcards verwenden kann, um auch Treffer zu erhalten, die nur Teile des Suchbegriffs enthalten.
Er gibt, wie in Abb. 4.6 zu sehen, „*wasser“ ein, um Dokumente zu erhalten, die auf „wasser“ enden.

Wichtiger Hinweis: Die Anzahl der Suchergebnisse in Abb. 4.6 weicht von den anderen Screenshots in diesem Kapitel ab, da dieser Screenshot aus Qualitätsgründen zu einem späteren Zeitpunkt neu erstellt werden musste.

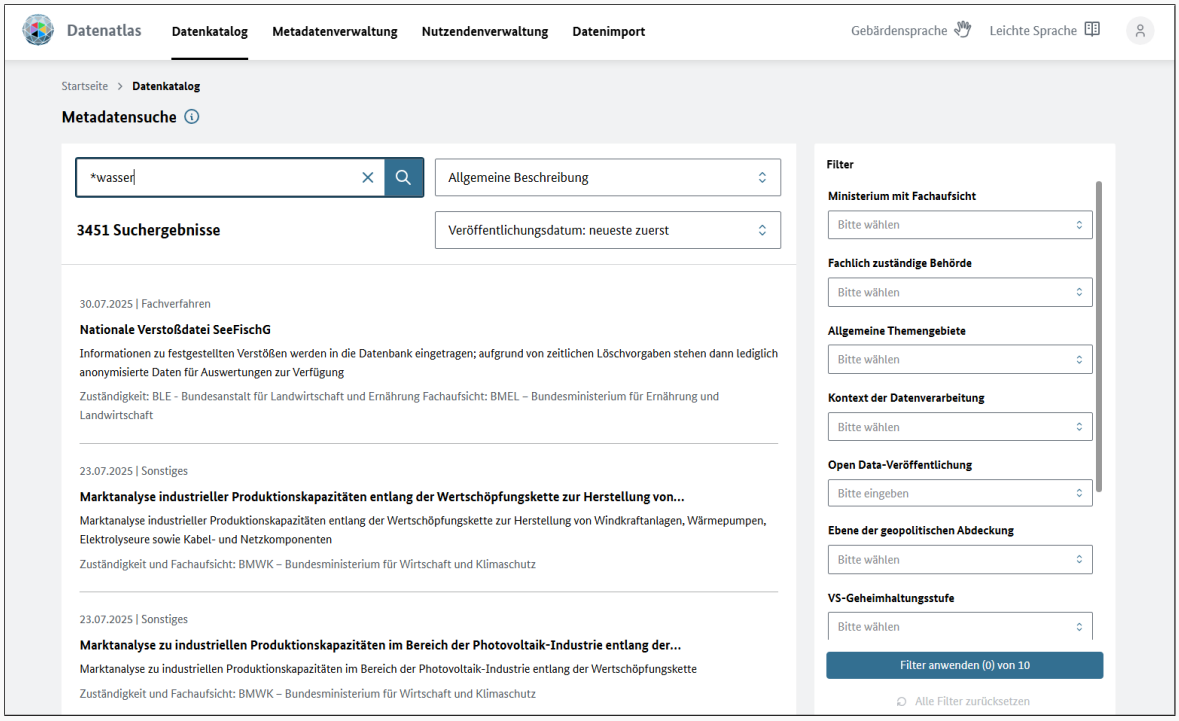


Abbildung 4.6: Datenkatalog; Suchbegriff „*wasser“

1. Verletzt den *Stand der Technik*, weil keine Wildcards oder Trunkierung genutzt werden können
2. Verletzt den *Stand der Technik*, weil keine Booleschen Operatoren unterstützt werden
3. Verletzt den *Stand der Technik*, weil keine Metadaten-Feld-basierte Suche ermöglicht wird
4. Verletzt den *Stand der Technik*, weil keine Phrasensuche zur Verfügung steht
5. Verletzt den *Stand der Technik*, weil keine Fuzzy Search nutzbar ist

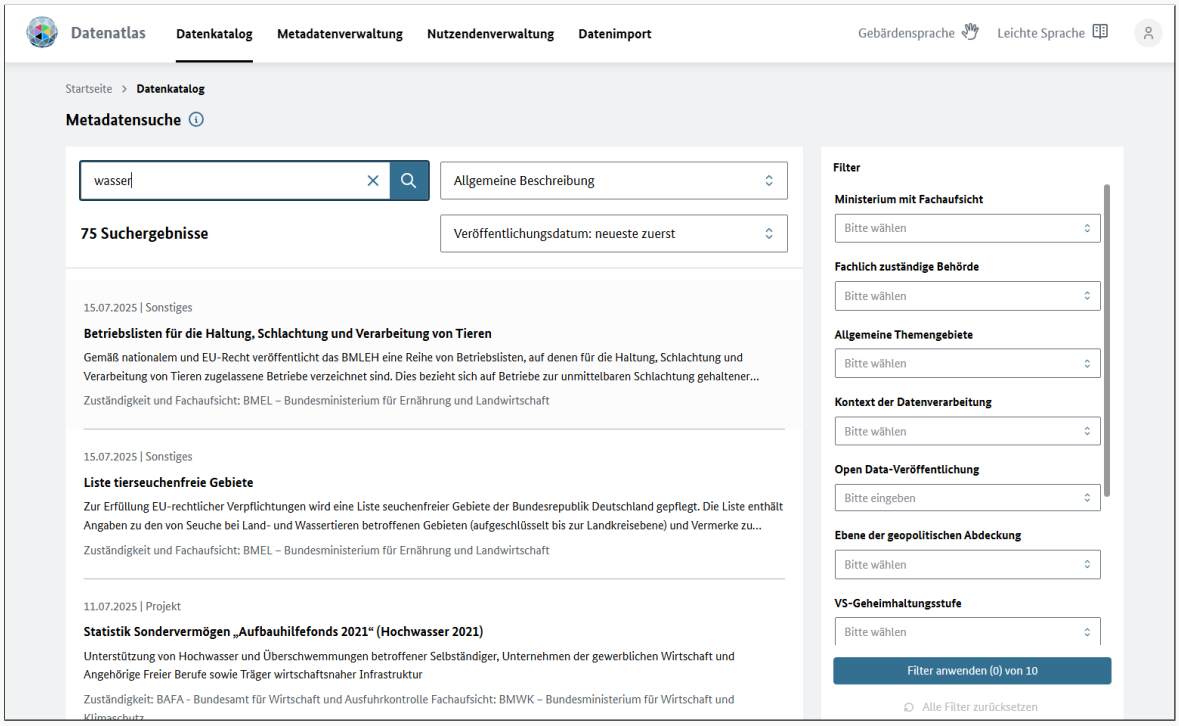


Abbildung 4.7: Trefferliste zum Suchbegriff „*wasser“

Nach dem Start der Suche erhält Uwe 75 Treffer, deren erste Einträge in Abb. 4.7 zu erkennen sind. Wie die Treffer zustande kommen, kann er sich nicht direkt erklären. Während er sich beim ersten Treffer herleiten kann, wie der Wasserbezug entsteht, fällt ihm das beim zweiten Eintrag schwer. Der dritte Treffer bestätigt ihm aber, dass die Suche wahrscheinlich korrekt funktioniert.

Er wundert sich jedoch, dass der einleitende Stern seines Suchbegriffs durch die Anwendung entfernt wurde.

1. Verletzt den *Stand der Technik*, weil keine Hervorhebung von Treffern in den Datensätzen vorhanden ist
2. Stärkt „Informatives Feedback“ (*Usability*) – Anzahl der Suchergebnisse wird ausgewiesen
3. Verletzt „Informatives Feedback“ (*Usability*) – die Zusammensetzung der Suchergebnisse ist nur teilweise nachvollziehbar
4. Stärkt „Abgeschlossene Aktionen“ (*Usability*) – die Präsentation der Suchergebnisse impliziert den Aktionsabschluss
5. Verletzt „Kontrollierbarkeit“ (*Usability*) – das System nimmt selbständig Änderungen an der Anfrage vor
6. Verletzt „Informatives Feedback“ (*Usability*) – über die Modifikation der Suche erfolgt keine Rückmeldung

Informationsrecherche IV – Schlagwort-basierte Suche:
3. Iteration - Sortierung

Schlagwort-basierte Suche: 3. Iteration - Sortierung

Nachdem Uwe einige Suchergebnisse inspiziert hat und nicht vollständig zufrieden ist, probiert er eine neue Anfrage aus (siehe Abb. 4.8).

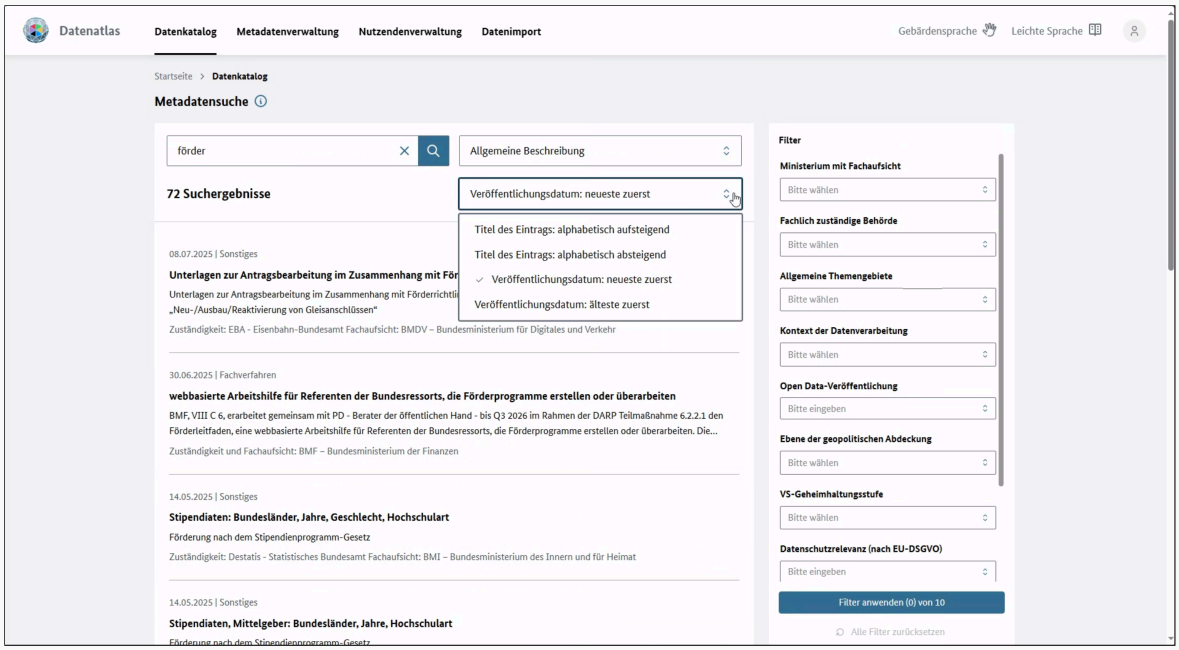


Abbildung 4.8: Trefferliste zum Suchbegriff „förder“; Sortierkriterien

Da er 72 Suchergebnisse erhält und auf den ersten Blick nicht erkennen kann, wie sich diese zusammensetzen, probiert er die Sortierung dieser aus. Das dafür nötige Auswahlmenu findet er an der erwarteten Stelle.

Uwe sucht nach der Auswahloption, bei der ihm besonders relevante Dokument zuerst angezeigt werden. Diese Option existiert nicht, so dass er sich für eine Sortierung nach dem Veröffentlichungsdatum entscheidet, um die neuesten Datensätze zuerst zu sehen.

1. Stärkt „Konsistenz“ (*Usability*) – Menuauswahl der Sortierungskriterien an üblicher Position
2. Verletzt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – es bleibt unklar, ob dies die Einstellung im DATENATLAS oder das Erstellungsdatum des ursprünglichen Datensatzes betrifft
3. Stärkt „Aktualität“ (*Datenqualitätsdimension*) – eine Sortierung nach dem Veröffentlichungsdatum impliziert Aktualität, obwohl sich dies nicht zweifelsfrei belegen lässt (s.o.)
4. Verletzt den *Stand der Technik*, weil keine Relevanzsortierung möglich ist; ein alleiniger Bedarf an den angebotenen Sortierkriterien ist fraglich

Informationsrecherche V – Schlagwort-basierte Suche: Filterung

Schlagwort-basierte Suche: Filterung

Da die Sortierung Uwe nicht bei der Verringerung seines Arbeitsaufwands geholfen hat, entscheidet sich Uwe, die Suchergebnisse zu filtern, um weniger potenziell relevante Dokumente durchsehen zu müssen.

Mit einem Klick kann er die Ergebnisse auf ein Ministerium einschränken (siehe Abb. 4.9).

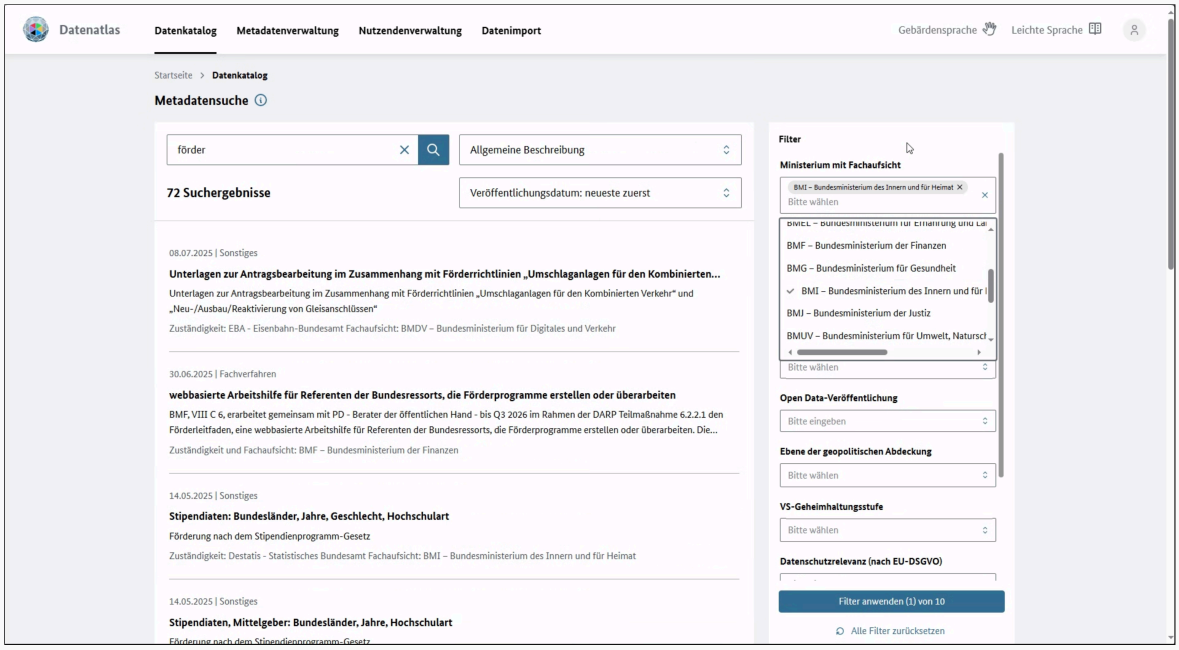


Abbildung 4.9: Filteransicht „Ministerium mit Fachaufsicht"

1. Stärkt „Konsistenz" (*Usability*) – übliche Filterauswahl durch Menuauswahl, verständliche Schaltflächen (X) zum Deaktivieren von Filtern
2. Stärkt „Informatives Feedback" (*Usability*) – gewählte Filter werden visualisiert
3. Verletzt „Informatives Feedback" (*Usability*) durch Verzicht auf Facettierung bleibt unklar, ob sich die Aktivierung eines Filters lohnt, da keine zu erwartenden Treffer visualisiert werden
4. Stärkt „Abgeschlossene Aktionen" (*Usability*) – „Filter anwenden"-Schaltfläche schließt Aktion klar ab
5. Verletzt „Fehlervermeidung" (*Usability*) – durch den Verzicht auf Facettierung bleibt unklar, ob sich die Aktivierung eines Filter lohnt, da keine zu erwartenden Treffer visualisiert werden; es besteht das Risiko, Filter zu streng bzw. widersprüchlich zu setzen
6. Stärkt „Einfache Umkehrbarkeit von Aktionen" (*Usability*) – alle Filter lassen sich zurücksetzen, einzelne Filter lassen sich mittels (X) einzeln entfernen
7. Stärkt „Geringe Belastung des Arbeitsgedächtnisses" (*Usability*) – gewählte Filter werden visualisiert
8. Stärkt „Kontrollierbarkeit" (*Usability*) – es muss explizit „Filter anwenden" angeklickt werden, um gefilterte Ergebnisliste zu erhalten

9. Verletzt den *Stand der Technik*, weil keine Facettierung existiert; es ist unklar, ob Ergebnisse zu erwarten sind

Nachdem Uwe den Filter „Ministerium mit Fachaufsicht“ auf BMI eingestellt und die „Filter anwenden“-Schaltfläche aktiviert hat, erhält er eine gefilterte Trefferliste, die in Abb. 4.10 dargestellt ist.

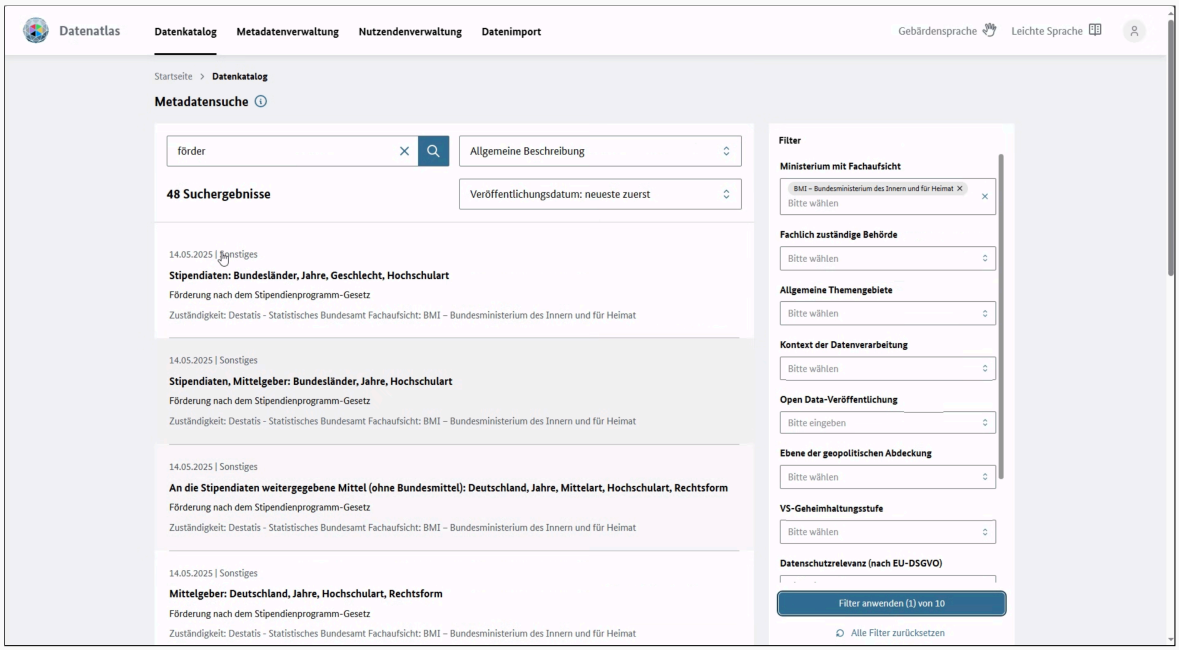


Abbildung 4.10: Ergebnisliste nach Filterung „Ministerium mit Fachaufsicht“

1. Stärkt „Konsistenz“ (*Usability*) – gewählte Filter bleiben sichtbar und wie zuvor modifizierbar
2. Stärkt „Informatives Feedback“ (*Usability*) – gewählte Filter bleiben sichtbar

Informationsrecherche VI – Schlagwort-basierte Suche: Verfeinerung der Filterung

Schlagwort-basierte Suche: Verfeinerung der Filterung

Da die Filterung gut funktioniert hat, möchte Uwe diese weiter verfeinern. Er erinnert sich, dass er einmal von einem Förderprogramm gehört hat, welches vom BMI verwaltet wurde und zu dem es ein Urteil, wahrscheinlich vom BUNDESARBEITSGERICHT, gab. Hierzu müssten sich Dokumente finden lassen. Er wählt den entsprechenden Filter, wie in Abb. 4.11 zu sehen ist, direkt zusätzlich aus und weist das System an, die neue Filterung anzuwenden.

Anmerkung des Autors Das BUNDESARBEITSGERICHT ist als Gericht unabhängig, jedoch unter der Dienstaufsicht des BUNDESMINISTERIUMS FÜR ARBEIT UND SOZIALES. In diesem Beispiel hätte dies Uwe auffallen können, dass hier das BMI keine Fachaufsicht hat. Ob diese Annahme jedoch auf jede Kombination aus obersten und nachgeordneten Bundesbehörden zutrifft, ist schon allein aufgrund der Komplexität kritisch zu hinterfragen.

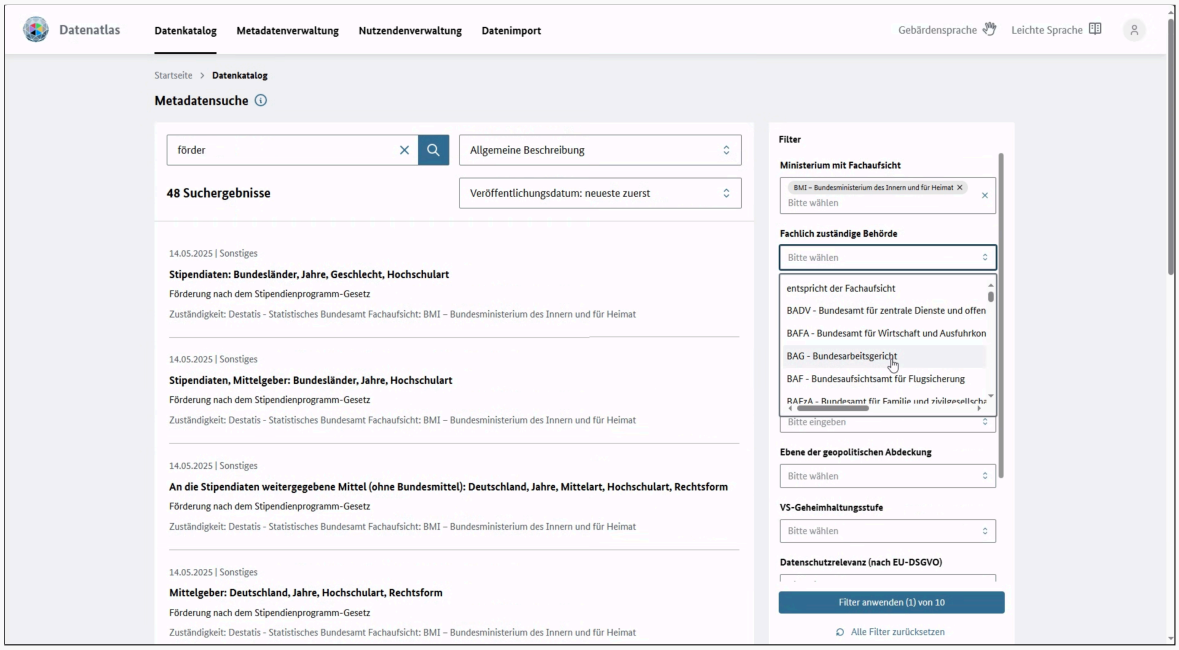


Abbildung 4.11: Trefferliste zum Suchbegriff „förder“, Sortierkriterien

1. Verletzt „Fehlervermeidung“ (*Usability*) – es ist möglich, „inkompatible“ Behörden zu wählen
2. Verletzt „Fehlervermeidung“ (*Usability*) – trotz sich logisch ausschließender „inkompatibler“ Behörden bleiben diese auswählbar
3. Stärkt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – gewählte Filter werden visualisiert
4. Verletzt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – externes Wissen über die Organisation der Bundesverwaltung muss genutzt werden, um gültige Filterkombinationen setzen zu können

1. Verletzt „Informatives Feedback“ (*Usability*) – das Feedback ist nicht hilfreich, da die „Inkompatibilität“ der Bundesbehörden zweifelsfrei durch das System hätte bestimmt werden können (siehe oben)
2. Verletzt „Informatives Feedback“ (*Usability*) – gerade bei komplexeren Filterkombinationen ist die Fehlermeldung sehr vage
3. Verletzt „Fehlervermeidung“ (*Usability*) – leere Ergebnisliste wird nicht im Vorfeld verhindert (siehe oben), Eingabe ungültiger Filter-Kombinationen ist möglich
4. Verletzt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – externes Wissen über die Organisation der Bundesverwaltung muss genutzt werden, um gültige Filterkombinationen setzen zu können
5. Verletzt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – Behebung des Fehlers ist nur mit Domänenwissen möglich
6. Verletzt „Kontrollierbarkeit“ (*Usability*) – es entsteht ein subjektiver Kontrollverlust durch unerwartetes, leeres Ergebnis

7. Verletzt den *Stand der Technik*, weil im Falle leerer Suchergebnisse verwandte Suchen etc. angeboten werden könnten

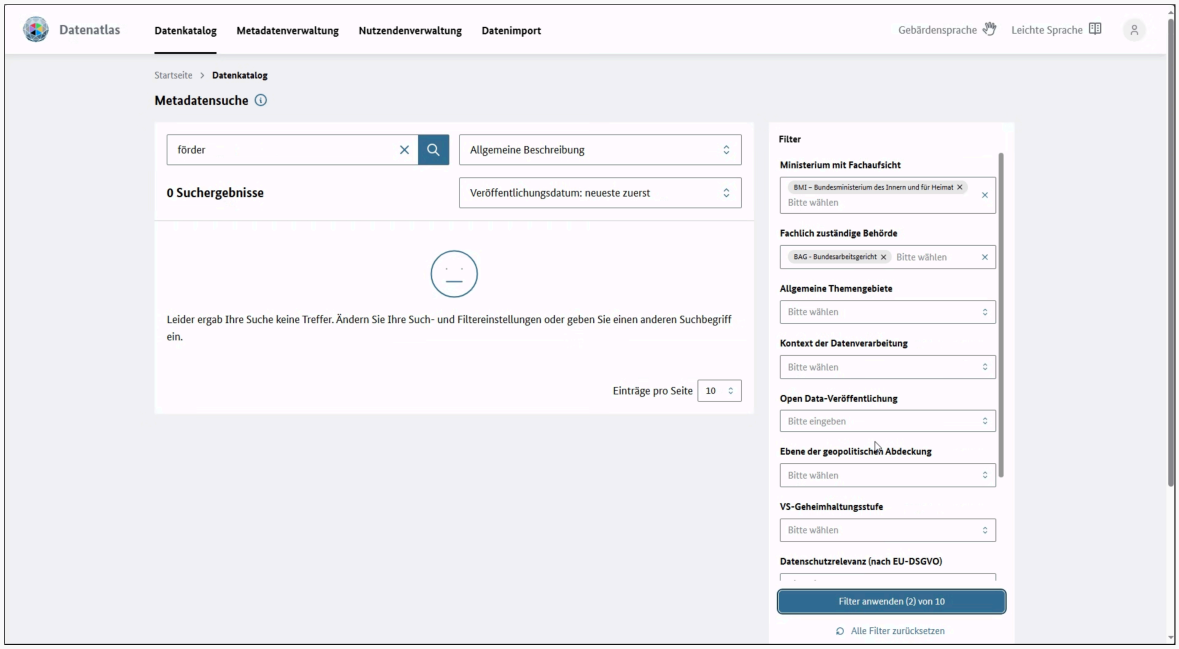


Abbildung 4.12: Ergebnisliste nach Filterung „Ministerium mit Fachaufsicht“ unter Hinzunahme des Filters „Fachlich zuständige Behörde“ bei nicht existierender Fachaufsicht

Hiweis des Autors Dieser und der folgende Prozess-Schritt dienen nur zur Illustration der prinzipiellen Interaktion mit dem DATENATLAS und ergeben, fachlich mit Bezug auf die Arbeitsaufgabe der Persona gesehen, wenig Sinn.

Letztendlich schafft es Uwe, die Filter passend zu wählen und erhält eine Liste an Suchergebnissen (siehe Abb. 4.13), die er sich für später abspeichern möchte.

Die dafür nötigen Schaltflächen kann er nicht entdecken, so dass er sich seinen Suchbegriff und die gewählten Filter auf Papier notiert, um die Suche später wieder reproduzieren zu können.

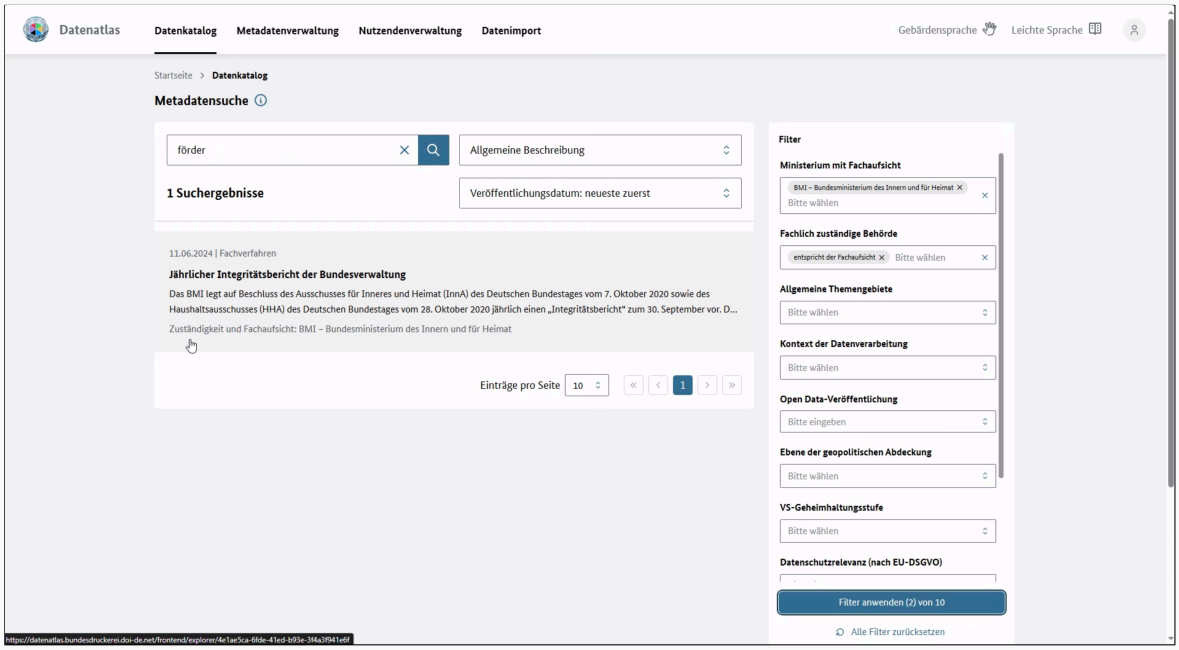


Abbildung 4.13: Ergebnisliste nach Filterung „Ministerium mit Fachaufsicht" unter Hinzunahme des Filters „Fachlich zuständige Behörde" bei existierender Fachaufsicht

Informationsrecherche VII – Ansicht Metadatensatz

Ansicht Metadatensatz

Uwe hat ein gesuchtes Dokument entdeckt, welches er sich detailliert anschauen möchte. Durch einen Klick in die Trefferliste kommt er wie erwartet in die Detailansicht (siehe Abb. 4.14).

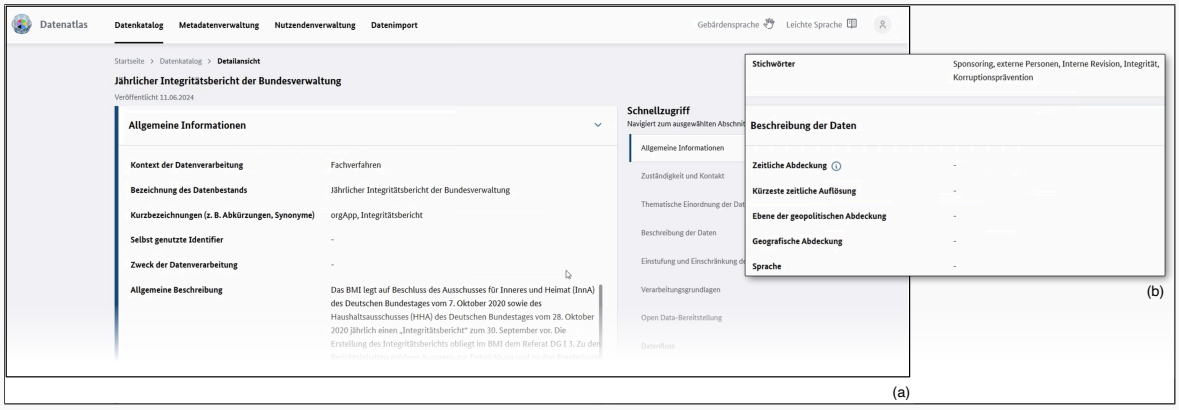


Abbildung 4.14: Detailansicht eines Metadatensatzes; Komposition zur besseren Lesbarkeit

Da Uwe den DATENATLAS nur unregelmäßig nutzt, kann er sich nur in Teilen daran erinnern, wofür die einzelnen Metadatenfelder stehen (Abb. 4.14; (a), links). Er ist unsicher, was sich hinter der Abkürzung „orgApp" bzw. dem „Zweck der Datenverarbeitung" verbirgt.
Es ist ihm nicht klar, warum viele Datenfelder nicht belegt sind.

1. Stärkt „Konsistenz" (*Usability*) – Detailansicht lässt sich durch Anklicken direkt in der Ergebnisliste öffnen

2. Verletzt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – die Bedeutung der Metadatenfelder bzw. der verwendeten Einträge im Datenbereich müssen bekannt sein, eine Online-Hilfe existiert häufig nicht
3. Verletzt „Genauigkeit“ (*Datenqualitätsdimension*) – müsste es nicht zumindest einen „Zweck der Datenverarbeitung“ geben? ohne Domänen-Expertise ist eine abschließende Bewertung nicht möglich
4. Verletzt „Vollständigkeit“ (*Datenqualitätsdimension*) – offensichtlich sind nicht alle Metadatenfelder mit Daten versehen
5. Verletzt „Konsistenz“ (*Datenqualitätsdimension*) – die Kurzbezeichnungen wirken wie Freitext unter Ausschluss von *kontrollierten Vokabularen* oder *Thesauri* (ansonsten wäre eine erläuternde Verlinkung sinnvoll)
6. Verletzt „Gültigkeit“ (*Datenqualitätsdimension*) – ergibt sich aus den oben genannten Defiziten

Trotzdem scrollt er weiter, um weitere Beschreibungen des Datensatzes zu finden, die beispielhaft in Abb. 4.14; (b) zu erkennen sind. Er versucht die Metadaten für die spätere Nutzung zu exportieren.

1. Verletzt „Konsistenz“ (*Usability*) – das Metadatenfeld „Zeitliche Abdeckung“verfügt über ein ℹ-Icon, der Rest nicht
2. Verletzt „Informatives Feedback“ (*Usability*) – Stichwörter werden nicht erläutert bzw. sind nicht anklickbar; wahrscheinlich keine *kontrollierte Vokabulare* oder *Thesauri* hinterlegt
3. Verletzt „Gültigkeit“ (*Datenqualitätsdimension*) – kaum Pflichtfelder, dafür viele Freitext-Felder; Stichwörter werden nicht erläutert bzw. sind nicht anklickbar, wahrscheinlich keine *kontrollierte Vokabulare* oder *Thesauri* hinterlegt
4. Verletzt den *Stand der Technik*, weil kein Export der Metadaten oder kein Bookmarking, möglich sind

4.2 *Statistische Auswertung – Informationsrecherche*

Die User Journey der ersten Persona, Uwe, zeigt einige Stärken und Schwächen der angebotenen Funktionen des DATENATLAS bzw. seiner Subkomponente „Datenkatalog“ auf, welche auf Grundlage der in Kapitel 3 vorgestellten Kriterien in Ausschnitten bewertet wurden. Die Ergebnisse der Kurzevaluierung werden einander in Tabelle 4.1 gegenübergestellt.

Tabelle 4.1 weist dabei „Stärkungen“ und „Verletzungen“ anhand verschiedener Qualitätsbereiche, der *User Experience* (UX), der *Datenqualität* (DQ) und des allgemeinen *Stands der Technik* aus. Wie eingangs erläutert, werden wiederkehrende „Stärkungen“ und „Verletzungen“ in den jeweiligen Prozess-Schritten nicht immer wieder erneut genannt oder aufsummiert, um die Lesbarkeit des Gutachtens zu erhöhen.

EIN HOHER WERT im Bereich der „Stärkungen“ ist nicht unein-

geschränkt als positiv zu bewerten. Er drückt lediglich aus, dass das System in bestimmten Punkten den *Stand der Technik* erreicht, was im Rahmen eines Entwicklungsauftrags die Mindesterwartung (siehe unten) darstellt.

Tabelle 4.1: Statistiken zur User Journey „Informationsrecherche“

Typ	Gesamt[8]	UX	DQ	Stand der Technik
„Stärkungen“	22	21	1	1
„Verletzungen“	48	28	5	15

[8] Insgesamt wurden in der User Journey „Informationsrecherche“ 7 Prozess-Schritte betrachtet.

Besonderes Augenmerk ist jedoch auf die „Verletzungen“ zu richten, da diese Defizite der Teilkomponenten des Datenatlas aufzeigen, an denen angesetzt werden muss, um die Potenziale des Datenatlas besser nutzbar machen zu können. Unter Umständen stellen diese Punkte Sachmängel am System dar, da in diesen Bereichen der *Stand der Technik* nicht erreicht wurde.

Im allgemeinen Bereich *Stand der Technik* ist leicht erkennbar, dass hier kaum „Stärkungen“ aufgeführt werden. Dies darf nicht überraschen, da davon auszugehen ist, dass dieser Stand zu erreichen ist, damit kein Sachmangel vorliegt[9]. „Verletzungen“ am *Stand der Technik* sind aus Sicht des Gutachters als Sachmangel zu werten, da ihm kein Vertrag vorliegt, der dies ausschließt.

[9] Siehe §434 BGB.

Die Tabelle verdeutlicht, dass Mitte Juli 2025 einer Erreichung des Stands der Technik durch den Datenatlas der 2,18-fache Wert an Mängeln gegenübersteht.

4.3 *Metadatenverwaltung*

In dieser User Journey ist es Elins (siehe Abb. 4.1; rechts) Aufgabe, einen neuen Metadatensatz im Datenatlas im Zuständigkeitsbereich ihres Ministeriums anzulegen. Die zugrundeliegenden Metadaten wurden durch eine nachgeordnete Behörde erhoben und per E-Mail übermittelt, da im Bundesamt gerade eine Stellenvakanz im zuständigen Datenmanagement-Referat existiert.

Es ist unklar, ob die nachgeordnete Behörde den Metadatensatz bereits im Datenatlas angelegt hat.

Metadatenverwaltung I – Übersicht Metadatenverwaltung

Übersicht Metadatenverwaltung

Im ersten Schritt ruft Elin das Modul „Metadatenverwaltung“ auf, um sich einen ersten Überblick zu verschaffen (siehe Abb. 4.15).

In ihrem Ministerium sind 88 Einträge gelistet, sie erkennt jedoch keine Datensätze der nachgeordneten Behörden, da diese aktuell nichts veröffentlicht haben.

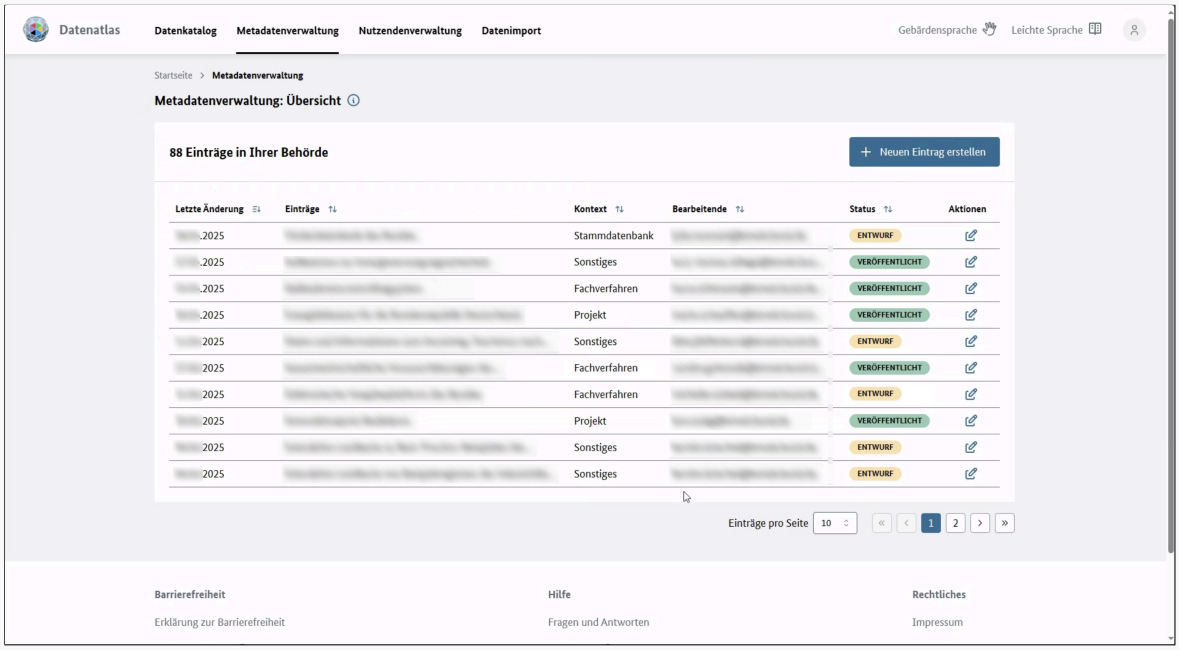


Abbildung 4.15: Übersicht Metadatenverwaltung (personenbezogene Daten wurden aus Datenschutzgründen unkenntlich gemacht)

1. Stärkt „Konsistenz“ (*Usability*) – die Tabellenansicht folgt üblichen Web-Standards
2. Stärkt „Kontrollierbarkeit“ (*Usability*) – die Sortierung der Spalten erfolgt nach Wunsch der Nutzenden
3. Verletzt „Vollständigkeit“ (*Datenqualitätsdimension*) – da Datensätze der nachgeordneten Behörden nicht gelistet werden, kann dies nicht durch Nutzende bewertet werden; weitere Implikationen auf die menschzentrierte Entwicklung werden in Abschnitt 5.3 thematisiert
4. Verletzt „Aktualität“ (*Datenqualitätsdimension*) – da Datensätze der nachgeordneten Behörden nicht gelistet werden, kann dies nicht bewertet werden

Metadatenverwaltung II – Neuanlage von Metadaten

Neuanlage von Metadaten

Da Elin keinen existierenden Datensatz finden kann, entscheidet sie sich, auf Grundlage der eingegangenen E-Mail einen neuen Eintrag zu erstellen.

Abbildung 4.16 zeigt den Informationsdialog, welchen Elin vor der Eingabe bestätigen muss. Auch wenn sie sich unsicher ist, ob sie die juristischen Aspekte korrekt einschätzen kann, stimmt sie der Belehrung zu.

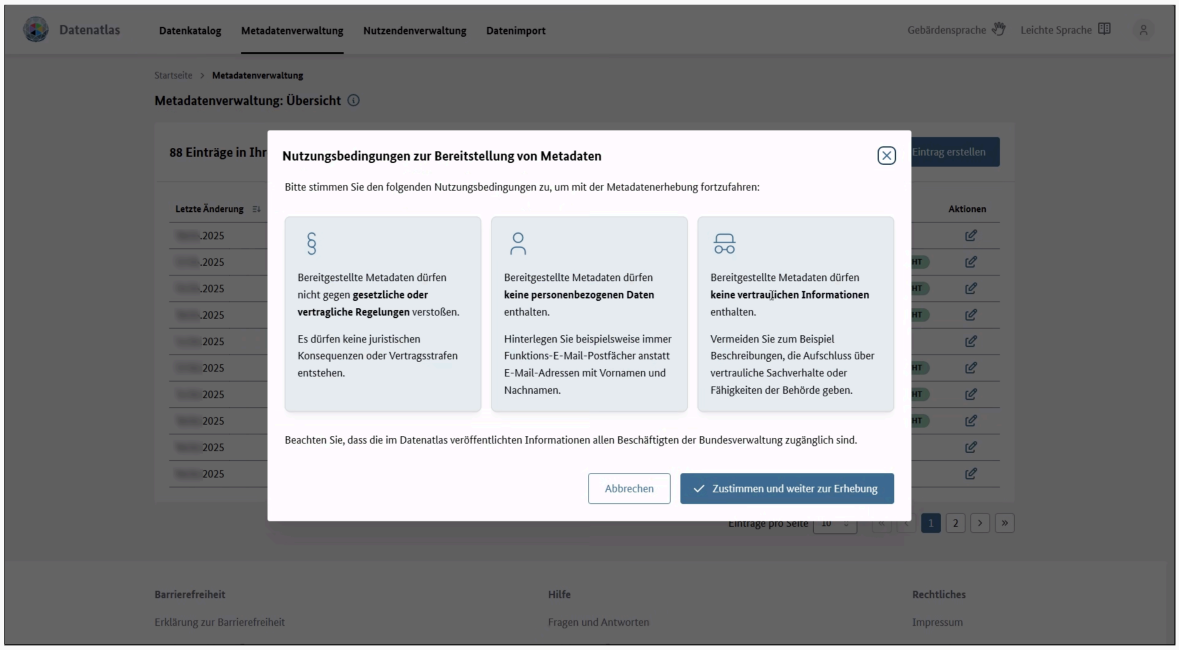


Abbildung 4.16: Neuanlage von Metadaten

1. Verletzt „Konsistenz“ (*Usability*) – typischer Web-Dialog, aber in Apple-Reihenfolge, obwohl die BUNDESVERWALTUNG primär Windows nutzt (`https://www.nngroup.com/articles/ok-cancel-or-cancel-ok/`; Letzter Abruf: 01.08.2025)
2. Stärkt „Informatives Feedback“ (*Usability*) – klares Informationsangebot, Konsequenz des Handelns wird klar
3. Stärkt „Abgeschlossene Aktionen“ (*Usability*) – klare Arbeitssequenz
4. Stärkt „Fehlervermeidung“ (*Usability*) – gegeben
5. Verletzt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – es ist unklar, wie Nutzende prüfen sollen, ob juristische oder Vertragsstrafen entstehen; hierzu sind weiterführende Recherchen nötig
6. Stärkt „Kontrollierbarkeit“ (*Usability*) – gegeben, Abbruch bzw. Fortfahren jederzeit möglich
7. Verletzt den *Stand der Technik*, weil das System automatisiert prüfen könnte, welche Arten von E-Mails angegeben werden (z.B. über Namenslisten)

Metadatenverwaltung III – Eingabe Metadatensatz

Eingabe Metadatensatz

Nach der Zustimmung öffnet sich die Eingabemaske zur Erstellung eines neuen Datensatzes (siehe Abb. 4.17).

Im ersten Moment ist Elin angesichts der großen Anzahl an Eingabefeldern überfordert. Sie findet sich jedoch schnell zu Recht, da ihr der Schnellzugriffsbereich dabei hilft zu erkennen, was von ihr erwartet wird. Pflichtfelder sind zudem mit einem Sternchen gekennzeichnet. Ihr fällt

schnell auf, dass es nicht besonders viele Pflichtfelder und relativ häufig Freitextfelder ohne Eingabebeschränkung gibt. Sie rechnet damit, die Daten schnell eingeben zu können, auch wenn sie nicht immer zu 100% sicher ist, was sie in jedes Feld eingeben soll. Zeitgleich befürchtet sie aufgrund ihres fachlichen Hintergrunds, dass diese Freiheit generell die *Datenqualität* beeinträchtigen könnte.

Diese *„Freitext-Problematik“* wird separat in Abschnitt 4.7 thematisiert.

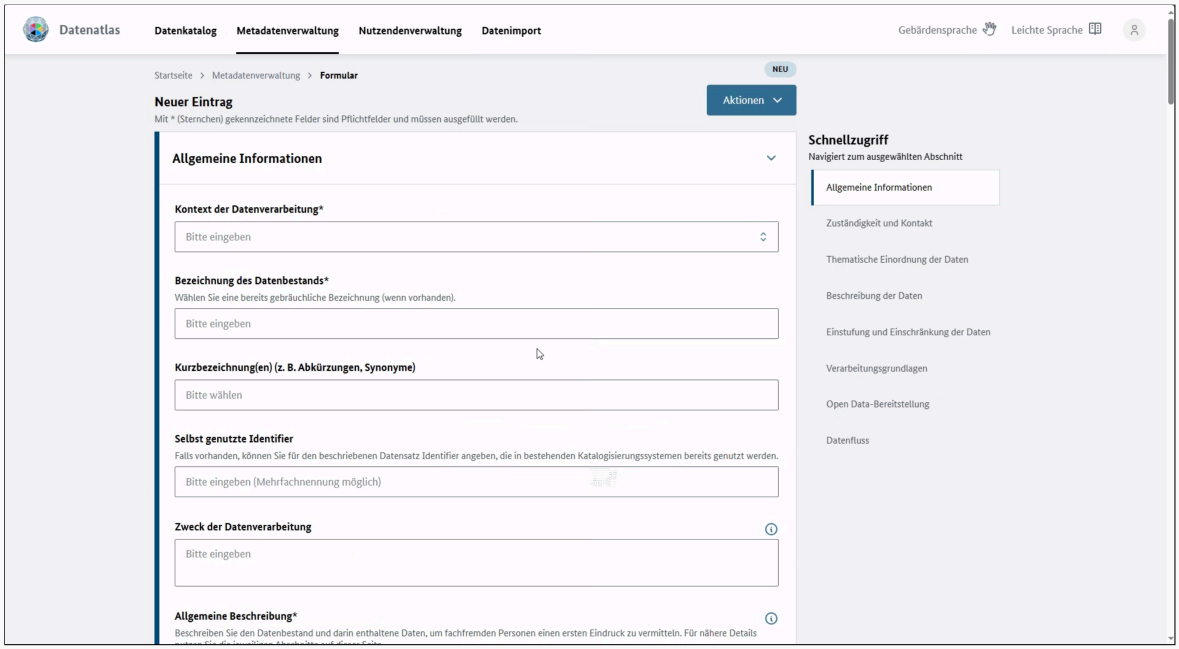


Abbildung 4.17: Eingabe Metadatensatz; Teilausschnitt

1. Stärkt „Konsistenz“ (*Usability*) – übliches Format einer Eingabemaske
2. Verletzt „Konsistenz“ (*Usability*) – inkonsistente Nutzung von Erläuterungen und Online-Hilfetexten bzw. ℹ-Icons
3. Verletzt „Informatives Feedback“ (*Usability*) – inkonsistente Nutzung von Erläuterungen und Online-Hilfetexten bzw. ℹ-Icons
4. Stärkt „Abgeschlossene Aktionen“ (*Usability*) – Eingaben werden durch Aktionen-Dropdown ausgelöst
5. Verletzt „Fehlervermeidung“ (*Usability*) – sehr viele Freitextfelder begünstigen Fehleingaben (*„Freitext-Problematik“*)
6. Verletzt „Einfache Umkehrbarkeit von Aktionen“ (*Usability*) – bei Freitexten sind Korrekturen recht aufwändig durchzuführen, wenn Fehleingaben erkannt werden (*„Freitext-Problematik“*); unproblematisch bei Dropdown-Menus
7. Verletzt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – Erfassende müssen wissen, wie die Felder konkret zu befüllen sind, da Hilfestellungen inkonsistent umgesetzt sind (*„Freitext-Problematik“*)
8. Stärkt „Kontrollierbarkeit“ (*Usability*) – Eingaben werden durch Aktionen-Dropdown ausgelöst

9. Verletzt „Genauigkeit“ (*Datenqualitätsdimension*) – die Vielzahl an Freitext-Feldern begünstigt Fehleingaben (*„Freitext-Problematik“*)
10. Verletzt „Vollständigkeit“ (*Datenqualitätsdimension*) – durch ungeprüfte Freitextfelder können auch bei Pflichtfeldern inkorrekte Daten eingegeben werden (*„Freitext-Problematik“*)
11. Verletzt „Konsistenz“ (*Datenqualitätsdimension*) – wird durch Freitext-Felder gefährdet, z.B. in Form von Vertippern bzw. Fehleingaben (*„Freitext-Problematik“*)
12. Verletzt „Gültigkeit“ (*Datenqualitätsdimension*) – eine automatisierte Formatprüfung auf Freitext ist kaum möglich (*„Freitext-Problematik“*)

Nachdem Elin eine Vielzahl an Eingaben vorgenommen hat, kommt sie langsam zum Ende und widmet sich dem Bereich „Open Data-Bereitstellung“. Bei der Dateneingabe stößt sie auf keine weiteren Probleme, da sie mittels Dropdown-Menus passende Optionen wählen kann.

Als sie zum Feld „Datennutzungslizenz“ gelangt, reagiert sie überrascht. Aus ihrem Studium erinnert sie sich daran, dass es sich bei der Apache-Lizenz und der BSD-Lizenz um traditionelle Lizenzen für die Bereitstellung von *Open Source-Software* handelt.

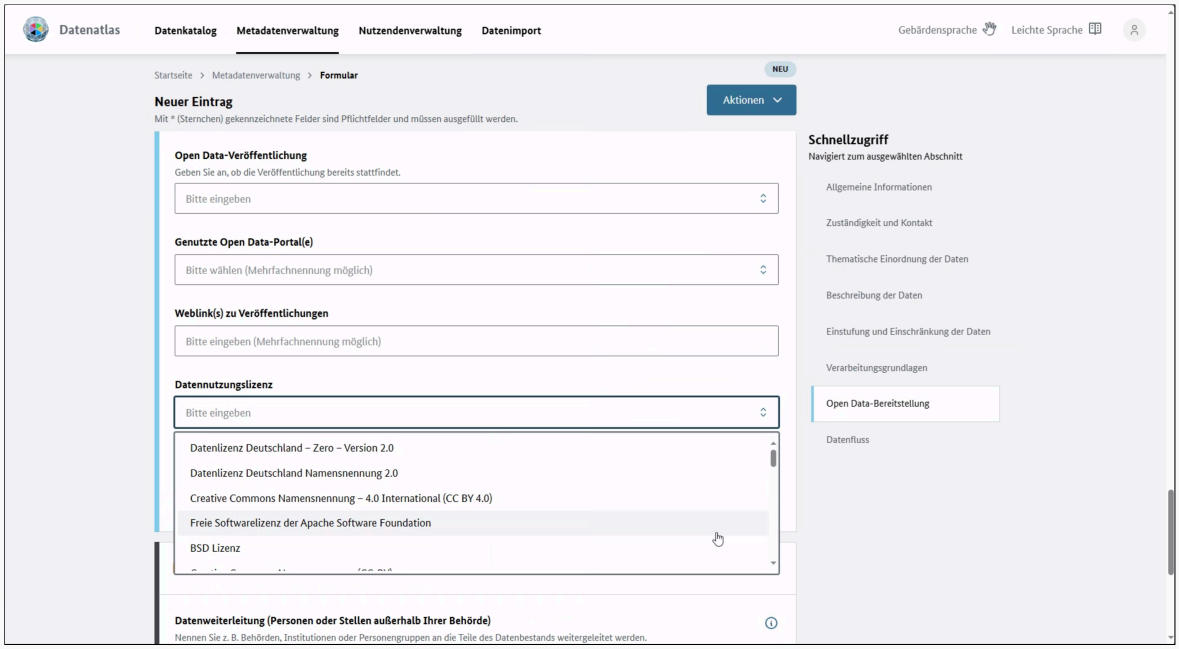


Abbildung 4.18: Eingabe Metadatensatz; Open Data-Veröffentlichung, Lizenzauswahl

1. Verletzt „Fehlervermeidung“ (*Usability*) – es können nicht zutreffende Lizenzen gewählt werden
2. Verletzt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – während der Eingabe müssen die Lizenzen und ihr Anwendungsbereich bekannt sein
3. Verletzt „Genauigkeit“ (*Datenqualitätsdimension*) – durch unpassende Einträge (z.B. im Bereich Lizenzen) kann es zu Fehleingaben kommen

4.4 *Statistische Auswertung – Metadatenverwaltung*

Tabelle 4.2 listet die bisherigen Erkenntnisse auf. Eine Interpretationshilfe der Tabelle befindet sich in Abschnitt 4.2.

Die Tabelle verdeutlicht, dass Mitte Juli 2025 einer Erreichung des Stands der Technik durch den DATENATLAS der 1,88-fache Wert an Mängeln gegenübersteht.

Tabelle 4.2: Statistiken zur User Journey „Metadatenverwaltung“

Typ	Gesamt[10]	UX	DQ	Stand der Technik
„Stärkungen“	9	9	0	0
„Verletzungen“	17	9	7	1

[10] Insgesamt wurden in der User Journey „Metadatenverwaltung“ 3 Prozess-Schritte betrachtet.

4.5 *Datenimport*

Da Elin (siehe Abb. 4.1; rechts) Kenntnis darüber hat, dass es möglich ist, Datensätze aus GovData in den Datenatlas zu importieren, ist es ihr Ziel, diese Funktion zu verwenden, um ihr bekannte Datensätze schnell im Datenatlas weiterverwenden und anderen bereitstellen zu können.

Als *Data Scientist* ist es ihr dabei wichtig, die *Datenqualität* im Blick zu behalten und zu verstehen, was beim Datenimport passiert, damit sich auch weniger versierte Anwendende auf den Datenatlas verlassen können.

Datenimport I – Datensuche

Datensuche

Elins Recherche beginnt im Teilmodul „Datensuche“ des Datenatlas, dessen Einstiegsseite in Abb. 4.19 abgebildet ist.

Der Sucheinstieg präsentiert sich übersichtlich, so dass sich Elin schnell zurechtfindet. Aufgrund der iterativen Entwicklung des Datenatlas steht aktuell nur GovData als Importquelle zur Verfügung.

Durch das Drop-Down-Menu der „Datenkategorien“ verschafft sie sich einen Überblick über den Inhalt der Datensätze, die für den Import zur Verfügung stehen.

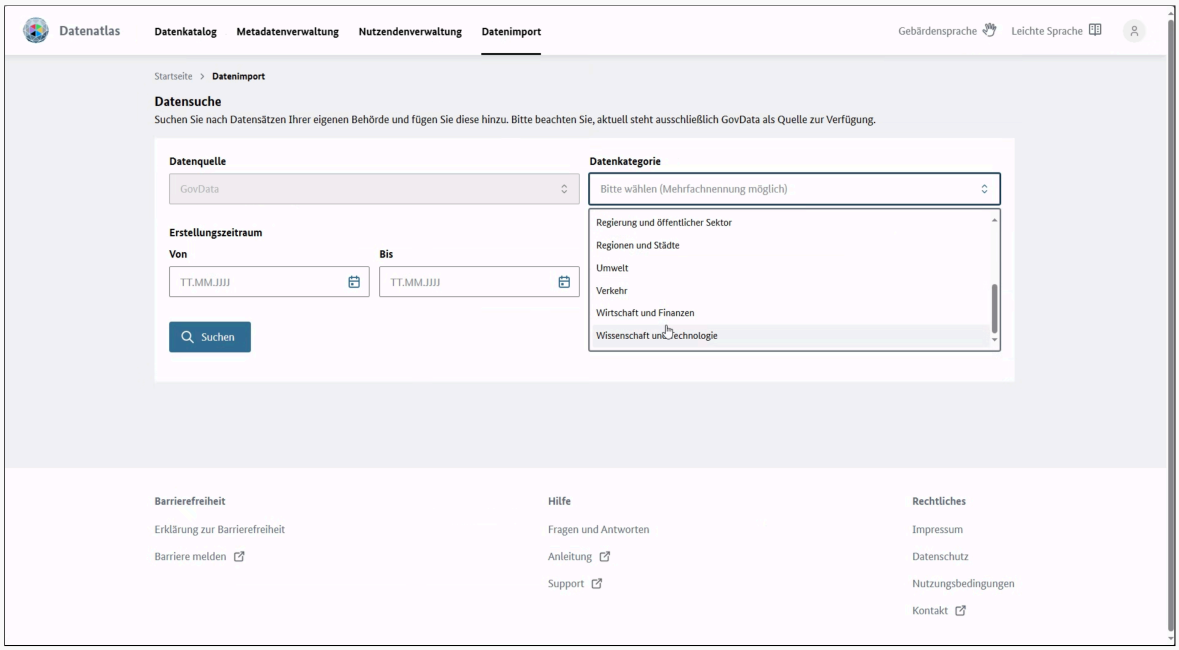


Abbildung 4.19: Datenimport; Datensuche, Auswahl Datenkategorie

1. Stärkt „Kontrollierbarkeit“ (*Usability*) – die Parameter-Setzung wird erst durch den Klick auf den „Suchen“-Button aktiv
2. Entspricht dem *Stand der Technik*, weil ein explorativer Sucheinstieg angeboten wird

3. Verletzt den *Stand der Technik*, weil keine gerichtete Suche unterstützt wird

Nachdem sie die Suche startet, erhält sie schnell eine Liste von 29 Datensätzen, die ihrer gewählten Kategorie entsprechen.

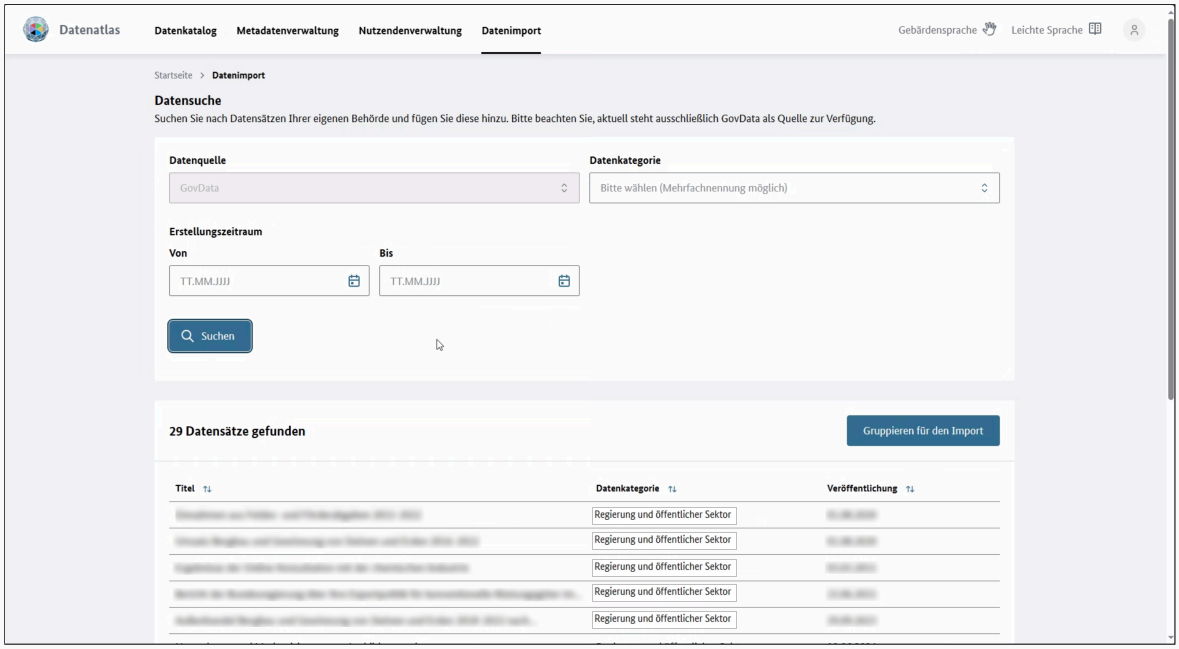


Abbildung 4.20: Datenimportansicht; Ergebnisliste – 29 Datensätze (Daten, welche die Zuordnung zu einzeln Behörden ermöglichen, wurden unkenntlich gemacht.)

Ein Ausschnitt der Ergebnisse ist in Abb. 4.20 zu sehen. Elin ist unsicher, ob sie einzelne Datensätze für den Import auswählen kann und anhand welcher Kriterien eine Gruppierung für den Import vorgenommen wird. Sie kann keine Gruppierungskriterien einstellen.

Sie entscheidet sich letztendlich aber doch „Gruppieren für den Import“ anzuklicken, da sie damit rechnet, die Aktion jederzeit rückgängig machen zu können.

1. Verletzt „Konsistenz“ (*Usability*) – Operationen, die sich auf Gruppen von Datensätzen bzw. Mehrfachauswahlen beziehen, sind in der Regel durch Checkboxen o.ä. selektierbar
2. Verletzt „Informatives Feedback“ (*Usability*) – die Auswirkung der Aktion „Gruppieren für den Import“ wird nicht erläutert
3. Stärkt „Abgeschlossene Aktionen“ (*Usability*) – der Import findet nur durch explizite Aufforderung statt
4. Verletzt „Fehlervermeidung“ (*Usability*) – ob das Gruppieren rückgängig gemacht werden kann, ist für Nutzende nicht erkennbar
5. Verletzt „Kontrollierbarkeit“ (*Usability*) – die Gruppierung erfolgt anhand intrasparenter Kriterien, nur der Aktionsstart ist kontrollierbar

Datenimport II – Ablauf Datenimport

Ablauf Datenimport

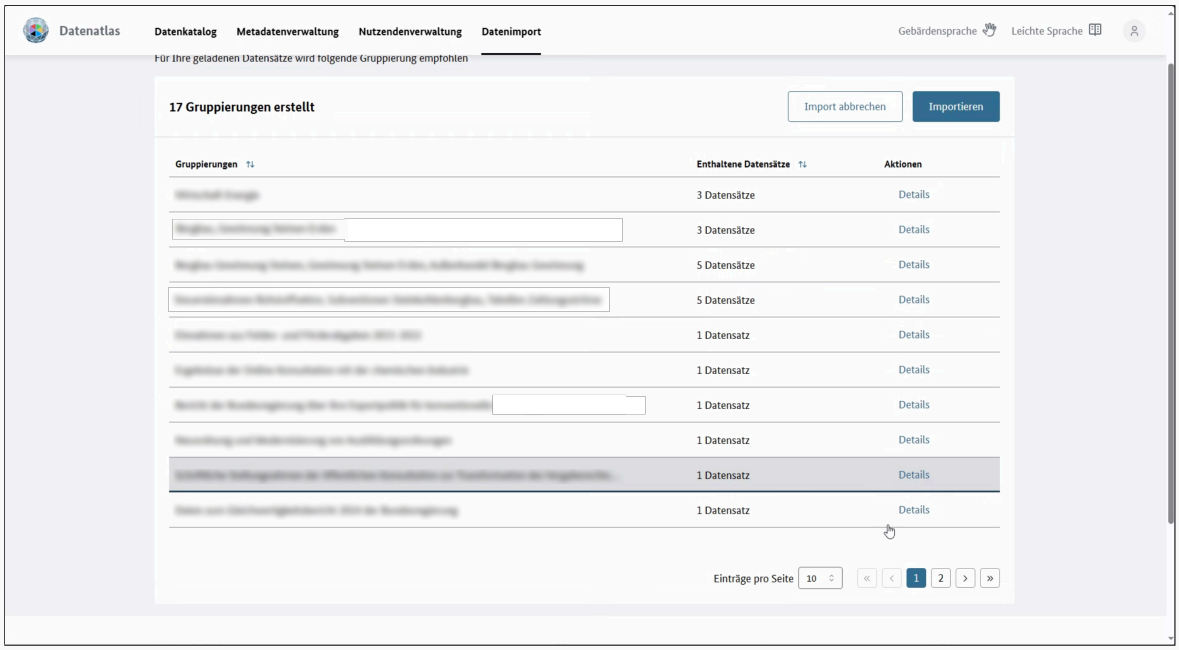


Abbildung 4.21: Datenimport; Ergebnisansicht

Nach kurzer Wartezeit wird Elin ein Ergebnis (siehe Abb. 4.21) präsentiert. Sie kann aufgrund des klaren Feedbacks erkennen, dass aus zuvor 29 Datensätzen 17 Gruppen erstellt wurden.

Gleichzeitig fragt sie sich noch immer, welches Kriterium zur Gruppierung genutzt wurde. Immerhin kann sie in der Liste deutlich erkennen, aus welcher Anzahl an Dokumenten die Einzelgruppen bestehen.

1. Verletzt „Informatives Feedback“ (*Usability*) – es bleibt unklar, wie die Gruppierung arbeitet
2. Stärkt „Fehlervermeidung“ (*Usability*) – der Import kann abgebrochen werden
3. Stärkt „Einfache Umkehrbarkeit von Aktionen“ (*Usability*) – der Import kann mittels eines Klicks abgebrochen werden
4. Verletzt „Kontrollierbarkeit“ (*Usability*) – es ist unklar, auf welcher Grundlage die Gruppierung erfolgt
5. Verletzt „Genauigkeit“ (*Datenqualitätsdimension*) – aufgrund unklarer Gruppierungskriterien kann die *Datenqualität* negativ beeinflusst
6. Verletzt „Vollständigkeit“ (*Datenqualitätsdimension*) – aufgrund unklarer Gruppierungskriterien kann die *Datenqualität* negativ beeinflusst
7. Verletzt „Konsistenz“ (*Datenqualitätsdimension*) – aufgrund unklarer Gruppierungskriterien kann die *Datenqualität* negativ beeinflusst
8. Verletzt „Aktualität“ (*Datenqualitätsdimension*) – aufgrund unklarer Gruppierungskriterien kann die *Datenqualität* negativ beeinflusst

9. Verletzt „Gültigkeit“ (*Datenqualitätsdimension*) – aufgrund unklarer Gruppierungskriterien kann die *Datenqualität* negativ beeinflusst

Um näheres über die Gruppierung zu erfahren, klickt Elin auf den „Details“-Link eines Eintrags (siehe Abb. 4.22), um weiteres über die Gruppierung zu erfahren.

Der Liste kann sie entnehmen, welche Einträge der Gruppe zugeordnet wurden, warum dies der Fall ist, muss sie jedoch selbst ermitteln.

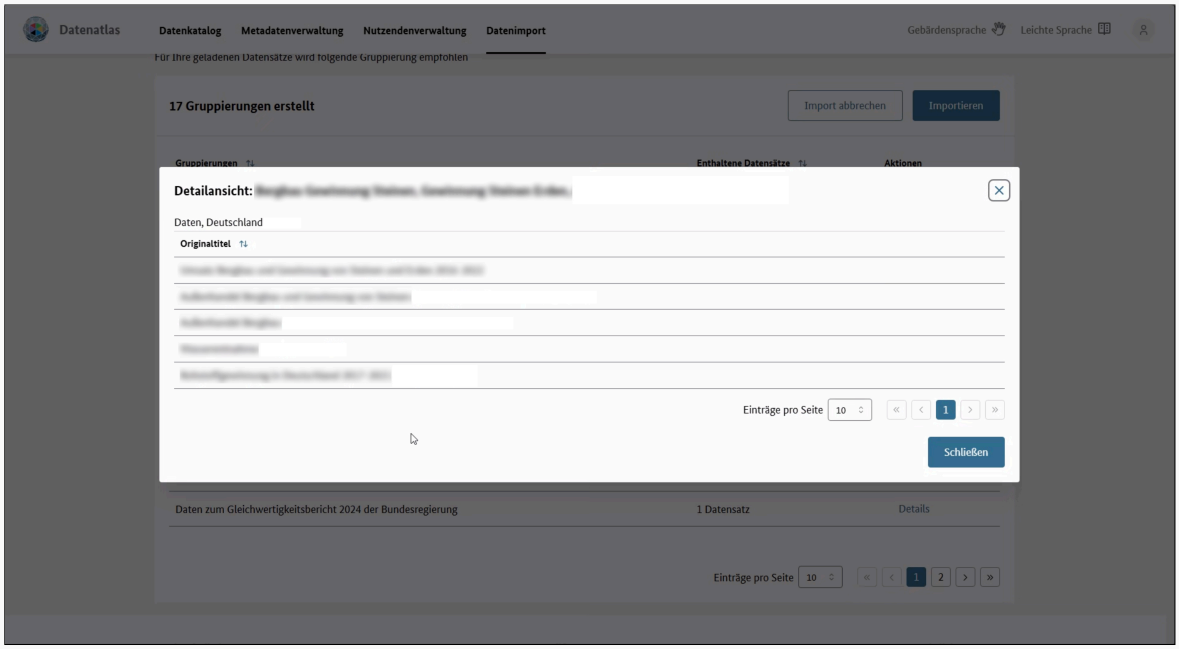


Abbildung 4.22: Detailansicht aggregierter Import-Datensätze (Daten, welche die Zuordnung zu einzeln Behörden ermöglichen, wurden unkenntlich gemacht.)

1. Verletzt „Informatives Feedback“ (*Usability*) – es liegt keine Erklärung für die Zusammensetzung der Gruppe vor
2. Verletzt „Geringe Belastung des Arbeitsgedächtnisses“ (*Usability*) – die Zusammensetzung der Gruppe kann wahrscheinlich nur mittels externer Hilfe bestimmt werden

4.6 *Statistische Auswertung – Datenimport*

Tabelle 4.3 enthält die bisherigen Erkenntnisse. Eine Interpretationshilfe der Tabelle befindet sich in Abschnitt 4.2.

Die Tabelle verdeutlicht, dass Mitte Juli 2025 einer Erreichung des Stands der Technik durch den DATENATLAS der 3,5-fache Wert an Mängeln gegenübersteht.

Tabelle 4.3: Statistiken zur User Journey „Datenimport“

Typ	Gesamt[11]	UX	DQ	Stand der Technik
„Stärkungen“	4	4	0	1
„Verletzungen“	14	8	5	1

[11] Insgesamt wurden in der User Journey „Datenimport“ 2 Prozess-Schritte betrachtet.

4.7 *Statistische Auswertung aller User Journeys*

Eine Interpretationshilfe für Tabelle 4.4 befindet sich in Abschnitt 4.2.

Die Tabelle zeigt, dass Mitte Juli 2025 einer Erreichung des Stands der Technik durch den DATENATLAS der 2,25-fache Wert an Mängeln gegenübersteht, wenn man die drei unterschiedlichen User Journeys gemeinsam betrachtet, welche die *minimalen Use Cases der Bundesverwaltung* abbilden[12].

[12] Siehe Kapitel 2.

Tabelle 4.4: Global-Statistik aller User Journeys

Typ	Gesamt[13]	UX	DQ	Stand der Technik
„Stärkungen“	35	34	1	2
„Verletzungen“	79	45	17	17

[13] Insgesamt wurden in der Gesamtheit aller User Journeys 12 Prozess-Schritte betrachtet.

DER EVALUIERUNGSMETHODE, d.h. der *heuristischen Inspektion*[14] und der Datengrundlage in Form von *Screenshots* ist geschuldet, dass die Ergebnisse der Auswertung in Richtung der Usability-Probleme hin verzerrt sind.

[14] Siehe Abschnitt 3.4.

Die verwendete Inspektionsmethode entdeckt schwerpunktmäßig mehr dem Bereich der *User Experience* zuordenbaren Probleme, was in Tabelle 4.4 deutlich sichtbar wird. Hier entfallen ca. 69,29% der Beobachtungen auf dieses Feld, während der Rest auf die Bereiche *Datenqualität* und genereller *Stand der Technik* entfallen.

DIE DATENQUALITÄTSASPEKTE anhand von Screenshots zu bewerten, gestaltet sich als schwierig, da hier primär Indizien angeführt werden können, welche sich auf die *Datenqualität*, einen zentralen Aspekt des DATENATLAS, auswirken.

Die Indizien bieten hier jedoch wenig Interpretationsspielraum und lassen sich im Wesentlichen auf die *„Freitext-Problematik“*, d.h. die sehr häufige Verwendung unreglementierter Freitextfelder, die weder auf *kontrollierte Vokabulare* oder *Thesauri* setzen, zurückführen. Daher muss eine tiefergehende Überprüfung am Live-Systeme erfolgen, wenn der DATENATLAS produktiv eingesetzt werden soll.

Freitext-Problematik

Aktuell entfallen allein 7 von 17, d.h. ca. 41%, die *Datenqualität* betreffende Probleme auf die *„Freitext-Problematik“*.

5
Bewertung und Desiderata

Die in diesem Kapitel vorgestellten Desiderata werden aus den Verbesserungspotentialen, die während der in Kapitel 4 beschriebenen User Journeys ausgemacht wurden, im Abgleich mit dem *Stand der Technik*[1] abgeleitet.

[1] Siehe Kapitel 3.

Ein Implementierungsniveau auf dem Stand der Technik kann dem Datenatlas auf dieser Analysegrundlage, wenn auch mit Abstrichen in der *Usability*, nur bei der Gestaltung der Nutzerschnittstelle (UI) bescheinigt werden. Diese wirkt zeitgemäß, wenngleich wichtige technische Eigenschaften wie die Barrierefreiheit nicht betrachtet werden konnten[2].

Die Bundesdruckerei bescheinigt sich jedoch selbst die Gewährleistung von Barrierefreiheit für den Datenatlas[3].

[2] Siehe Abschnitt 3.1; Limitationen des Gutachtens.

[3] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. `https://tinyurl.com/bdr-pm1`. Letzter Abruf: 21.07.2025

Funktional gesehen wird der Datenatlas dem Gestaltungsniveau des User Interfaces nicht gerecht und bietet viele Ansatzpunkte, in denen Verbesserungen Voraussetzung dafür sind, den *Stand der Technik* erreichen zu können.

Für die Bewertung, welche Desiderata praktisch umgesetzt werden sollten, sei zuvor auf Abschnitt 6.3 verwiesen, welcher konkrete Handlungsempfehlungen beinhaltet.

5.1 Technische Einordnung des Datenatlas

Auch wenn sich der Datenatlas nach Eigendarstellung als *Metadaten-Portal* u.a. in Ergänzung zu GovData sieht[4], so handelt es sich aus technischer Sicht zweifelsfrei um ein *Repository-System* für die interne Nutzung in der Bundesverwaltung.

[4] Ebenda.

Die vorgestellte Kurzevaluierung zeigt erhebliche Defizite des Datenatlas auf, so dass die Mitte Juli 2025 bereitgestellten Funktionen nicht auf Augenhöhe mit denen der in Abschnitt 3.2 vorgestellten Systemen bzw. dem *Stand der Technik* zu sehen sind.

Folglich ergibt sich ein recht rudimentär umgesetztes *Repository-System*, welches auch die *minimalen Use Cases der Bundesverwaltung*[5] nur in Teilen umzusetzen vermag.

[5] Siehe Kapitel 2.

Dem gegenüber steht die zeitgemäße Gestaltung der UI, welche

sich wohl auf die Verwendung des Matine-Frameworks zurückführen lässt[6]. Diese Aussage kann als gesichert gelten, da der Autor des Gutachtens Einblick in den Quellcode der Webseiten des DATENATLAS nehmen konnte.

Weitere Aussagen zur technischen Umsetzung lassen sich aus Beobachtungen nur indirekt herleiten.

Eine im Kontext der Lehrtätigkeit des Autors gestellte Anfrage, die hier ggf. zu mehr Erkenntnissen hätte führen können, lief ins Leere[7].

[6] Hierbei handelt es sich um eine React-Komponenten-Bibliothek, welche die Programmierung von zeitgemäßen Webseiten vereinfacht (`https://mantine.dev`; Letzter Abruf: 01.08.2025).

[7] Siehe Anhang A.5.

5.2 *Informationsrecherche – Desiderata*

In Abschnitt 5.1 wurde der DATENATLAS als rudimentäres *Repository-System* klassifiziert, welches nur in wenigen Aspekten dem *Stand der Technik* im Bereich der *Informationsrecherche* entspricht.

Informationssuchstrategien

Der DATENATLAS unterstützt *fast ausschließlich* gerichtete Suchstrategien mittels eines *Exact Matchings* von Anfragen.

Wie bereits in Abschnitt 3.5 ausführlich dargelegt wurde, setzt dies die Annahme voraus, dass die Nutzenden korrekte Anfragen formulieren können bzw. den Datenbestand kennen (siehe Abb. 5.1).

Dass die erste Annahme für einen Großteil der Beschäftigten der BUNDESVERWALTUNG nicht zutreffen wird, wurde in Abschnitt 3.5 und Kapitel 4 sowohl auf Basis der Erkenntnisse aus umfangreichen, wissenschaftlichen Studien als auch praktisch gezeigt.

Würden die Nutzenden den Datenbestand genau kennen, so ist es durchaus vorstellbar[8], dass sie bereits über die teilweise sehr spezifischen Datensätze, z.B. auf einem Abteilungslaufwerk, verfügen.

Es stellt sich also die Frage, warum keine weiteren *Informationssuchstrategien* im DATENATLAS für die Datenrecherche realisiert wurden, obwohl dies seit vielen Jahrzehnten als *best practice* gilt.

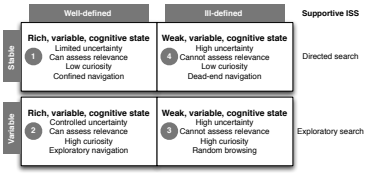


Abbildung 5.1: Matrix der vier intrinsischen Extremfälle an Informationsbedürfnis-Ausprägungen nach Ingwersen (1996)

[8] Siehe dazu auch die Desiderata in Abschnitt 5.3.

POSITIV ANZUMERKEN IST, dass der DATENATLAS zumindest die Filterung der Ergebnismenge ermöglicht, um die Suchergebnisse anhand einiger, weniger Anforderungen der Nutzenden einzuschränken.

Eine nachvollziehbare Relevanz-Sortierung, optimalerweise inkl. Highlighting der Treffer, wäre *Stand der Technik*, steht jedoch nicht zur Verfügung.

Leider geschieht die Umsetzung der Filterung nicht dem *Stand der Technik* folgend mittels der facettierten Suche[9], so dass Nutzende nicht voraussehen können, ob ihre Filterung überhaupt zu Ergebnissen (siehe Abb. 4.11) führen wird[10].

Verschärfend kommt hinzu, dass keine *Plausibilitätsprüfung* der Filtereinstellungen implementiert ist, was frustrierende Fehlbedie-

[9] Siehe Abschnitt 3.5.

[10] Die daraus resultierenden Auswirkungen auf die Usability werden in Abschnitt 5.3 vorgestellt.

nungen des DATENATLAS begünstigt.

Durch den Rückgriff auf *Linked (Open) Data*-Technologien wäre es ohne großen Aufwand möglich gewesen, die Hierarchie der Bundesverwaltung zu modellieren und aktuell zu halten. Diese Aspekte, die auch Auswirkungen auf die Usability des DATENATLAS hat, werden in Abschnitt 5.4 diskutiert.

Die iterative Verfeinerung im Rahmen einer explorativen Suche wird so im Ergebnis erheblich erschwert – wenn nicht sogar aufgrund des hohen Frustpotentials verhindert.

Gepaart mit dem *Exact Matching* des DATENATLAS führt dies aktuell dazu, dass nur der Use Case der *Known-Item-Search*, also die Suche nach bereits bekannten Dokumenten, für Anwendende effektiv nutzbar ist.

Wie der folgende Abschnitt zeigt, geschieht dies jedoch aus diversen Gründen nicht auf dem *Stand der Technik*.

DIE ENTSCHEIDUNG, *keine* EXPLORATIVEN Suchstrategien für einen Großteil der Nutzenden zu implementieren, steht im direkten Widerspruch zur eigenen Implementierungsskizze der BUNDESDRUCKEREI, welche u.a. einen Metadaten-Browser[11] darstellt.

Sie widerspricht aucj der zweiten Pressemeldung, wonach ein Ziel des DATENATLAS die „ressortübergreifende Suche und Exploration der Datenbestände“[12] ist.

[11] Siehe Abb. 2.1 bzw. Bundesdruckerei. *Erstes Vollständiges Datenmodell Der Bundesverwaltung - Pressemitteilung*, 2022. https://tinyurl.com/bdr-pm3. Letzter Abruf: 21.07.2025

[12] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025

DER DATENIMPORT, welcher sich an Expertinnen und Experten richtet,[13] unterstützt einen explorativen Suchzugang.

Gerade bei dieser Nutzendengruppe kann davon ausgegangen werden, dass sie in der Lage wäre auch eine *Known-Item-Search* durchzuführen, zumal das entsprechende Modul auch so angelegt ist, dass es die Suche in Behörden-eigenen Datensätzen unterstützten soll (siehe Abb. 4.20).

Warum gerade für diese Gruppe nicht zusätzlich die davor angebotene *Gerichtete Suche* implementiert wurde, ist kaum nachvollziehbar.

[13] Siehe Abschnitt 4.5.

Anfragemöglichkeiten und -ergebnisaufbereitung

In Abschnitt 3.2 wird ein *Mindestmaß der Anfrage-Formulierungsmöglichkeiten* benannt, die seit mehr als zwei Jahrzehnten bei der *Informationsrecherche Stand der Technik* sind.

Tabelle 5.1 fasst diese Möglichkeiten, ergänzt um *Exact* bzw. *Best Matching* während der Relevanzbewertung[14], zusammen und erlaubt den direkten Vergleich des Funktionsumfangs der in Abschnitt 3.2 vorgestellten Systeme mit dem DATENATLAS, um diesen einfacher bezüglich seines Funktionsumfangs bewerten zu können.

Die Tabelle zeigt deutlich, dass der DATENATLAS trotz seines jungen Alters bezüglich seines Funktionsumfangs im Vergleich sehr deutlich zurückfällt.

Rein funktional gesehen platziert sich das älteste System aus

[14] Siehe Abschnitt 3.3.

	StabiKat classic	*StabiKat*	*HoWeR*	*CrossAsia*	DATENATLAS
Veröffentlichung	1998	2010	2010☆	2016	2025
Systemklasse	OPAC	Discovery	Discovery	Repository	Repository
Anfragetyp					
Suchterme	✓	✓	✓	✓	✓
Boolesche Operatoren	✓	✓	✓	✓	–
Wildcard-Operatoren	✓	✓	✓	✓	–
Proximity-Operatoren	✓	✓	✓	✓	–
Unscharfe Suche	✓	✓	✓	✓	–
In-/Exklusion von Begriffen	✓	✓	✓	✓	–
Definition von Suchbegriffen für einzelne Metadaten-Felder	✓	✓	✓	✓	–
Phrasensuche	✓	✓	✓	✓	–
Exact Matching	✓	(✓)	(✓)	(✓)	✓
Best Matching	(✓)	✓	✓	✓	–

Geklammerte Häkchen geben das sekundäre Matching-Paradigma der *Information-Retrieval-Systeme* an, welches explizit durch Nutzende, z.B. durch die Selektion geeigneter Operatoren, ausgewählt werden muss.
☆ Grundlage ist der EBSCO Discovery Service

Tabelle 5.1: Mindestmaß der Anfrage-Formulierungsmöglichkeiten (nach Systemen)

dem Jahr 1998 deutlich vor dem DATENATLAS und unterstützt außerdem noch weitere Matching-Paradigmen.

Für ein *27 Jahre jüngeres* System wie den DATENATLAS gibt es keine Sachgründe, sich derart weit vom *Stand der Technik* zu entfernen, zumal es sich bei *Information-Retrieval-Systemen* nicht um ein exotisches Aufgabengebiet innerhalb der Informatik handelt, wie in Abschnitt 3.2 anhand einiger repräsentativer Beispiele dargelegt wurde.

DIE UNTERSTÜTZUNG DER ANFRAGEFORMULIERUNG, wie sie z.B. bei HoWeR (siehe Abb. 3.4(b)) mittels einer Autovervollständigung anhand der vorhandenen Datenbasis bzw. der Suchverläufe anderer Suchender erfolgt, wird durch den DATENATLAS nicht unterstützt.

SUCHERGEBNISSE WERDEN UNGEORDNET durch den DATENATLAS bereitgestellt. Im Nachgang kann eine Filterung durch Nutzende erfolgen (s.o.).

Je nach Anfrage oder Filterung verhindert der DATENATLAS nicht das Auftreten leerer Ergebnislisten bzw. federt deren Auswirkung auf die Nutzendenzufriedenheit durch Kontext-sensitive Empfehlungen ab, wie es beispielsweise der StabiKat classic (siehe Abb. 3.2 (a)) tut.

Die Sortierung von Ergebnissen ist im DATENATLAS prinzipiell möglich. Es werden jedoch nur elementare Sortierungen, d.h. eine alphabetische Sortierung nach dem Titel der Datensätze oder dem Veröffentlichungsdatum, angeboten. Beide Sortierungen sind auf-

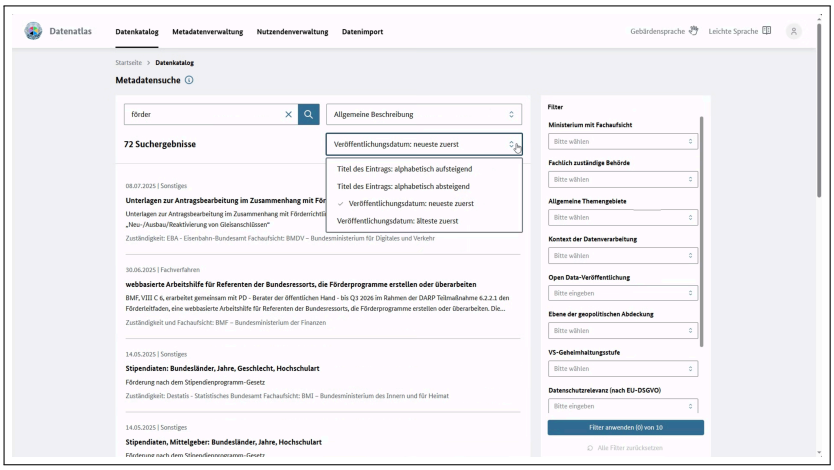


Abbildung 5.2: Trefferliste zum Suchbegriff „förder“; Sortierkriterien

und absteigend möglich (siehe Abb. 5.2).

Ein eigentlich zu erwartendes Relevanz-Ranking nach dem *Stand der Technik*[15] oder die Hervorhebung von Treffern des Suchbegriffs in den Datensätzen ist nicht vorhanden.

[15] Siehe Abschnitt 3.3.

Komplexere Anfragen, wie die Recherche nach Datensätzen aus den ehemaligen Ost-Bundesländern oder ähnliche, strukturelle Anfragen stehen im Datenatlas durch die Vernachlässigung von Technologien aus dem *Linked (Open) Data*-Umfeld nicht zur Verfügung, wie detailliert in Abschnitt 5.4 dargestellt wird.

Die Verwendung gängiger Techniken wie die Anfrageerweiterung mittels Synonymen oder die Einbeziehung verwandter Suchen konnte nicht ermittelt werden.

Da die genannten Mängel allerdings so grundlegend sind, scheint es wahrscheinlich, dass diese aufgrund der bereits benannten basalen Mängel ebenfalls nicht implementiert wurden.

Zwischenfazit

Die Beobachtung dieser teils eklatanten Mängel im Bereich der Anfragemöglichkeiten und -ergebnisaufbereitung für ein System, welches der *Informationsrecherche* dienen soll, ist für den Autor des Gutachtens trotz seiner Berufserfahrung überraschend, da fast alle in Tabelle 5.1 gelisteten Funktionen *spätestens* seit dem Jahr 2006 in Apache Lucene 1.9 (wahrscheinlich mit einer Ausnahme[16]) verfügbar sind.

Das Lucene-Paket bildet, natürlich in seiner aktuellen Form, welches alle der in Tabelle 5.1 genannten Funktionen bietet, die Basis moderner *Information-Retrieval-Systeme* bzw. Suchmaschinen wie Elasticsearch oder Solr, welche durchaus als arrivierte technologische Lösungen mit *weltweiter Verbreitung und Bekanntheit* angesehen werden können.

[16] Die Bereitstellung von Proximity-Operatoren konnte bei der kursorischen Prüfung des Quellcodes nicht sicher festgestellt werden. Eine Ermittlung der erstmaligen Bereitstellung der anderen Anfragetypen auf Grundlage des im Online-Archiv des Apache Lucene-Projekts bereitgestellten Quellcodes ist jedoch nur bis zu dieser Version möglich(`https://t1p.de/nboho`; Letzter Abruf: 31.07.2025). Im Quellcode finden sich weitere Hinweise, dass dieses Funktionen teilweise bereits 2004 implementiert wurden.

Schliesslich liegt aufgrund der sehr grundlegenden Abweichungen vom *Stand der Technik* die *Vermutung* nahe, dass bereits beim technischen Entwurf des Datenatlas teilweise ungeeignete Software-Komponenten für die Umsetzung der minimalen Use Cases[17] ausgewählt wurden.

[17] Siehe Kapitel 2.

Es ist anzunehmen, dass die Datenhaltung des Datenatlas in einer relationalen Datenbank erfolgt, was *per se* eine nachvollziehbare Entscheidung darstellt, die jedoch mit Hinblick auf *Linked (Open) Data* Folgeprobleme erzeugt. Diese Probleme werden in Abschnitt 5.4 gesondert beschrieben.

Offenbar wird jedoch auch der Retrieval-Prozess – also die Anfrageverarbeitung – innerhalb dieses *Datenbankmanagement-Systems* realisiert. Bei den eigentlich leicht zu ermittelnden Anforderungen an die *Informationsrecherche*[18] bleibt unklar, warum nicht, wie bereits in Abschnitt 2.2 beschrieben, auf die Verbindung mit einem *Information-Retrieval-System* gesetzt wurde, um die Nutzenden bestmöglich bei der *Informationsrecherche* unterstützen zu können.

[18] Siehe Abschnitt 5.3.

Die Hypothese wird zudem durch die Charakteristik der Ergebnismenge gestützt, welche offensichtlich durch ein *Exact Matching* entsteht. Auch wenn Abb. 5.2 das Vorliegen einer geordneten Liste suggeriert, scheint es sich jedoch eher um eine Multimenge zu handeln, wie sie für relationale *Datenbankmanagement-Systeme*[19] typisch ist.

[19] Siehe Abschnitt 2.2.

Gestärkt wird dieser Eindruck durch die angebotenen Sortierkriterien, die in Abb. 5.2 zu sehen sind. Diese lassen sich trivial in SQL umsetzen, wie die folgenden Befehle illustrieren:

```
1 SELECT * FROM datenatlas WHERE titel LIKE 'förder%';
2 SELECT * FROM datenatlas WHERE titel LIKE 'förder%' ORDER BY title;
3 SELECT * FROM datenatlas WHERE titel LIKE 'förder%' ORDER BY title DESC;
```

Das SQL-Statement in Zeile 1 fragt sämtliche Datensätze in der fiktiven Tabelle `datenatlas`[20] ab, deren `titel` mit der Buchstabenfolge „förder“ beginnt und auf die dann Null oder beliebig viele Zeichen mehr folgen. Dies entspricht der ungeordneten Ergebnismenge und der Suchanfrage, die in Abb. 5.2 dargestellt ist.

[20] Natürlich lassen sich diese Interna des Datenatlas nicht mittels der Screenshots ermitteln. Allerdings ist anzunehmen, dass das Retrieval beim Datenatlas prinzipiell gleich abläuft.

Die aufsteigende Sortierung nach dem Titel-Feld ist in Zeile 2 zu sehen. Zeile 3 stellt die Anfrage mit absteigender Reihenfolge dar. Der Parameter `ASC` kann in Zeile 2 entfallen, da SQL standardmäßig aufsteigend sortiert. Dieses Beispiel lässt sich einfach auf das Veröffentlichungsdatum übertragen.

Eine Relevanz-Sortierung wird, wie gesagt, nicht angeboten und wäre aufgrund des zugrundeliegenden Booleschen Retrievalmodells auch mathematisch unmöglich[21,22].

[21] Siehe Abschnitt 3.3.

[22] Welche zusätzlichen Arbeiten an einem *relationalen Datenbankmanagement-System* erfolgen müssen, um diese Limitation zu überwinden, wird in Anhang A.2 skizziert.

Hinzu kommen weitere negative Auswirkungen auf die Retrievalqualität, da die oben genannten Systeme per se wenig für die Verarbeitung natürlicher Sprache geeignet sind und häufig gängige Mechanismen wie Stemming[23] nicht direkt unterstützen.

[23] Siehe Abschnitt 2.2.

Die Abschaltung der Relevanz-Sortierung bei einem Apache Lucene-basierten System (oder jedem anderen *Information-*

Retrieval-System) ist ungleich komplizierter, da das Matching eine Kernfunktionalität eines jeden *Information-Retrieval-Systems* darstellt[24]. Es ist äußerst unwahrscheinlich, dass eine qualifizierte Programmiererin oder ein Programmierer sich entscheidet, ein *Information-Retrieval-System* einzusetzen, um dann seine Kernfunktionen zu deaktivieren.

[24] Siehe Abschnitt 2.2.

ABSCHLIESSEND LÄSST SICH diese Frage jedoch nur durch den interaktiven Zugriff auf den DATENATLAS beantworten, wenngleich die präsentierten Indizien wenig Interpretationsspielraum bieten. Die Entscheidung für die direkte Kombination eines relationalen DBMS mit einer Website zur Steuerung wäre in jedem Fall suboptimal, um die minimalen Use Cases des DATENATLAS nutzerfreundlich umzusetzen.

Verbesserung der technischen Umsetzung des Datenatlas

Aufgrund der oben aufgezeigten Mängel, die sich bereits aus der kursorischen Prüfung im Zeitumfang von ca. 30 Minuten ergeben, muss kritisch geprüft werden, inwiefern der DATENATLAS technisch verbessert werden kann.

Die Implementierung eines *Best Matchings* ist zwingend, damit mehr Use Cases als die *Known-Item-Search* mit dem DATENATLAS realisiert werden können und dieser damit auf höhere Nutzungsakzeptanz trifft.

Es liegt auf der Hand, dass die anzunehmende Zielgruppe den DATENATLAS auch ohne Vorkenntnisse zur Befriedigung ihres *Informationsbedürfnisses* nutzen können muss.

DIES HÄTTE ZUR FOLGE, dass der Systemkern des DATENATLAS – zumindest die Funktionen, die unmittelbar mit dem Retrieval-Prozess verbunden sind – ausgetauscht werden müssen.

Hierbei ist anzunehmen, dass dieser Schritt weitere Auswirkungen auf das User Interface hat und damit Folgerisiken beinhaltet.

Die Sinnhaftigkeit dieses Refactorings muss unter Einbeziehung der in Abschnitt 5.8 präsentierten Punkte bewertet werden, zumal weitere als die bisher diskutierten Mängel bestehen.

GENERELL STELLT SICH DIE FRAGE, warum nicht bereits von Vornherein geeignete, bestehende Software-Komponenten für die Implementierung des DATENATLAS nachgenutzt wurden.

Wie bereits in Abschnitt 3.2 erwähnt, existieren im Bibliotheks- und Archivsektor seit vielen Jahrzehnten vergleichbare Anforderungen[25] wie die des DATENATLAS, welche dort mittels Software-Systemen erfolgreich umgesetzt wurden.

[25] Siehe Seite 20.

Die gleiche Aussage lässt sich auf den Bereich *Open Data* übertragen. Hier liegen sogar diverse Erfahrungen aus den Ländern und auf Bundesebene in Form von GOVDATA vor.

In diesem Abschnitt sollen einige dieser Software-Komponenten

vorgestellt werden.

Aus der Perspektive Recherchierender könnte die Unterstützung verschiedener *Informationssuchstrategien* durch die Verwendung eines bereits bestehenden *Open Source-Discovery-Systems*, wie z.B. VuFind[26] gewährleistet werden.

Die *minimalen Use Cases*[27] umfassen jedoch auch die Erfassung von Metadaten, welche typischerweise als Kernfunktion in *Repository-Systeme* integriert ist.

[26] https://vufind.org/vufind/; Letzter Abruf: 21.07.2025.

[27] Siehe Kapitel 2.

Die prinzipielle Funktionsweise von Repository-Systemen wird stellvertretend am Beispiel von DSpace[28] in Abb. 5.3 dargestellt.

Weitere *Repository-Systeme* und mit ihnen verbundene Komponenten werden im Nachgang vorgestellt.

[28] Ein *Open Source-Repository-System* mit üblichem Funktionsumfang auf dem *Stand der Technik*.

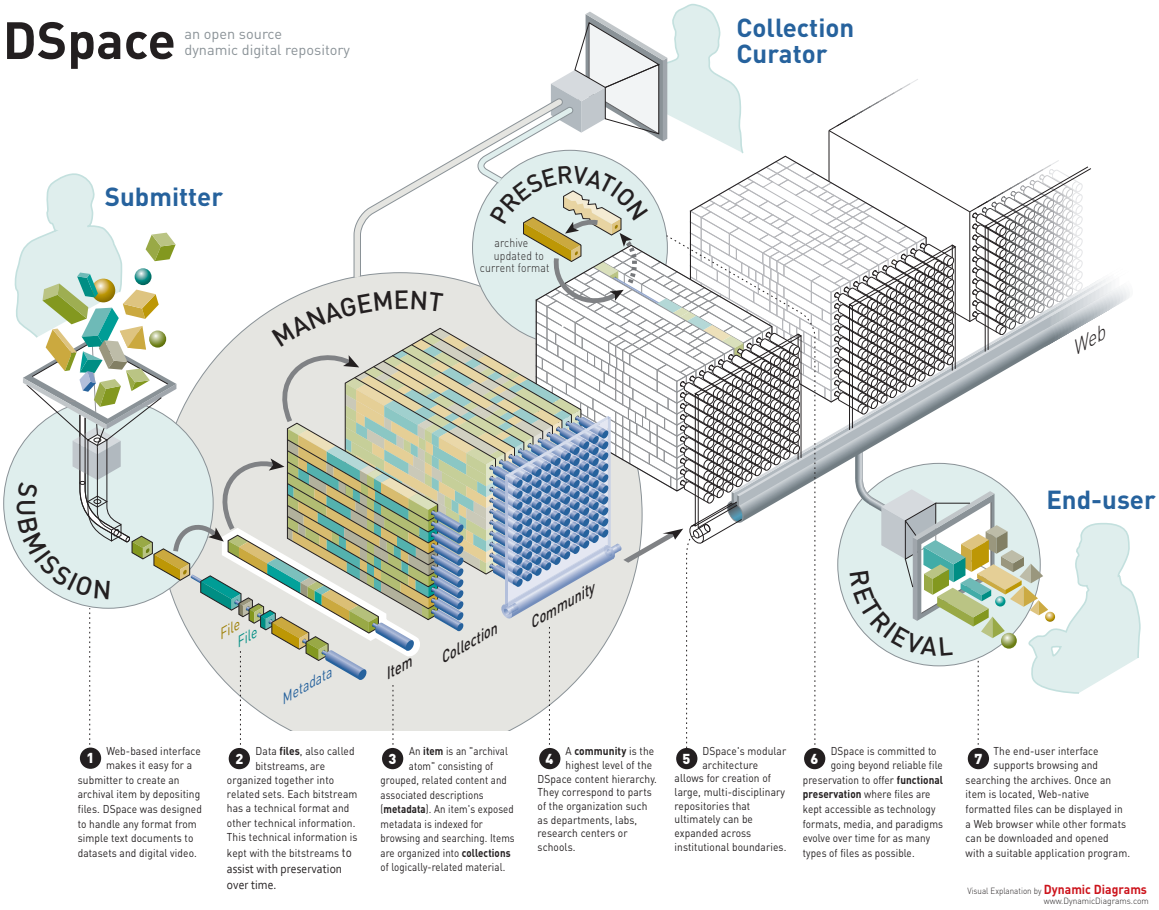


Abbildung 5.3: Prinzipielle Funktionsweise eines Repositorys; https://duraspace.org/wp-content/uploads/dspace-files/DSpace_Diagram.pdf; Letzter Abruf: 25.07.2025 Original nicht länger verfügbar – Rekonstruktion mittels der Wayback Machine am 25.07.2025

Aus der Abbildung wird deutlich, dass die Interaktion mit *Repository-Systemen* i.d.R. mittels drei Rollen geschieht, welche Nutzende einnehmen können:

Submitter legen Dokumente im Repository ab und erfassen zugehörige Metadaten.

Im Rahmen des Datenatlas sind dies Mitarbeitende der obersten und nachgeordneten Bundesbehörden.

Collection Curator kuratieren Datensätze und fassen diese z.B. nach Verwendungszwecken zusammen, definieren Export- und Interoperabilitätsmöglichkeiten und legen Archivierungsansätze etc. fest.

Diese Rolle wird beim DATENATLAS voraussichtlich durch Mitarbeitende der Datenlabore der obersten Bundesbehörden wahrgenommen.

End-User umfassen alle Mitarbeitenden der BUNDESVERWALTUNG, welche den DATENATLAS zur Recherche nutzen.

Durch die Unterstützung dieser Rollen sind bereits die *minimalen Use Cases*[29] für den DATENATLAS vollständig abgedeckt.

[29] Siehe Kapitel 2.

Funktionsumfang von Repository-Systemen Darüber hinausgehend bieten *Repository-Systeme* eine Vielzahl weiterer Funktionen, die nach Kenntnis des Autors für den DATENATLAS nicht bzw. unnötigerweise neu implementiert wurden, wie z.B. ein Rechte-Rollen-Konzept.

Beispiele an typischen Repository-Funktionen sind Kernfunktionen wie der Dokumenten-Upload, der nach Kenntnis des Autors nicht Teil des Entwicklungsauftrags des DATENATLAS ist, oder die Bereitstellung einer Schnittstelle zur *Langzeitarchivierung*, deren Bedarf sich rechtlich begründen lässt und die deshalb separat in Abschnitt 5.7 betrachtet wird.

Typische Funktionsumfänge von *Repository-Systemen* können den ursprünglichen Entwicklungszielen von FEDORA[30] vom Ende der 1990er-Jahre stellvertretend für andere Systeme entnommen werden:

[30] Ein weiteres *Open Source-Repository-System*; https://fedorarepository.org/core-attributes-of-fedora-repository-enable-complex-modeling-of-data-and-objects-for-re-use-in-a-wide-variety-of-instances/; Letzter Abruf: 01.08.2025

> *Identifiers:* provision of persistent identifiers; unique names for all resources without respect for machine address
>
> *Relationships:* support for relationships between objects
>
> *Tame Content:* normalization of heterogeneous content and metadata based on an extensible object model
>
> *Integrated Management:* efficient management by repository administrators not only of the data and metadata in a repository, but also of the supporting programs, services and tools that make presentation of that data and metadata possible
>
> *Interoperable Access:* provision of interoperable access by means of a standard protocol to information about objects and for access to object content; discovery and execution of extensible service operations for digital objects
>
> *Scalability:* provision of support for >10 million objects
>
> *Security:* provision of flexible authentication and policy enforcement
>
> *Preservation:*[31] provision of features to support longevity and archiving, including text-based serialization of objects and content versioning
>
> *Content Recon:* reuse of objects including object content being present in any number of contexts within a repository; repurposing of objects allowing dynamic content transformations to fit new presentations requirements[32]

[31] Langzeitarchivierung; siehe Abschnitt 5.7.

[32] Eine Übersetzung dieser Entwicklungsziele findet sich in Anhang A.3.

Die Liste dieser Kernfunktionen macht deutlich, dass sich diese nicht nur auf das Frontend, d.h. die *Informationsrecherche*, beschränken, sondern alle Bedarfe des Betriebs abdecken.

Üblich ist es vielmehr, dass moderne *Repository-Systeme* von Haus aus Sicherheits- bzw. Zugriffssteuerungskonzepte, interoperable Schnittstellen, *Persistent Identifier* und Funktionen zur Anbindung an Backup- und Langzeitarchivierungssysteme implementieren.

Zur Umsetzung dieser Kernfunktionen unterstützen *Repository-Systeme* i.d.R. vielfältige Möglichkeiten zur Anbindung diverser relationaler *Datenbankmanagement-Systeme*, wie z.B. PostgreSQL oder MySQL, die häufig mit *Information-Retrieval-Systemen* wie Solr oder Elasticsearch zur Vebesserung der Retrieval-Funktionalitäten verbunden werden können. Hinzu kommen häufig Triplestores zur Unterstützung von *Linked (Open) Data*.

Abbildung 5.4 illustriert diesen modularen Aufbau am Beispiel von Fedora.

Durch die modulare Systemarchitektur wird es möglich, einzelne Komponenten des Gesamtsystems zu ersetzen oder zu ergänzen, wenn dies aus Betriebsgründen nötig wird.

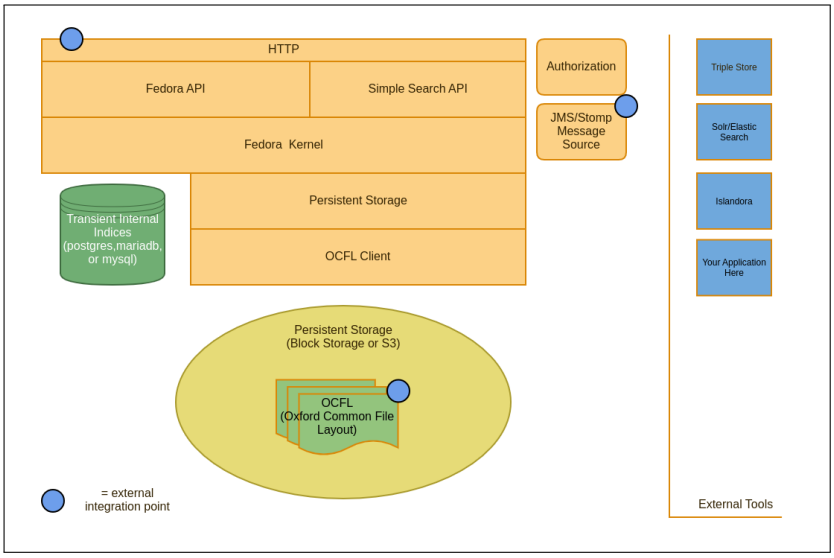


Abbildung 5.4: Interne Repository-Systemarchitektur am Beispiel von Fedora 6.x CC BY LYRASIS https://wiki.lyrasis.org/display/FEDORA6x

Tabelle 5.2 listet eine kleine Auswahl gängiger *Open Source-Repository-Systeme* mitsamt typischer Funktionen – bzw. Desiderata für den Datenatlas – und durch diese genutzte Datenbank- bzw. Information-Retrieval-Systeme inklusive deren Erscheinungsjahr[33] auf.

Die Tabelle kann als repräsentativ für den *Stand der Technik* in diesem Bereich gesehen werden, da sie einen Zeitraum von zwei Jahrzehnten der *Repository-System*-Entwicklung umfasst.

Sie erhebt jedoch keinen Anspruch auf Vollständigkeit, da dies den Umfang des Gutachtens bei weitem sprengen würde.

[33] Das Alter der Systeme wurde anhand der jeweiligen Projektseite oder des passenden Wikipedia-Artikel, wenn verfügbar, ermittelt. Lagen diese Informationen nicht vor, wurde das Quellcode-Repository geclont und das Alter manuell mittels `git log -reverse` über den ältesten Log-Eintrag bestimmt. Die Systeme können deshalb ggf. noch älter sein, wenn z.B. die Code-Versionierung von SVN auf Git umgestellt wurde.

Name	Seit...	Unterstützt, u.a. ...	URL
Repository-Systeme			
FEDORA	1998	Bietet größte Flexibilität und umfangreichsten Funktionsumfang: Support für verschiedene Datastores (u.a. POSTGRESQL, div. Triplestores). Suchmaschinen (i.d.R. SOLR); Dateimanagement, Zugriffskontollsteuerung, *Persistent Identifier*, Betriebssicherheit (u.a. Disaster Recovery, Data Integrity), *Linked (Open) Data* u.v.m.★	https://fedorarepository.org
DSPACE	2002	Typischer Funktionsumfang eines *Repository-Systems* (s.o.), Unterstützung div. Datastores, Suche mittels SOLR, *Persistent Identifier*	https://dspace.org
CKAN	2007	Typischer Funktionsumfang eines *Repository-Systems* (s.o.), Suche mittels SOLR, Datastore in POSTGRESQL; Geo-Daten, Unterstützung einer Vielzahl an Community-Erweiterungen, z.B. für die direkte Verarbeitung von DCAT-AP oder DOI	https://ckan.org
ZENODO	2011	Spezialisierung auf Forschungsdaten und Publikationen, Betrieb durch das CERN, u.a. Basis für „EU Open Research Repository“✡; Datastore in Kombination aus POSTGRESQLund REDIS, Suche mittels ELASTICSEARCH, *Persistent Identifier*	https://github.com/zenodo/
PIVEAU	2018	Spezialisierung auf den Bereich der ÖFFENTLICHEN VERWALTUNG mit integrierter DCAT-AP- und *Linked (Open) Data*-Unterstützung, Datastore im Virtuoso-Triplestore, Suche mittels ELASTICSEARCH, *Persistent Identifier*; u.a. Basis des Europäischen Datenportals✣ und der *Open Data*-Portale Bayerns und Brandenburgs	https://www.piveau.de
Information-Retrieval-Systeme			
APACHE LUCENE	2000	Alle in Tabelle 5.1 genannten Funktionen plus Dokumentenindizierung.	https://lucene.apache.org
SOLR	2004☆	Basiert auf APACHE LUCENE, ergänzt um div. Webzugriffs- und Managementfunktionen	https://solr.apache.org
ELASTICSEARCH	2010	Basiert auf APACHE LUCENE, ergänzt um div. Webzugriffs- und Managementfunktionen	https://www.elastic.co
Relationale Datenbankmanagement-Systeme			
POSTGRESQL	1995	Relationales DBMS mit höchster SQL-Standardkonformität und größtem Funktionsumfang, u.a. Support für nicht-relationale bzw. semi-strukturierte Daten (JSON, XML), erweiterbar um Geo-Daten†- oder RDF✦-Unterstützung	https://www.postgresql.org
MYSQL	1995	Relationales DBMS mit weitgehender Unterstützung des SQL-Standards, grundlegende Unterstützung von JSON und Geo-Daten	https://www.mysql.com
Sonstige Datenbankmanagement-Systeme (NoSQL)			
REDIS	2009	Key-Value-Store (In-Memory)	https://redis.io

★ https://fedorarepository.org/technical-specifications/; Letzter Abruf: 26.07.2025
✡ https://zenodo.org/communities/eu/; Letzter Abruf: 26.07.2025
✣ https://data.europa.eu/data/datasets; Letzter Abruf: 26.07.2025
☆ 2006 erfolgte die Übertragung an die *Apache Software Foundation (ASF)*, einer gemeinnützigen Stiftung zur Förderung von Open-Source-Entwicklungsprojekten, die auch u.a. die Weiterentwicklung von APACHE LUCENE koordiniert.
† https://postgis.net; Letzter Abruf: 31.07.2025 ✦ https://pgxn.org/dist/rdf_fdw/; Letzter Abruf: 22.05.2025

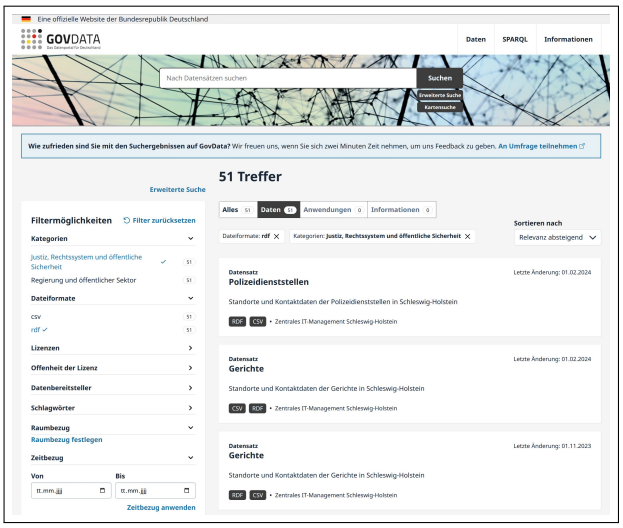
Tabelle 5.2: Auswahl aktiv weiterentwickelter Software-Komponenten zur *Informationsrecherche* unter Open-Source-Lizenz

Fazit Informationsrecherche

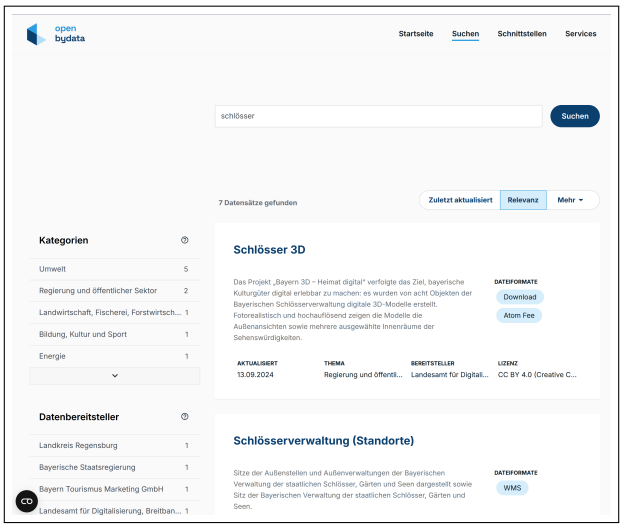
Betrachtet man das Umfeld des DATENATLAS, so zeigt sich eine Vielzahl vergleichbarer Systeme, die ähnliche Use Cases umsetzen und mittels oberflächlicher Suche über gängige Web-Suchmaschinen schnell zu finden sind.

Das bereits angesprochene *Open Data*-Portal Deutschlands – GOVDATA[34] – basiert auf CKAN und setzt eine Vielzahl der Desiderata im Bereich der *Informationsrecherche* um, wie auf Abb. 5.5 (a) zu erkennen ist.

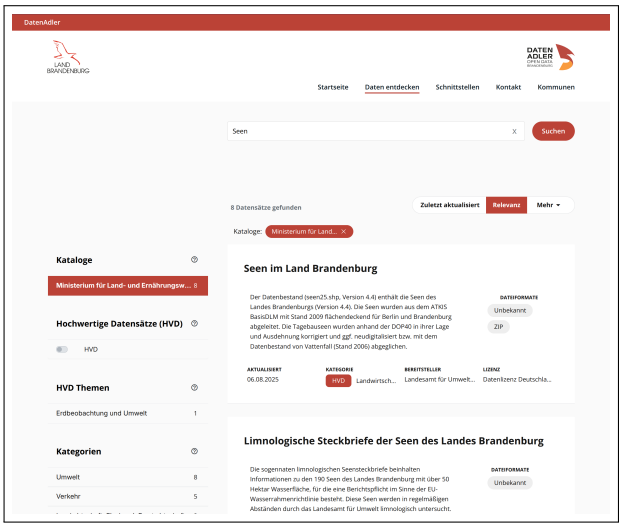
[34] https://www.govdata.de; Letzter Abruf: 21.07.2025



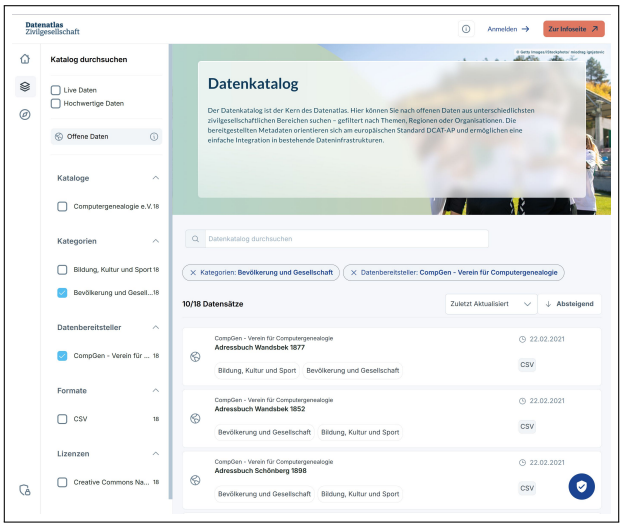
(a) GOVDATA; https://www.govdata.de; Letzter Abruf: 19.08.2025



(b) OPEN.BYDATA; https://open.bydata.de; Letzter Abruf: 19.08.2025



(c) DATENADLER; https://datenadler.de; Letzter Abruf: 19.08.2025



(d) Datenatlas Zivilgesellschaft; https://datenatlas-zivilgesellschaft.de; Letzter Abruf: 19.08.2025

Abbildung 5.5: Gegenüberstellung existierender *Repository-Systeme* der Verwaltung und Zivilgesellschaft

Das System bietet u.a. ein *Best Matching*, *Facettierte Navigation* und weitere Funktionen auf dem *Stand der Technik*.

Teilweise positioniert sich GOVDATA am oberen Ende des Stands der Technik, beispielsweise durch die Karten-basierte Suchfunktion.

Die übrigen in Abb. 5.5 (b-d) dargestellten *Repository-Systeme* basieren allesamt auf PIVEAU und reichen funktional nur beinahe

an GovData bzw. kaum an CrossAsia heran; sie sind jedoch auch wesentlich jünger.

Die Systeme eint, dass sie alle *Linked (Open) Data* unterstützen. Die Wichtigkeit der Unterstützung von *Linked (Open) Data* kann dadurch unterstrichen werden, dass selbst GovData mittelfristig zu Piveau wechseln wird, welches Triplestores nativ einbindet. Beim aktuell verwendeten CKAN geschieht dies nur mittels eines Plug-Ins. Dieser Migrationspfad wurde dem Gutachter durch den zuständigen Product-Owner bestätigt.

Bei CrossAsia mag der wesentlich größer Funktionsumfang im Vergleich mit den anderen Systemen nicht überraschen, da dieses *Repository-System* wesentlich mehr Anforderungen, u.a. im Volltextbereich, abdecken muss und Wissenschaftlerinnen und Wissenschaftler als Zielgruppe hat.

Diese Zielgruppe ist anspruchsvoller, denn sie muss anders als bei den anderen drei Systemen, welche sich an das Gros der Nutzerinnen und Nutzer richten, diverse Use Cases bedienen.

Nichtsdestotrotz setzen die anderen Systeme den *Stand der Technik* um und sollten als Mindestmaßstab für den Funktionsumfang des Datenatlas dienen.

Die enge Verwandtschaft zwischen den *Open Data*-Portalen der Länder Bayern und Brandenburg – Open.ByData und Datenadler – ist augenscheinlich.

Ursache hierfür ist, dass Brandenburg die User-Interface-Anpassungen Bayerns für das zugrundeliegende Piveau nachnutzt.

Auch das *Repository-System* der Zivilgesellschaft – der Datenatlas Zivilgesellschaft[35] – unter der Finanzierung der Bertelsmann-Stiftung setzt auf eine Erweiterung von Piveau: Civora UI[36], ein *Repository-System* mit Fokus auf die Integration von urbanen Daten aus GIS[37] oder dreidimensionalen Daten.

Alle vorgestellten *Repository-Systeme* liegen unter einer *Open Source*-Lizenz vor.

[35] https://datenatlas-zivilgesellschaft.de/; Letzter Abruf: 24.07.2025

[36] https://opencode.de/de/software/civora-ui-5560; Letzter Abruf: 18.08.2025

[37] Ein Geoinformationssystem ist ein Informationssystem zur Verwaltung geographischer Daten wie z.B. Karten.

5.3 *Menschzentrierte Entwicklung – Desiderata*

Laut Eigenaussage nutzt die Bundesdruckerei ein agiles Vorgehen zur Entwicklung des Datenatlas[38]. Diese Entscheidung entspricht dem *Stand der Technik*.

Typische Vorteile des agilen Vorgehens sind eine höhere Qualität der Software, eine bessere Integration von Nutzendenanforderungen sowie eine erhöhte Team-Produktivität, die sich in schnellen Produkt-Release-Zyklen niederschlägt[39].

Die während der Entwicklung erstellten und stetig überarbeiteten Prototypen helfen dem Entwicklungsteamm, den Nutzungskontext des entstehenden Produkts besser zu verstehen und sich besser über Anforderungen auszutauschen, u.a. zwischen der Entwicklung, dem Design-Team und den Endanwendenden[40].

Mit einer mehr als zwei Dekaden andauernden Entwicklungs-

[38] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025

[39] Matharu et al. (2015)

[40] Larrea et al. (2024)

geschichte sind agile Methoden mittlerweile in vielen Bereichen Standard und konnten wiederholt mit ihren oben genannten Vorzügen überzeugen – auch in der BUNDESVERWALTUNG[41].

IM KONTRAST DAZU steht das in Abschnitt 4.7 zusammengefasste Ergebnis der Evaluierung des DATENATLAS anhand dreier User Journeys, die gerade im Usability- und UX-Bereich erhebliche Mängel aufzeigen.

Dies ist mit Hinblick auf dessen sehr lange Entwicklungszeit, welche laut der Bundesdruckerei (2022) Ende 2021 begonnen hat, überraschend. Berücksichtigt man einen gewissen Projektvorlauf, so befindet sich der DATENATLAS sicherlich seit Mitte 2022 und damit, Stand Mitte Juli 2025, seit ca. drei Jahren in der Entwicklung.

OB DIE DATENLABORE mit ihrer fachlichen Expertise in die nutzerzentrierte Entwicklung des DATENATLAS einbezogen wurden, ist aufgrund der Aussagen einzelner Vertreter der Datenlabore nicht zu bestimmen, wie bereits in der Einleitung dargelegt wurde.

Bezüglich der individuellen Wahrnehmung der Beteiligung ergibt sich ein diverses Meinungsbild, welches bis hin zur Enttäuschung reicht. Ob eine strukturierte Erfassung der Use Cases oder Erhebung der wesentlichen User Requirements[42] erfolgte und wie es Stand der Technik wäre[43], ist dem Autor nicht bekannt.

Laut der Pressemitteilung[44] der BUNDESDRUCKEREI vom Oktober 2022 wird jedoch ein „nutzerzentrierter Datenatlas“ konzipiert, der einen „Meilenstein auf dem Weg zur datengetriebenen Verwaltung“ darstellen soll.

Da davon auszugehen ist, dass in den Datenlaboren „Information Professionals“ beschäftigt sind – also Personen mit einer hohen Expertise in Datenmanagement und -recherche, die um den *Stand der Technik* wissen müssen, erscheint es unglaubwürdig, dass nicht bereits bei frühesten Prototypen das Fehlen üblicher Suchmöglichkeiten, wie in Abschnitt 3.2 dargestellt, bemängelt wurde – gerade wenn dem DATENATLAS eine solch zentrale Bedeutung für die gesamte BUNDESVERWALTUNG zugedacht wird.

ÜBLICHERWEISE müssten außerdem konkrete Endanwenderinnen und Endanwender, welche die größte Zielgruppe des DATENATLAS darstellen, in die nutzendenzentrierte Entwicklung eines solchen Werkzeugs eingebunden worden sein. Dies dient der kontinuierlichen Überprüfung der Usability und der Nutzendenerwartung. Da der DATENATLAS zum Betrachtungszeitpunkt teils gravierende Usability-Probleme aufweist, erscheint dies unwahrscheinlich.

DURCH USABILITY-TESTS hätten eklatante Probleme, wie z.B. leere Ergebnislisten[45], frühzeitig erkannt werden müssen.

Inwiefern solche kostengünstigen und wirkungsvollen Methoden[46] auf dem *Stand der Technik* Verwendung fanden, ist dem Autor nicht bekannt.

[41] David Zellhöfer. Agilität und Weltkulturerbe: Erfahrungen mit sieben Jahren agilen Methoden an der Staatsbibliothek zu Berlin – Eine Fallstudie I/II. *ABI Technik*, 41(3):194–201, 2021. ISSN 2191-4664, 0720-6763. DOI: 10.1515/abitech-2021-0032. https://www.degruyter.com/document/doi/10.1515/abitech-2021-0032/html. Letzter Abruf: 16.04.2025

[42] *Nutzungsanforderungen*; d.h. alle konkreten Anforderungen der Nutzenden.

[43] Siehe u.a. Geis und Polkehn (2018); Sommerville (2003).

[44] Bundesdruckerei. *Erstes Vollständiges Datenmodell Der Bundesverwaltung - Pressemitteilung*, 2022. https://tinyurl.com/bdr-pm3. Letzter Abruf: 21.07.2025

[45] Siehe Abschnitt 4.1 Informationsrecherche VI – Schlagwort-basierte Suche: Verfeinerung der Filterung.

[46] Siehe Abschnitt 3.4.

Die Einbindung von Nutzendenfeedback ist sowohl Teil agiler Vorgehensmethoden, die im Rahmen des Entwicklungsprojekts angeblich Anwendung finden, als auch ein zentrales Element des menschzentrierten Gestaltungsprozesss nach DIN EN ISO 9241-210 (2020) oder des *Servicestandards*[47] – der in diesem Bereich einschlägigen Normen und Prozessmodelle. Aus Sicht des Gutachtens lässt der Funktionsstand des DATENATLAS Mitte Juli 2025 nur zwei Vermutungen zu:

[47] Siehe Abschnitt 3.4.

- entweder wurden kein Feedback Nutzender eingeholt oder
- die Umsetzung der Nutzendenanforderungen wurde seitens der Projektsteuerung ignoriert oder in ihrer Bedeutung verkannt.

Trifft der letztgenannte Punkt zu, wäre es die Aufgabe des technischen Dienstleisters gewesen, hier über den *Stand der Technik* aufzuklären, wie es z.B. die *Ethischen Leitlinien der Gesellschaft für Informatik*[48] fordern – zumal der *Stand der Technik*, wie in Kapitel 3 dargelegt, in diesem Bereich sehr leicht zu ermitteln ist.

[48] `https://gi.de/ueber-uns/organisation/unsere-ethischen-leitlinien/`; Letzter Abruf: 01.08.2025

HÄTTE MAN, wie in Abschnitt 5.2 argumentiert, bestehende Komponenten nachgenutzt, wären bestimmte Usability-Probleme, die auch teilweise Auswirkung auf die *Datenqualität*[49] des DATENATLAS haben, von vornherein vermieden worden und man hätte mit einem reiferen MVP in den Nutzendentest gehen können.

[49] Siehe Abschnitt 5.4.

GERADE DIE FREITEXT-PROBLEMATIK ist ein solches Problem[50], welches durch einige Usability-Desiderata unmittelbar Auswirkung auf die *Datenqualität* hat, wie beispielsweise Abschnitt 4.3 Metadatenverwaltung III – Eingabe Metadatensatz aufzeigt.

[50] Siehe Abschnitt 4.7.

Werden Freitext-Felder fehlerhaft oder unvollständig ausgefüllt, wirkt sich das sowohl auf die *Datenqualität* als auch die Wiederauffindbarkeit von Datensätzen mittels der primär durch den DATENATLAS umgesetzten *Gerichteten Suche* aus[51].

[51] Siehe Abschnitt 3.6.

Dabei ist anzumerken, dass sich die *„Freitext-Problematik"* nicht einfach auf Ebene der UI beheben lässt, sondern vor allem konzeptionelle Überlegungen auf Ebene der Datenmodellierung voraussetzt. Diese werden im Folgeabschnitt Datenqualität und -semantik – Desiderata separat adressiert.

Es finden sich weitere potenzielle Desiderata, welche durch eine Befragung der Nutzenden ermittelt werden könnten. So werden in Abschnitt 4.3 Metadatenverwaltung I – Übersicht Metadatenverwaltung ausschließlich Datensätze der eigenen Behörde gelistet, was die Durchführung der Fachaufsicht seitens der Ministerien oder den Support durch die dort angesiedelten Datenlabore erschwert. Das jeweilige Ministerium fungiert zwar als Administrator, kann jedoch nur veröffentlichte Datensätze der nachgeordneten Behörden sehen, was die technische Unterstützung unnötig erschwert.

Ebenso sind Warndialoge, wie der in Abschnitt 4.3 Metadatenverwaltung II – Neuanlage von Metadaten dargestellte, kaum rechtssicher durch Nutzende zu bestätigen. Dies liegt vor allem

daran, dass sie erhebliches externes und juristisches Fachwissen voraussetzen, um aus Sicht der Nutzenden sicher ausschließen zu können, dass „juristische Konsequenzen oder Vertragsstrafen entstehen".

Diese Art der Expertise seitens der Zielgruppe, den Beschäftigten der BUNDESVERWALTUNG in ihrer Breite, vorauszusetzen, erscheint gewagt. In jedem Fall müssten begleitend zur Einführung des DATENATLAS erhebliche Zusatzaufwände in den diesbezüglichen Kompetenzaufbau, z.B. in Form von Schulungen, investiert werden, wenn sich diese Risiken nicht sogar mittels Automatisierung bewerten lassen. Eine reine Übertragung des Risikos an die Nutzenden ohne Software-seitige Hilfestellung erscheint wenig zeitgemäß.

Fazit Menschzentrierte Entwicklung

Wie bereits in Abschnitt 5.2 dargelegt wurde, unterstützt der DATENATLAS im Wesentlichen die *Known-Item-Search* ohne *Best Matching* für den zentralen Bereich der *Informationsrecherche*, so dass aufgrund der Studienlage davon auszugehen ist, dass er am Bedarf der Nutzenden vorbei entwickelt wurde.

Es erscheint unplausibel, dass sich gerade die Mitarbeitenden der ÖFFENTLICHEN VERWALTUNG anders als der Rest der Bevölkerung verhalten, zumal etwas mehr als jeder zehnte Erwerbstätige in Deutschland in diesem Bereich arbeitet[52].

Selbst wenn die übrigen Usability-Probleme behoben werden können, ergibt sich aus Sicht des Autors aktuell kein erkennbarer Anwendungsfall für den DATENATLAS, welcher auf breite Nutzendenakzeptanz treffen wird. Denn es wird nur eine *Known-Item-Search* für die Mehrheit an Nutzenden auf Grundlage einer wenig gesicherten *Datenqualität* bereitgestellt.

[52] https://www.destatis.de/DE/Themen/Staat/Oeffentlicher-Dienst/_inhalt.html; Letzter Abruf: 21.08.2025

5.4 Datenqualität und -semantik – Desiderata

Da der DATENATLAS als das zentrale System für die Recherche von Datenbeständen der BUNDESVERWALTUNG dienen soll, muss besonderes Augenmerk auf die Aspekte der Datenbereitstellung, -aufbereitung und -qualität geworfen werden.

Als relativ spät in die Entwicklung übergegangenes Projekt kann der DATENATLAS davon profitieren, sich bereits an zahlreichen Vorarbeiten[53] und existierenden Lösungen mit Verwaltungsbezug, wie z.B. GOVDATA, orientieren zu können um den dort gesetzten Mindeststand zu erreichen.

[53] Siehe Abschnitt 3.2.

Der Fokus dieses Abschnittes liegt auf Desiderata, welche die *Datenqualität* als auch die Implementierung von *Linked (Open) Data* im DATENATLAS betreffen. Diese Aspekte haben zumeist auch eine direkte Auswirkung auf die Nachnutzbarkeit von Daten, welche in Abschnitt 5.5 separat thematisiert wird.

Datenqualität

Bereits in Abschnitt 3.6 wurde die Bedeutung einer ausreichenden *Datenqualität* für die ÖFFENTLICHE VERWALTUNG diskutiert, um Datensätze auffinden oder nutzen zu können.

Da die ÖFFENTLICHE VERWALTUNG Daten nutzen soll, um Entscheidungen zu treffen, muss die *Datenqualität* derart verlässlich sein, dass Verwaltungshandeln nachvollziehbar und korrekt erfolgt.

Die Datenbasis des DATENATLAS wird aktuell durch zwei Datenquellen gespeist:

1. die individuelle Erfassung einzelner Datensätze und
2. den Datenimport von mehreren Datensätzen aus GOVDATA.

Perspektivisch ist es vorgesehen, weitere Datenquellen vergleichbar zu GOVDATA anzubinden.

THEORETISCH GEFÄHRDET der Import aus GOVDATA bereits jetzt die im DATENATLAS vorliegende *Datenqualität*, da der Gesetzgeber Behörden davon freistellt, *Open Data* auf „Richtigkeit, Vollständigkeit, Plausibilität oder in sonstiger Weise“ §12a(8) EGovG (2024) zu prüfen[54].

[54] Siehe Abschnitt 3.7.

Nach dem Import solcher Daten in den DATENATLAS lässt sich kaum mehr erkennen, welchen Ursprung einzelne Datensätze haben oder welchen Datenqualitätsdimensionen[55] sie genügen.

[55] Siehe Abschnitt 3.6.

Auf dieser Grundlage verlässliche Entscheidungen treffen zu können erscheint zweifelhaft, selbst wenn – ohne Zweifel – ein hoher Anteil qualitativ hochwertiger *Open-Data-Datensätze* in GOVDATA existiert.

Neben dieser Problematik, die vor allem auf regulatorischer Ebene zu lösen ist, jedoch auch durch eine entsprechende Gestaltung der GUI, welche die Provenienz der Datensätze hervorhebt, abgeschwächt werden kann, bestehen weitere Desiderate, welche *unbedingt* adressiert werden müssen, um auf dem Bereich der *Datenqualität* den *Stand der Technik* erreichen zu können.

DIE INTRANSPARENTE AGGREGATION der zu importierenden Datensätze[56] stellt ein Risiko im Bereich der Datenqualitätssicherung dar. Da kein manueller Eingriff in diese Aggregation vorgesehen ist, kann weder bestimmt werden, ob die Zusammenfassungen im Ergebnis *genau, vollständig, konsistent, eindeutig* oder *gültig* sind.

[56] Siehe Abschnitt 4.5.

In der Folge besteht das erhebliche Risiko, dass unentdeckt bereits fünf von sechs der DAMA-Datenqualitätsdimensionen negativ beeinflusst werden[57].

[57] Siehe Abschnitt 3.6.

Obwohl dem Gutachter alle Parameter des in Abschnitt 4.5 (Datenimport II – Ablauf Datenimport) beschriebenen Imports in den DATENATLAS vorlagen, war es nicht möglich auf GOVDATA (also der verwendeten Datenquelle) ein vergleichbares Aggregationsergebnis zu erzielen. Dies gelang auch nicht, nachdem u.a. der Product Owner dieses Portals hinzugezogen wurde.

Eine notwendige Sicherung der Datenqualität mittels gängiger Methoden ist im Datenatlas nicht implementiert.

In Kapitel 4 wurde im Rahmen der User Journeys wiederholt auf die sogenannte *Freitext-Problematik*[58] hingewiesen.

[58] Siehe Abschnitt 4.7.

Tatsächlich stellt diese ein zentrales Problem des Datenatlas bzgl. der *Datenqualität* dar.

Dadurch, dass die Metadaten-Felder des Datenatlas häufig mit frei wählbaren Texten belegt werden können, wird die Auffindbarkeit einzelner Datensätze erschwert, da z.B. Schreibfehler nicht korrigiert werden. Hinzu kommt, dass nicht sichergestellt wird, dass die Semantik einzelner Begriffe gleich bleibt, was die Recherche und Interpretation weiter erschwert.

Üblicherweise werden im Daten- und Informationsmanagement *kontrollierte Vokabulare* oder *Thesauri* verwendet, um die oben genannte Probleme zu lösen. Dieser Ansatz ist im Archiv- und Bibliotheksbereich seit Jahrhunderten bewährte Praxis. Diese Methode wird durch internationale Normen wie ISO 25964-1 (2011) empfohlen und hat auch längst Einzug in Rechtsnormen gefunden, wie in Abschnitt 3.6 dargelegt wurde.

Auch DCAT-AP fordert die Verwendung europäischer und nationaler *kontrollierter Vokabulare*[59], so dass nicht nachvollziehbar ist, warum diese gängige *best practice* bei der Implementierung des Datenatlas keine Verwendung findet und damit eine Minderung der *Datenqualität* in Kauf genommen wird.

[59] `https://t1p.de/25hut`; Letzter Abruf: 28.07.2025

Durch den Verzicht auf *kontrollierte Vokabulare* wird es ferner unmöglich, Fehleingaben zu verhindern und Anwendenden Orientierung bei Erfassung und Recherche zu bieten.

Der häufige Rückgriff auf Freitexte erschwert bzw. verhindert ebenso die automatisierte Validierung von Datensätzen im Datenatlas[60], so dass keine umfassende Bewertung der *Datenqualität* innerhalb des Datenatlas möglich wird. Dadurch wird die konkrete Nützlichkeit des Software-Systems in Frage gestellt.

[60] Siehe Abschnitt 3.6.

Nach Kenntnis des Autors nutzt der Datenatlas eine Teilmenge von DCAT-AP, stellt jedoch keine Mittel zur automatisierten Schema-Validierung bereit, was die Bewertung der *Datenqualität* weiter erschwert. Es ist mindestens zu erwarten, dass das verwendete Schema publiziert wird, um ein Mindestmaß an Interoperabilität[61] zu ermöglichen.

[61] *Interoperabilität*; die Fähigkeit verschiedener (IT-)Systeme zusammenzuwirken und Daten miteinander auszutauschen.

Die Gültigkeit von Datensätzen (siehe Kriterium 6 (Datenqualität)) kann zum aktuellen Zeitpunkt deshalb für *keinen* Datensatz im System überprüft werden.

Es ist davon auszugehen, dass die im Datenatlas vorgehaltenen Datensätze kaum – wenn überhaupt – *maschinenlesbar* sind und sich dementsprechend schlecht für Anwendungen der *Künstliche Intelligenz* oder sonstige Nachnutzungsszenarien eignen werden[62].

[62] Siehe Abschnitt 5.6.

Die Verletzung der Datenqualitätsdimension *Vollständigkeit* (siehe Kriterium 2 (Datenqualität)) lässt sich auf zwei Ebenen beobach-

ten. Die bereits in Abschnitt 2.2 präsentierte Abb. 5.6 zeigt deutlich, dass unvollständige Datensätze bereits zum Beobachtungszeitpunkt existieren.

Andererseits soll der DATENATLAS das „erste vollständige Datenmodell der Bundesverwaltung“[63] darstellen. Da keine Bezugsgrößen oder Metriken zur Messung der Vollständigkeit ermittelt werden konnten, kann aktuell nicht bewertet werden, inwiefern die Aussage zutrifft.

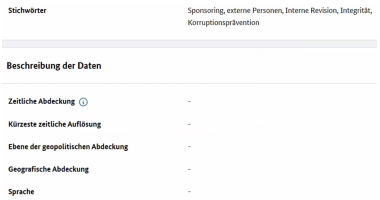


Abbildung 5.6: Auszug aus den Metadaten eines Beispieldatensatzes des DATENATLAS

[63] Bundesdruckerei. *Erstes Vollständiges Datenmodell Der Bundesverwaltung - Pressemitteilung*, 2022. https://tinyurl.com/bdr-pm3. Letzter Abruf: 21.07.2025

AUFGRUND DES FEHLENS von Plausibilitätsprüfungen während der Dateneingabe ist es möglich, dass das Kriterium der *Eindeutigkeit* (siehe Kriterium 4 (Datenqualität)) verletzt wird, da potenziell mehrere Datensätze im DATENATLAS angelegt werden können, die jeweils die gleichen Objekte oder Ereignisse der Realwelt beschreiben.

Datensemantik

Die Bedeutung der semantische Datenkontextualisierung und *Linked (Open) Data* im Besonderen wurde bereits in Abschnitt 5.2 thematisiert. So stellen alle dort diskutierten Datenportale *Linked (Open) Data*-Funktionen bereit.

Die Idee, Daten in einen semantische Kontext zu bringen, wurde in Abschnitt 3.6 vorgestellt. Hierbei werden Technologien und Konzepte des *Semantic Web* wie RDF oder *Persistent Identifier* (PID) genutzt, um maschinell verarbeitbare Zusammenhänge zwischen Daten im Sinne eines „Web of data“ (d.h. als *Linked (Open) Data*) zu modellieren. Die Nutzung dieser Technologie ist im Bereich von Datenportalen ohne Zweifel als *Stand der Technik* zu bezeichnen.

Die Berücksichtigung von *Linked (Open) Data-Prinzipien* während der Konzeption von Datenportalen manifestiert sich in der Regel in technischen Implikationen. Einerseits müssen zur effizienten Verarbeitung und Bereitstellung von *Linked (Open) Data* geeignete Datenbanksysteme in Form von Triplestores genutzt werden, andererseits stellen *Linked (Open) Data-Frameworks* einen Reihe an Kernfunktionen bereit, die direkt verwendet werden können.

Dies kann am Beispiel von GOVDATA illustriert werden. Die CKAN-basierte Implementierung bietet u.a. einen direkten Export beliebiger Datensätze oder Suchanfrage-Ergebnissen u.a. als RDF an. So können mittels des Aufrufs einer URL[64] sämtliche Datensätze mit Bezug zum Suchbegriff „kindergarten“ als RDF exportiert werden.

Durch die Nachnutzung von CKAN und der bereitstehenden Plug-ins werden außerdem weitere übliche Datenformate aus dem Bereich *Linked (Open) Data*, wie Turtle oder JSON-LD, unterstützt.

Durch den Rückgriff auf einen Triplestore – Apache Jena Fuseki im Fall von GOVDATA[65] – werden außerdem Anfragemöglichkeiten über SPARQL[66] verfügbar, was die Anschlussfähigkeit des Systems an andere Anwendungen des *Semantic Web* weiter erhöht.

[64] https://ckan.govdata.de/api/3/action/dcat_catalog_search?q=kindergarten&format=rdf; Letzter Abruf: 15.07.2025

[65] https://github.com/GovDataOfficial/GovDataPortal/blob/master/INSTALL.md; Letzter Abruf: 15.07.2025 y

[66] *SPARQL*; SPARQL Protocol And RDF Query Language, eine Graphen-basierte Anfragesprache für RDF. welche es ermöglicht, logische Aussagen zu formulieren auf welche maschinell geschlossen werden kann.

Trotz der Existenz dieser Lösungen bietet der Datenatlas keine vergleichbaren Funktionen.

Diese Einschätzung basiert auf Aussagen einzelner Mitarbeitender der Datenlabore als auch den im Rahmen der User Journeys[67] präsentierten Indizien.

[67] Siehe Kapitel 4.

Die Vorteile der Nutzung von *Linked (Open) Data* beim Datenatlas lassen sich leicht begründen. Die Öffentliche Verwaltung bzw. die Bundesverwaltung sind streng hierarchisch untergliedert. Zusätzlich werden je nach Legislaturperiode einzelne Arbeitsbereiche z.B. zwischen Bundesministerien verschoben bzw. in neue Ministerien überführt. Dieser Fall lässt sich aktuell bei der Gründung des Bundesministeriums für Digitales und Staatsmodernisierung (BMDS) beobachten.

Außerdem werden Behörden im Rahmen der Ressortzuschnitte umbenannt, auch wenn sich ihr Arbeitsbereich nur minimal verschiebt und ein Großteil der Fachabteilungen dort verbleibt. So war beispielsweise das Bundesministerium des Innern bis 2025 als Bundesministerium des Innern und für Heimat bekannt.

Die Datenmodellierung solcher Zusammenhänge lässt am besten mithilfe von Graphen umsetzen. Zur besseren Nachvollziehbarkeit werden Entitäten wie ein Bundesministerium, üblicherweise mittels *Persistent Identifier* (PID)[68] modelliert und der jeweils gültige Name über eine Kante im Graph hergeleitet.

[68] Siehe Abschnitt 3.6.

Es ist leicht zu erkennen, dass es sich hierbei um Kernkonzepte von *Linked (Open) Data*[69] handelt.

[69] Siehe Abschnitt 3.6.

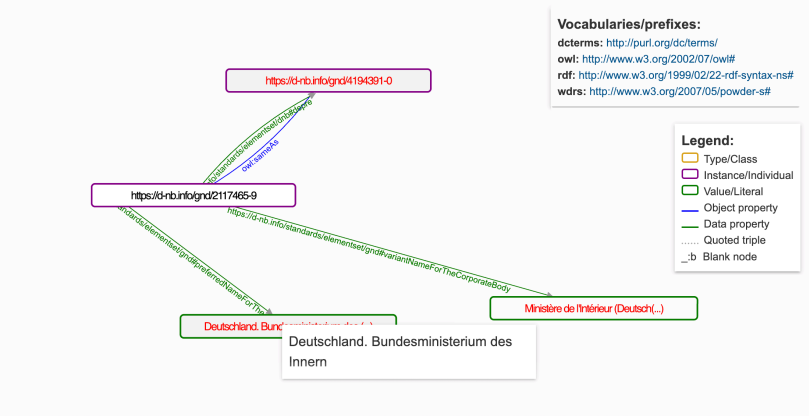


Abbildung 5.7: RDF-Graph am Beispiel des Bundesministeriums des Innern (Auszug)

Abbildung 5.7 illustriert einen solchen Graphen. Das BMI ist hier über die GND[70]-ID 2117465-9, einen PID, modelliert, welcher über Kanten u.a. mit der offiziellen deutschen als auch französischen Bezeichnung verbunden ist[71].

Die Abbildung von Hierarchien wie bei nachgeordneten Bundesbehörden unter oberste Bundesbehörden ist unter Zuhilfenahme eines Graphs, z.B. unter Verwendung von RDF, trivial.

[70] *GND*; Gemeinsame Normdatei, eine Normdatei für Personen, Körperschaften und weitere Entitäten.

[71] Die vollständige Darstellung des Graphen sowie seiner Generierung findet sich in Anhang A.6.

Die Nachteile durch den Verzicht auf eine geeignete Art der Datenmodellierung und -haltung in Form eines *Triplestores* kann am Beispiel eines Usability-Problems des Datenatlas aufgezeigt werden: So ist es im Arbeitsschritt 4.1 Informationsrecherche VI – Schlagwort-basierte Suche: Verfeinerung der Filterung möglich, ungültige Kombinationen aus dem Ministerium mit Fachaufsicht und den fachlich zuständigen Behörden zu wählen.

Wäre im Datenatlas eine entsprechende Ontologie, z.B. unter Verwendung des *Semantic Web*-Standards OWL[72], hinterlegt, würde dieses Szenarion unmöglich, da bereits der Klick auf ein Ministerium nur noch den Teilgraphen zur Verfeinerung ausgewählt hätte, welcher die unter der jeweiligen Fachaufsicht stehenden Behörden enthält. Das implizit durch den Datenatlas vorausgesetzte Domänenwissen wäre in diesem Fall nicht notwendig.

[72] *OWL*; Web Ontology Language, eine Logik-basierte, maschinell lesbare Auszeichnungssprache für Ontologien (siehe auch Abb. 3.20).

Dieses einfache Beispiel macht deutlich, wie sich eine suboptimale Konzeption des dem Datenatlas zugrundeliegenden Datenmodell und dem sich daraus ergebenden Datenmanagement *unmittelbar* negativ auf die Usability des Gesamtsystems auswirkt und eine effektive und effiziente Bedienung erschwert.

Auch wenn man diesen gerade beschriebenen zweistufigen Filterprozess im Sinne einer automatisierbaren Plausibilitätsprüfung nicht in der GUI implementiert hätte, hätte der Rückgriff auf eine Ontologie es sogar Nutzenden ermöglichen können, sich selbständig aus einer Fehlersituation heraus zu helfen – anstelle wie in Abb. 4.12 dargestellt – abstrakt auf die Änderung von Such- und Filtereinstellungen verweisen zu müssen.

Würde hier eine Graph-basierte Datenmodellierung vorliegen, wäre ein expliziter Hinweis auf den vorliegenden Widerspruch möglich gewesen.

Selbst wenn das entsprechende Domänenwissen über die Struktur der Bundesverwaltung nicht vorliegt, wäre diese Information leicht ermittelbar bzw. jederzeit automatisiert aus einem *Knowledge Graph* abrufbar, wie in Abschnitt A.4 skizziert wird.

Eine wesentliche Voraussetzung für die Nutzung von *Linked (Open) Data* stellt die Verwendung von *Persistent Identifier* (PID) dar, wie bereits erläutert wurde.

Dem *Stand der Technik* folgend, könnten PID im Datenatlas *zusätzlich* zur Adressierung einzelner Datensätze, zur Eingabe maschinell auswertbarer Daten in die jeweiligen Metadatenfelder, in Kombination mit *kontrollierten Vokabularen* zur Formulierung maschinenlesbarer Aussagen über Sachverhalte oder zur dauerhaften Adressierung von Suchergebnissen – wie sie in vergleichbaren Systemen üblich ist (siehe Abb. 5.8)– eingesetzt werden.

Zur Wahrung eines Mindestmasses an Datenqualität ist die Unterstützung von PID auch deshalb notwendig, weil die importierten Datensätze aus GovData häufig auf diese zurückgreifen. Durch den Import verlieren diese Datensätze diese semantische

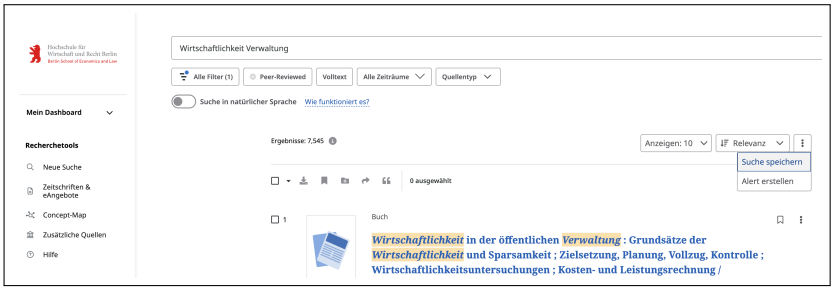


Abbildung 5.8: Speicher-Funktion eines Recherche-Ergebnisses am Beispiel des Discovery-Systems der HWR Berlin

Wertigkeit, was sie um bis zu zwei Stufen im 5-Sterne-Modell für offene Daten (siehe Abb. 3.23) zurückwirft.

Dies mindert die *Datenqualität* im DATENATLAS weiter.

Fazit Datenqualität und -semantik

Die Bewertung der Aspekte *Datenqualität* und *-semantik* – zwei wesentlichen Anforderungen, um verlässliche datenbasierte Entscheidungen in der BUNDESVERWALTUNG treffen zu können – fällt aufgrund der aufgezeigten Punkte *mangelhaft* aus.

ZUSAMMENFASSEND GESAGT verfügt der DATENATLAS zum Begutachtungszeitpunkt über keinerlei wirksam implementierte Mechanismen der *Datenqualitätssicherung*.

Allein durch die erfolgte Sichtprüfung ist davon auszugehen, dass der DATENATLAS bereits jetzt eine Vielzahl von ungültigen und unvollständigen Datensätzen enthält.

HINZU KOMMEN die sich durch den Verzicht auf *Linked (Open) Data-Technologien* ergebenden weiteren Datenqualitäts- und Usability-Probleme, die letztendlich zu einer *mangelhaften Maschinenlesbarkeit* der vorgehaltenen Daten führen.

QUALITATIV GESEHEN fällt der DATENATLAS in der vorliegenden Form hinter GOVDATA zurück, auch wenn hier keine regulativen Verpflichtung zur *Datenqualitätssicherung*[73] besteht. Dies betrifft sowohl die *Datenqualität* als auch die *Datensemantik*.

[73] Siehe Abschnitt 3.7.

Es bleibt offen, warum man sich nicht am konzeptionell reiferen Datenmodell von GOVDATA und seiner Validierungsmöglichkeiten orientiert hat und vergleichbare Schnittstellen zum Datenaustausch implementiert hat.

Die aktive Verringerung der Datenqualität von Datensätzen aus GOVDATA durch die Entfernung von *Linked (Open) Data-Bestandteilen* wie PID beim Import in den DATENATLAS ist inakzeptabel. Als Konsequenz verbaut dieses Vorgehen sogar den Export nach GOVDATA.

WENNGLEICH DER EXPORT von Datensätzen nach GOVDATA oder in andere Portale nicht Teil des Auftrags gewesen zu sein scheint,

wäre eine Vorbereitung dieser Funktionalität mehr als naheliegend gewesen. Dieses Desiderat lässt sich leicht aus Forderungen der nationalen Ebene, wie §12a EGovG (2024), bzw. der supra-nationalen Ebene in Form der EU-Richtlinie 2019/1024 herleiten, da davon auszugehen ist, dass ursprünglich nur für den verwaltungsinternen Einsatz erstellte Datensätze sich in *Open Data* wandeln werden[74].

[74] Siehe Abschnitt 3.6.

Aufgrund der Defizite bei der Maschinenlesbarkeit der Daten ist davon auszugehen, dass ein vollständig automatisierter Datenaustausch aktuell nicht möglich ist.

Diese Beispiele verdeutlichen, dass sich der Datenatlas bereits jetzt in eine *technische Sackgasse* manövriert hat, da die Defizite im Bereich der konzeptionellen Datenmodellierung derart fundamental sind, dass davon auszugehen ist, dass hier eine zentrale, technische Komponente im Ganzen auszutauschen wäre.

Diese Bewertung findet sich auch bereits in Abschnitt 5.2. Hinzu kommen die sich daraus ergebenden Usability-Probleme, welche in Abschnitt 5.3 diskutiert wurden.

Zukünftige Anwendungsszenarien wie die Nutzung der Datenbasis des Datenatlas im Rahmen von *KI-Anwendungen* sind aufgrund der Mängel im Bereich der *Datenqualität* und der *Maschinenlesbarkeit* zum aktuellen Stand auszuschließen. Diese werden potenziell erst durch aufwändige manuelle Datenaufbereitungsschritte denkbar.

5.5 *Nachnutzbarkeit und Digitale Souveränität – Desiderata*

Der Datenatlas stellt laut Eigendarstellung den „souveräne[n] Datenkatalog für die Bundesverwaltung“[75] dar.

[75] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025

Souveränität bedeutet, dass der Staat unabhängige Entscheidungen treffen kann. Dies gilt für den digitalen Raum, in welchem der Datenatlas verortet ist. In diesem Fall wird häufig der Begriff *Digitale Souveränität* genutzt, welcher weiter unten diskutiert wird.

Die Bewertung der Nachnutzbarkeit des Datenatlas ergibt sich sich u.a. aus Abschnitt 3.7 – also regulatorischen und Wirtschaftlichkeitsüberlegungen.

Nachnutzbarkeit

Der Begriff „Nachnutzbarkeit“ hat mindestens zwei Aspekte, die im Kontext des Datenatlas relevant sind: die Nachnutzbarkeit des Systems und die der verwalteten Daten.

Nachnutzung des Systems Die Nachnutzbarkeit des Datenatlas setzt voraus, dass dieser den Anforderungen der nachnutzenden Behörde genügt, was sowohl funktionale als auch operative Anforderungen mit einschließt.

Wie bereits in Kapitel 2 erwähnt wurde, wird aktuell die Nachnutzung des DATENATLAS durch das weitere Einrichtungen geprüft.

Aufgrund der bereits aufgezeigten, *schwerwiegenden Mängel* des DATENATLAS und der unbekannten funktionalen Anforderungen anderer Einrichtungen, kann jedoch zum aktuellen Zeitpunkt davon ausgegangen werden, dass sich hier keine Nachnutzungsperspektive ergibt.

Dies ist insbesondere auf die Probleme im Bereich der *Informationsrecherche*[76] zurückzuführen, da es sich hierbei um ein essenzielle Desiderata handelt, welche sicherlich auch seitens anderer Behörden genannt würden.

[76] Siehe Abschnitt 5.2.

Der Autor des vorliegenden Gutachtens hat keine Kenntnis über die Betriebsumgebung anderer Behörden noch über die tiefergehende technische Implementierung des DATENATLAS, so dass zum operativen Teilaspekt keine Aussage getroffen werden kann.

Auch die Lizenz der Software kann Auswirkungen auf deren Nachnutzbarkeit und Weiterentwicklung haben, wie im folgenden Teilabschnitt Digitale Souveränität diskutiert wird.

Nachnutzung der Daten Die Nachnutzung von Daten kann in zwei Bereiche untergliedert werden:

1. die Nachnutzung von Daten im Austausch zwischen zwei Systemen (*Interoperabilität*)
2. die Nachnutzung von Daten im Falle der Abschaltung des Systems (*Exit-Strategie*)

INTEROPERABILITÄT wird im DATENATLAS rein *konsumptiv* durch den Datenimport aus GOVDATA unterstützt.

Eine wesentliche Anforderung an *interoperable* Systeme stellt eine bewertbare *Datenqualität* dar. Wie in Abschnitt 5.4 ausführlich begründet wurde, ist diese beim DATENATLAS nicht sichergestellt. Wie in Abschnitt 5.4 dargelegt wurde, existiert ebenfalls kein Datenschema des DATENATLAS, welches sich automatisiert validieren lässt.

ALS ORIENTIERUNG für ein interoperables System hätte dem DATENATLAS GOVDATA dienen müssen – zumal aus diesem Portal Daten importiert werden.

So unterstützt GOVDATA zwei verschiedene Schnittstellen für die Datenrecherche und den Datenexport sowie drei gängige Exportformate (RDF, TURTLE und JSON-LD)[77].

[77] https://www.govdata.de/suche/daten/govdata-metadatenkatalog; Letzter Abruf: 21.07.2025

VERNACHLÄSSIGT MAN DIE INTEROPERABILITÄT kann ein Mindestmaß an Daten-Nachnutzung durch strukturierte Datenexporte, z.B. nach CSV oder RDF, erfolgen. Selbst solche niedrigschwellig zu implementierende Funktionen sind aktuell nicht im DATENATLAS verfügbar.

Im Falle der Abschaltung des Datenatlas müsste ein finaler Datenexport zur Verfügung stehen. Ob dieser implementiert ist, konnte nicht ermittelt werden.

Eine entsprechende *Exit-Strategie* kann deshalb im Rahmen des Gutachtens weder erstellt noch bewertet werden.

Aufgrund der im folgenden Abschnitt diskutierten Punkte, muss jedoch befürchtet werden, dass eine Loslösung vom existierenden Datenatlas – aus welchem Grund auch immer – mit erheblichen Hürden versehen sein wird und die Souveränität der Bundesverwaltung negativ beeinflusst.

Digitale Souveränität

Der Begriff der *Digitalen Souveränität* wird im aktuellen Diskurs derart unscharf genutzt, dass eine genaue Definition den Rahmen des Gutachtens bei weitem sprengen würde. Eine aktuelle Zusammenfassung der insbesondere Deutschland genutzten Bedeutungen und Narrative findet sich bei Lambach und Oppermann (2023).

Wie eingangs erwähnt, bezeichnet die Bundesdruckerei den Datenatlas als souveräne Lösung.

Betrachtet man die aktuelle Diskussion um die Bedeutung des Begriffs ist dieser Einschätzung klar zu widersprechen.

Der Datenatlas liegt als *Closed Source* vor, d.h. die Bundesverwaltung hat keinerlei Möglichkeit, den zugrundeliegenden Quellcode der Software zu prüfen, weiterzuentwickeln oder den Wissenstransfer im Rahmen einer souveränen Exit-Strategie sicherzustellen.

Da eine Anfrage zur konkreten Lizenzierung bei der Bundesdruckerei nicht beantwortet wurde[78], basiert diese Einschätzung auf einer Recherche mittels der Suchterme „datenatlas“ und „bundesdruckerei“ (inkl. Variationen) auf den gängigen Repositorys wie dem *Open Source*-Reposiory des Bundes, *openCode*[79], oder den großen kommerziellen Plattformen *GitHub*[80] und *GitLab*[81].

Als *Open Source-Lösung* findet sich GovData z.B. auf *GitHub*[82].

Der als Open Source bezeichnete freie Zugang zum Quellcode wird häufig als einer der Pfeiler *Digitaler Souveränität* betrachtet und ist deshalb seit längerem Bestandteil der einschlägigen Gesetze wie dem EGovG (2024)[83].

Der Antagonismus zwischen *Closed Source* und *Digitaler Souveränität* besteht darin, dass die Nichtverfügbarkeit des Quellcodes die staatliche *Digitale Souveränität* derart einschränkt, dass sie keinen anderen, geeigneten Dienstleister mit der Weiterentwicklung des Datenatlas beauftragen kann und somit dauerhaft an die Bundesdruckerei gebunden ist.

Dieser Effekt wird üblicherweise als *vendor lock-in* bezeichnet und steht dem Ziel der *Digitalen Souveränität* diametral gegenüber. Lambach und Oppermann (2023) führen den *vendor lock-in* explizit als konkrete Bedrohung für die Öffentliche Verwaltung auf.

[78] Siehe Anhang A.5.

[79] `https://opencode.de`; Letzter Abruf: 25.07.2025

[80] `https://github.com`; Letzter Abruf: 25.07.2025

[81] `https://gitlab.com`; Letzter Abruf: 25.07.2025

[82] `https://github.com/GovDataOfficial/GovDataPortal`; Letzter Abruf: 25.07.2025

[83] Siehe Abschnitt 3.7.

da dieser das staatliche Handeln gerade im Bereich der (Weiter-)Entwicklung und des Betriebs von IT-Systemen einschränkt.

Der *vendor lock-in* wirkt sich ebenso auf die verwalteten Daten aus und kann hier gegebenenfalls bis zum *Totalverlust* dieser führen, wenn keine standardisierte Datenhaltung implementiert wurde und der Dienstleister nicht länger für den Datenexport zur Verfügung steht.

Für die Verwendung von Open Source sprechen weitere Effekte wie das Community-Building, da sich Gemeinschaften von Nutzenden eher um weit verbreitete und *Open Source*-Anwendungen, wie die in Tabelle 5.2 vorgestellten Systeme, bilden. Dadurch dass der Datenatlas ausschließlich als *Closed Source* durch die Bundesverwaltung genutzt wird, ist die Möglichkeit der Weiterentwicklung durch Dritte unmöglich gemacht worden. Im Kontext der Öffentliche Verwaltung wird die Bedeutung des Community-Buildings am wachsenden Kreis der Piveau-Anwendenden deutlich, welcher mittelfristig auch GovData umfassen wird.

Folglich kann die Bundesverwaltung auch *nicht* von Weiterentwicklungen direkt profitieren, wie es die in Abschnitt 5.2 genannten Systeme – wie z.B. GovData, Open.ByData oder Datenadler – können und deshalb teilweise seit Jahrzehnten erfolgreich sind.

Die sich aus dem *Closed-Source-Ansatz* ergebenden negativen Auswirkungen auf die *Wirtschaftlichkeit* werden in Abschnitt 5.8 gesondert betrachtet.

Dabei muss angemerkt werden, dass etablierte *Repository-Systeme* wie Fedora oder CKAN mehr *Digitale Souveränität* versprechen, da Piveau nach Kenntnis des Autors aktuell ausschließlich von der Fraunhofer-Gesellschaft abhängt.

Perspektivisch könnte sich dies durch Anwender wie die Bundesländer Bayern (Open.ByData) und Brandenburg (Datenadler) ändern, wenn diese hier nennenswerte Entwicklungsressourcen aufbauen. Durch die Migration von GovData ist hier zukünftig mit einer noch breiteren Aufstellung zu rechnen.

Die nachhaltige Klärung der Betreiberfrage, z.B. in Form des Konsortialmodells von Fedora oder *ePayBL*[84] als Kooperationsmodell zwischen Bund und Ländern, würde sich positiv auf den Aspekt der *Digitalen Souveränität* bei der Nutzung von Piveau auswirken.

[84] https://www.epaybl.de

Fazit Nachnutzbarkeit und Digitale Souveränität

Auch wenn diese Überlegung klar außerhalb des Auftrags zur Entwicklung des Datenatlas liegt, sollte man sich die Frage stellen, ob die Beschränkung des Datenatlas – unter Vernachlässigung der präsentierten Mängel – auf die Bundesverwaltung zielführend ist.

Wie in Abschnitt 3.6 dargestellt wurde, erfolgt die Datenerfassung häufig auf Kommunal- oder Landesebene, so dass bei der Verwendung unterschiedlicher Software-Systeme mit Datenaustauschproblemen zu rechnen ist.

Der momentan existierende DATENATLAS spaltet hier weiter, da er nicht einmal technische Lösungen, die bereits auf Ebene des BUNDES existieren, wie z.B. GOVDATA bzw. dessen Basis CKAN, integriert oder darauf aufbaut.

Damit kann der DATENATLAS auch hier keinen Beitrag zur Weiterentwicklung leisten, zumal er, wie in Abschnitt 5.4 diskutiert wurde, nicht die Qualität des Datenmodells von GOVDATA erreicht.

DIE DIGITALE SOUVERÄNITÄT betreffend, etabliert der DATENATLAS neue Abhängigkeiten und schwächt diese damit weiter.

Würde ein bereits bestehendes System nachgenutzt, würde dies die *Digitale Souveränität* eher stärken, da gerade *Open-Source-Software* über existierende Communities, Entwicklungsressourcen und Community-getriebene Betreiber-Modelle verfügt. Dies trifft insbesondere auf FEDORA und in einem gewissen Maß auf PIVEAU oder CKAN zu.

Bezogen auf die *Digitale Souveränität* stellt die aktuelle Bindung der BUNDESVERWALTUNG an die BUNDESDRUCKEREI und deren Subunternehmern ein abzustellendes Risiko dar.

5.6 Anwendungsfälle der Künstlichen Intelligenz – Desiderata

Das Fazit aus Abschnitt 5.4 macht deutlich, dass sich die Nutzung der Datenbasis des DATENATLAS für *KI-Anwendungen* aufgrund der Defizite im Bereich der *Datenqualität* zum aktuellen Stand verbietet bzw. erst durch aufwändige manuelle Datenaufbereitungsschritte denkbar wird.

Dieser Umstand ist dem in Abschnitt 3.6 beschriebenen *EVA-Prinzip* geschuldet, welches auch gemeinhin als *„garbage in – garbage out“* überspitzt bezeichnet wird. Wie bei jeder IT-Anwendung ist es auch für die Ansätze der *Künstlichen Intelligenz* unabdingbar, dass ein Mindestmaß an *Datenqualität* zur Verfügung steht.

Dies gilt insbesondere für Anwendungen des *maschinellen Lernens*, da diese Verfahren beispielsweise – stark vereinfachend gesprochen – auf Grundlage einer bereitgestellten Datenmenge, den *Trainingsdaten*, Muster erkennen, welche sie auf neue, unbekannte Datenmengen übertragen und damit eine neue Ausgabe generieren.

Naturgemäß führt dies bei fehlerhaften oder für den Use Case unpassend gewählten Trainingsdaten dazu, dass überwiegend fehlerhafte Ausgaben generiert werden.

Dieses Phänomen ist als allgemein bekannt anzusehen und findet sich in Literatur, die sich eher an Studierende, wie z.B. Alpaydın (2008), oder Praktikerinnen und Praktiker, wie Géron (2018), wendet.

Trotz der angezeigten Mängel bezeichnet eine Pressemitteilung den Datenatlas als „Meilenstein auf dem Weg zur datengetriebenen Verwaltung: Basis für künftige Datenanalysen und KI-Anwendungen wie Maschinelles Lernen“[85].

Ohne eine wirkungsvolle Datenqualitätssicherung[86] rückt das Ziel, den Datenatlas mit Anwendungen der *Künstlichen Intelligenz* zu verbinden, in weite Ferne.

Tatsächlich ist aktuell eher anzunehmen, dass eine manuelle Datenqualitätskorrektur und Datenaufbereitung der im Datenatlas erfassten Datensätze einer Unterstützung durch *Künstliche Intelligenz* bedarf.

Aufgrund der diskutierten *„Freitext-Problematik“* ist damit zu rechnen, dass aktuell eine große Anzahl inkorrekter Daten aufgrund von Fehleingaben oder Rechtschreibproblemen Teil der Datenbasis sind.

Hier könnte ein Verfahren des *maschinellen Lernens*, das sogenannte *Clustering*, Verwendung finden, um sich zumindest einen ersten Überblick über die vorliegende *Datenqualität* zu verschaffen, in dem ähnliche Fehleingaben automatisiert zusammengefasst werden. Auf Grundlage dieses Clusterings kann dann eine manuelle Datenkorrektur erfolgen.

[85] Bundesdruckerei (2022)

[86] Siehe Abschnitt 5.4.

Anzunehmen gewesen wäre bei einer solchen Zielstellung in Richtung *Künstliche Intelligenz*, dass ein wohlbekannter Fakt bei der Entwicklung berücksichtigt worden wäre: gerade KI-basierte Anwendungen profitieren von einer Strukturierung von Daten[87] bzw. der Verwendung einer Graph-basierten Datenmodellierung[88] wie sie im Falle der Nutzung von Technologien des *Semantic Web* oder *Linked (Open) Data* vorgelegen hätten.

Aus diesem Grund nutzt z.B. Google spätestens seit 2012 einen sogenannten *Knowledge Graph*, um die Suchergebnisse zu verbessern[89].

[87] Eve Sauvage, Sabrina Campano, Lydia Ouali, und Cyril Grouin. Does the Structure of Textual Content Have an Impact on Language Models for Automatic Summarization? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, Seiten 280–285. Association for Computational Linguistics, 2024. DOI: 10.18653/v1/2024.acl-srw.25

[88] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, und Francesco Osborne. Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review*, 56(11):13071–13102, 2023. ISSN 0269-2821, 1573-7462. DOI: 10.1007/s10462-023-10465-9

[89] https://blog.google/products/search/introducing-knowledge-graph-things-not/; Letzter Abruf: 21.07.2025

Fazit Künstliche Intelligenz

Eine für die Bundesverwaltung gewinnbringende Nutzungsperspektive des Datenatlas im Bereich der *Künstlichen Intelligenz* ist aktuell nicht zu erkennen.

Die zu vermutende Verwendung eines relationalen DBMS[90] verhindert nicht nur sinnvolle Informationssuchstrategien sondern limitiert auch Anfragemöglichkeiten. Obwohl aus diversen Gründen eine Basis auf *Linked (Open) Data* für die Kombination mit Anwendungen der *Künstlichen Intelligenz* angezeigt gewesen wäre, setzt der Datenatlas dies nicht um.

[90] Siehe Abschnitt 5.2.

Die Hoffnung in diesem Fall etwas mithilfe einer *Künstlichen Intelligenz* „retten“ zu können, wird aller Wahrscheinlichkeit zerschlagen werden, da bereits bei grundsätzlichen Überlegungen – wie dem Datenmodell – scheinbar Entscheidungen getroffen wurden,

welche die darauf aufbauenden Schritte erschweren bis unmöglich machen.

Ferner bleibt zu vermuten, dass die manuell durchzuführenden Maßnahmen für die Bewertung und Korrektur der Datenbasis des DATENATLAS erheblich sein dürften, was kaum einen *wirtschaftlichen Betrieb* ermöglicht – es sei denn, es sind bisher nur wenige Behörden der BUNDESVERWALTUNG in diesem vertreten.

5.7 *Langzeitarchivierung und -verfügbarkeit – Desiderata*

In Abschnitt 3.7 wurde auf die Grundsätze der Aktenführung verwiesen, welche die BUNDESVERWALTUNG sowohl beim analogen als auch digitalen Arbeiten einhalten muss.

Wird der DATENATLAS im Rahmen der Herbeiführung einer behördlichen Entscheidung genutzt, ist es denkbar, dass generierte Ergebnisse oder einzelne Metadatensätze *aktenrelevant* werden und damit im Nachhinein nicht mehr aus der Akte entfernt werden dürfen. Unter Umständen ist der gesamte Geschehensablauf zu dokumentieren[91], was Recherchen im DATENATLAS einschließen dürfte.

[91] Deutscher Bundestag - Wissenschaftliche Dienste. *Grundsätze der Aktenführung in der Bundesverwaltung*, 2023. https://tinyurl.com/veraktung. Letzter Abruf: 26.07.2025

Dies bedeutet auch, dass die Suchergebnisse oder recherchierten Dokumente entweder *dauerhaft* im DATENATLAS verfügbar sein oder zur Veraktung exportierbar sein müssen. Auch im Falle des Exports ist sicherzustellen, dass diese Daten *dauerhaft* verfügbar bleiben, was bei der Verwendung eines E-Akten-Systems der Fall sein sollte.

Außerdem ist es nötig, die einzelnen Datensätze zu versionieren, d.h. Modifikationen an diesen nachvollziehbar zu machen, um diese sinnvoll verakten zu können.

Die genannten Funktionen existieren aktuell nicht im DATENATLAS und müssten ggf. mittels *persistenter Identifier*[92] und eines Systems, welches die *Langzeitarchivierung (LZA)* unterstützt, umgesetzt werden.

[92] Siehe Abschnitt 5.4.

ZUR UMSETZUNG DER LZA ist die ISO 14721 (2025) als einschlägige Norm zu nennen. Die Norm enthält das OAIS[93]-Referenzmodell, welches in Abb. 5.9 dargestellt ist.

[93] *Open Archival Information System*; Offenes Archiv-Informationssystem.

Die Norm beschreibt dabei OAIS ganzheitlich – bestehend aus Hard- und Software, den darin gespeicherten Informationen sowie den nötigen Richtlinien für Prozesse, welche eine Organisation zur Realisierung von LZA implementieren muss.

Fazit Langzeitarchivierung und -verfügbarkeit

Grundsätzlich stellt sich die Frage, ob die Veraktungsunterstützung im DATENATLAS mittels Exportfunktionalitäten oder der Nutzung einer LZA-Lösung realisiert werden sollte.

In beiden Fällen ist die Versionierung der im DATENATLAS vorgehaltenen Datensätze umzusetzen, welche dann entweder direkt

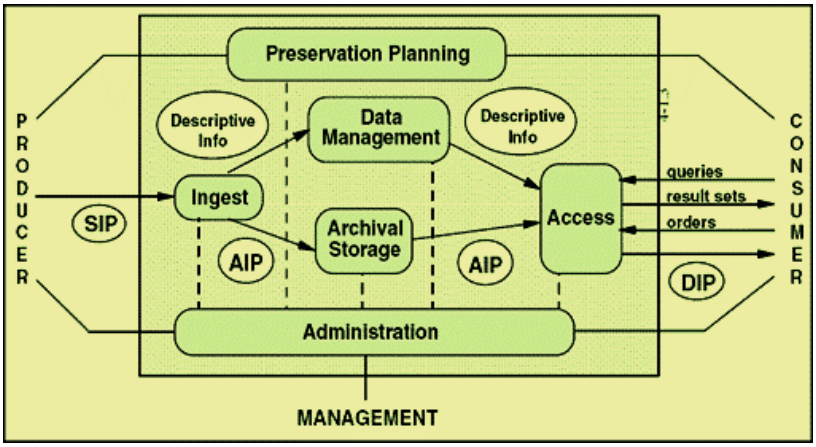


Abbildung 5.9: ISO OAIS-Referenzmodell

langzeitarchiviert oder exportiert werden können.

Hierbei ist sicherzustellen, dass sowohl die Daten als auch die Metadaten[94] des DATENATLAS versioniert werden. Die einzelnen Versionen sind mit PID zu versehen.

[94] Siehe Abschnitt 2.2.

Problematisch dabei ist, dass der DATENATLAS als reines Metadaten-Portal konzipiert ist, dessen verwaltete Daten wiederum Metadaten konkreter Datensätze (im Folgenden als *Content-Daten* bezeichnet) einzelner Bundesbehörden darstellen.

Content-Daten

Letztendlich müssten auch die *Content-Daten* der die Daten bereitstellenden Behörden archiviert werden, damit die entstehende Akte nachvollziehbar bleibt.

DIESES GRUNDSATZPROBLEM ist im Rahmen dieses Gutachtens nicht zu lösen, sollte jedoch zukünftig adressiert werden – zumal es auch GOVDATA betrifft.

GENERELL STELLEN VIELE *Repository-Systeme*[95] Möglichkeiten der Versionierung auf unterschiedlichen Ebenen bereit.

[95] Siehe Tabelle 5.2.

So ermöglicht CKAN beispielsweise nur die Versionierung auf Metadaten-Ebene, während FEDORA sowohl die Versionierung auf Daten- als auch auf Metadaten-Ebene ermöglicht.

Für PIVEAU wurde die Umsetzung dieser Funktion für das erste Quartal 2025 angekündigt[96], scheint aber nach der Prüfung der Dokumentation noch nicht erfolgt zu sein.

[96] `https://doc.piveau.io/community/roadmap/`; Letzter Abruf: 20.08.2025

Zum aktuellen Zeitpunkt lassen sich die für den DATENATLAS relevanten Versionierungsarten nur in FEDORA, DSPACE oder ZENODO umsetzen.

DIE UMSETZUNG DER LANGZEITARCHIVIERUNG kann am effizientesten durch die Verwendung eines bestehenden *Repository-Systems* erreicht werden. Beispielsweise FEDORA und DSPACE bieten geeignete Managementmöglichkeiten und Schnittstellen.

Persistent Identifier sind ebenfalls Teil des Standardfunktionsumfangs der gängigen *Repository-Systeme*.

5.8 Wirtschaftlichkeitsbetrachtung – Desiderata

Ohne die genaue Leistungsbeschreibung des DATENATLAS zu kennen, kann keine vollständige Wirtschaftlichkeitsbetrachtung nach den Anforderungen der VV-BHO (2025)[97] durchgeführt werden.

[97] Siehe Abschnitt 3.7.

Da die beiden einzigen Pressemitteilungen der BUNDESDRUCKEREI[98] bezüglich des umzusetzenden Leistungskatalogs recht vage bleiben, sollen stattdessen die Erstellungsaufwände der folgenden, vergleichbaren Systeme herangezogen werden, deren Aufwände durch den Autor ermittelt werden konnten.

[98] Bundesdruckerei (2025) und Bundesdruckerei (2022)

Dieses Vorgehen entspricht der Schätzung des Auftragswerts nach §3 der Vergabeordnung (VgV, 2016).

1. open.bydata – Das Open-Data-Portal Bayerns,
2. CrossAsia – ein Repository mit hohem Eigenentwicklungsanteil.

Diese Systeme werden einer Grobschätzung der Kosten für den DATENATLAS gegenübergestellt.

Berechnungsgrundlage

Für die Abschätzung der Wirtschaftlichkeit wird mit Tagessätzen freiberuflich Tätiger kalkuliert, die zum Jahresbeginn 2025 erhoben wurden. Grundlage der Erhebung bildet eine Befragung von ca. 6.000 IT-Freelancern durch die Plattformen freelancermap[99] und freelance.de[100].

[99] `https://www.freelancermap.de`

[100] `https://www.freelance.de`

Auch wenn diese Tagessätze höher als interne Tagessätze der BUNDESVERWALTUNG sind, können diese gut als Bezugspunkt für die gemeinsame Bewertung aller drei Systeme dienen.

Aus den Befragungen ergeben sich die in Tabelle 5.3 dargestellten durchschnittlichen Stundenlöhne für freiberuflich ausgeübte IKT-Berufstätigkeiten, die im Folgenden zur Abschätzung der Projektkosten herangezogen werden.

Tabelle 5.3: Durchschnittstagessätze IKT-Freelancer

Berufsfeld	Betrag[101]
Branchendurchschnitt	102,28 €
Beratung und Management	120,00 €

[101] Branchendurchschnitt: `https://www.heise.de/ratgeber/Freelancer-Studie-2025-Das-verdienen-IT-Freiberufler-in-Deutschland-9958271.html`; Letzter Abruf: 21.07.2025; „Beratung und Management“: `https://www.heise.de/news/Stundensaetze-gestiegen-Freelancer-liegen-im-Schnitt-bei-104-Euro-10324942.html`; Letzter Abruf: 21.07.2025

FÜR ALLGEMEINE PROJEKTMANAGEMENTANTEILE an den Projektgesamtkosten gibt die einschlägige Literatur einen Wert im Bereich von 10-20% (mit hoher Streuung je nach Projektkontext) an, während das [102] im Rahmen der „S-O-S-Methode für Großprojekte“ beispielsweise von einem Vollzeitstellenäquivalent (VZÄ) auf sieben Team-Mitglieder ausgeht.

[102] Bundesverwaltungsamt. *S-O-S-Methode Für Großprojekte*, 2021. `https://www.bva.bund.de/SharedDocs/Downloads/DE/Behoerden/Beratung/GrossPM/Handbuch/S-O-S_Handbuch.pdf?__blob=publicationFile&v=2`. Letzter Abruf: 18.08.2025

In diesem Tätigkeitsbereich ist der erhöhte Stundensatz aus Tabelle 5.3 anzusetzen.

Die Jahresarbeitszeit wird im Einklang der einschlägigen Leitlinie des Bundes[103] mit einem Bezugswert von 203,86 Tagen pro Jahr, d.h. 492 Minuten pro Tag bzw. 8,2 Stunden pro Tag für den höheren Dienst in Vollbeschäftigung festgelegt.

Es ist davon auszugehen, dass die Projektleitung und Implementierung des Datenatlas und der anderen Systeme auf dem Qualifikationsniveau des höheren Dienst erfolgt, so dass im weiteren Verlauf die entsprechende Jahresarbeitszeit angenommen wird.

Der Einfachheit halber werden die Jahresarbeitstage im Folgenden auf 203 Tage abgerundet.

[103] Siehe „Anlage 5 - Beispielvorlage des Beratungszentrums des Bundes zur Berechnung der Jahresarbeitszeit bei Vollzeitbeschäftigung“ https://www.orghandbuch.de/SharedDocs/downloads/Webs/OHB/DE/Anlage_5.xls?__blob=publicationFile&v=6; Letzter Abruf: 21.07.2025

Im weiteren Verlauf des Abschnitts wird generell mit gerundeten Ganzzahlen gerechnet.

open.bydata – Das Open-Data-Portal Bayerns

Die Bayerische Agentur für Digitales, *byte*, beschreibt den Erstellungsprozess ihres Open-Datal-Portals, Open.ByData, sehr transparent[104].

Im Projektverlauf wurde eine agile Vorgehensweise genutzt, um zuerst ein MVP zu entwickeln und darauf aufbauend eine Produktivversion zu entwickeln.

Die Agentur *byte* stellte dabei intern einen Product Owner und entwickelte kollaborativ das UX/UI-Design. Die Software-Implementierung erfolgte extern durch Fraunhofer FOKUS auf Basis von Piveau.

Hiermit ist der Entwicklungsprozess mit dem des Datenatlas vergleichbar[105].

[104] https://www.byte.bayern/was-wir-machen/open-data-bayern; Letzter Abruf: 18.08.2025

[105] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025

Als zeitlichen Ablauf gibt *byte* folgendes an: „in wenigen Monaten haben wir Anfang 2023 unser Minimum Viable Product (MVP) aufgebaut und Ende Mai live genommen“[106].

Die Produktivversion von Open.ByData wurde binnen vier Monaten entwickelt und wird seitdem interativ anhand von Nutzendenanforderungen weiterentwickelt.

[106] https://www.byte.bayern/was-wir-machen/open-data-bayern; Letzter Abruf: 18.08.2025

Die Nachnutzbarkeit des Portals ist möglich. So bietet *byte* jeder Verwaltungsstelle Bayerns eine eigene Präsenz innerhalb von Open.ByData an.

Der Entwicklungszeitraum von Open.ByData wird zum Zeitpunkt der Erstellung des Gutachtens auf 2,5 Jahre festgelegt.

Zum Zeitpunkt der Erstellung des Gutachtens beschäftigt *byte* fünf VZÄ[107]. Laut Eigenangabe ist eine Person als Product Owner gebunden, während weitere Personen zusammen mit dem Dienstleister an der UX arbeiten (s.o.).

[107] https://open.bydata.de/home

Der Einfachheit halber wird im Folgenden ein Personaleinsatz von fünf VZÄ angenommen, welche den des Dienstleisters einschließt. Für den Betrieb und die Weiterentwicklung werden zwei VZÄ angenommen.

Für die Erstellung des ersten MVP erscheinen vier Monate nach den internen Angaben plausibel. Hinzu kommen weitere vier Mo-

nate für die Erstellung des Produktivsystems. Insgesamt hat *byte* demnach acht Monate in die Erstellung von OPEN.BYDATA investiert und betreibt es seitdem. Daraus ergibt sich folgende Kostenschätzung.

Tabelle 5.4: Kostenschätzung open.bydata

Zeitaufwand (Erstellung; Feb. 2023 - Sep. 2023)	Betrag
8 Monate mit 5 VZÄ	5.560 h
davon Entwicklung (80%)	4.448 h
davon Projektmanagement (20%)	1.112 h
Stundensatz (5 VZÄ)	Kosten
Entwicklung: Branchendurchschnitt (102 €)	453.696 €
PM: Beratung und Management (120 €)	133.440 €
Zwischensumme:	*587.136 €*
Zeitaufwand (Betrieb; Okt. 2023 - Jul. 2025)	Betrag
2023 (3 Monate)	417 h
2024 (12 Monate)	1.665 h
2025 (6 Monate)	833 h
Stundensatz (2 VZÄ)	Kosten
davon Entwicklung (80%)	475.728 €
davon Projektmanagement (20%)	139.920 €
Zwischensumme:	*615.648 €*
Gesamtkosten zum Juli 2025:	**1.202.784 €**

CrossAsia – ein Repository mit hohem Eigenentwicklungsanteil

Die Aufwandsermittlung erfolgt auf Grundlage der Kenntnisse des Autors des Gutachtens, der die IT-seitige Verantwortung für CROSSASIA über ca. sechs Jahre trug. Die aktuelle Version des Systems wird seit ca. 2016 agil entwickelt und basiert auf FEDORA[108].

Intern ist ein VZÄ dauerhaft für den Betrieb und die Weiterentwicklung des *Repository-Systems* gebunden. Hinzu kommen Projektbezogene Entwickler, welche nicht über die gesamte Projektlaufzeit aktiv sind.

Zur besseren Nachvollziehbarkeit der Kostenkalkulation wird dieses zusätzliche Engagement mit 1,25 VZÄ über die gesamte Lebenszeit von CROSSASIA angesetzt, so dass sich 2,25 VZÄ ergeben.

DER ENTWICKLUNGSZEITRAUM von CROSSASIA wird auf 9 Jahre zum Zeitpunkt der Erstellung des Gutachtens festgelegt, um Vorlaufzeiten im Jahr 2016 zu berücksichtigen.

Aufgrund des Alters und der stetigen Weiterentwicklung des Systems kann jedoch nicht mehr zwischen Einführungs- und Betriebskosten wie bei OPEN.BYDATA unterschieden werden.

Die Kostenschätzung für CROSSASIA folgt der bereits etablierten Methodik und ergibt das in Tabelle 5.5 dargestellte Ergebnis. Auf-

[108] David Zellhöfer. Agilität und Weltkulturerbe: Erfahrungen mit sieben Jahren agilen Methoden an der Staatsbibliothek zu Berlin – Eine Fallstudie I/II. *ABI Technik*, 41(3):194–201, 2021. ISSN 2191-4664, 0720-6763. DOI: 10.1515/abitech-2021-0032. https://www.degruyter.com/document/doi/10.1515/abitech-2021-0032/html. Letzter Abruf: 16.04.2025; and David Zellhöfer, Oliver Schöner, und Gerrit Gragert. *Building Library Information Systems in Times of Vanishing Developer Resources – A Fedora-4-based Approach*. 40th European Library Automation Group Systems Seminar, Fachvortrag, 2019

grund der internen Kenntnisse des Autors wird der Projektsteuerungsaufwand allerdings mit 25% der Gesamtkosten angesetzt.

Der Gesamtbetrag bezieht sich ausschließlich auf die IT-Entwicklung und lässt bibliothekarische und weitere Aufwände außen vor.

Tabelle 5.5: Kostenschätzung CrossAsia

Zeitaufwand (2,25 VZÄ; Jun. 2016 - Jul. 2025)	Betrag
9 Jahre (à 1.665 h)	33.716 h
davon Entwicklung (75%)	25.287 h
davon Projektmanagement (25%)	8.429 h
Stundensatz	**Kosten**
Entwicklung: Branchendurchschnitt (102 €)	2.579.274 €
PM: Beratung und Management (120 €)	1.011.480 €
Gesamtkosten zum Juli 2025:	**3.590.754 €**

Der Datenatlas – eine weitgehende Eigenentwicklung

Da zum DATENATLAS kaum belastbare Informationen zu ermitteln sind, muss die folgende Grobschätzung mit der entsprechenden Zurückhaltung aufseiten der Lesenden interpretiert werden.

Laut der Bundesdruckerei (2022) startete das Projekte Ende 2021. Um eventuelle Projektvorlaufzeiten berücksichtigen zu können und das Alter des Systems festlegen zu können, erscheint ein Entwicklungszeitraum von 3 Jahren plausibel – also ein Start in Mitte 2022.

DER ENTWICKLUNGSZEITRAUM des DATENATLAS kann demnach auf drei Jahre zum Zeitpunkt der Erstellung des Gutachtens festgesetzt werden.

Auch wenn der DATENATLAS nicht die Reife der anderen vorgestellten Systeme erreicht, erscheint es plausibel anzunehmen, dass ein ähnlich hoher Personaleinsatz wie bei OPEN.BYDATA gerechtfertigt ist.

Nach Kenntnis des Autors befindet sich der DATENATLAS im Juli 2025 noch in aktiver Entwicklung, so dass ein Abschmelzen der beteiligten Personalressourcen wie bei OPEN.BYDATA nicht anzunehmen ist.

Aus den Pressemitteilungen der BUNDESDRUCKEREI und Gesprächen mit Mitarbeitenden der Datenlabore wurde ermittelt, dass begleitend zur Systementwicklung eine Dateninventur in der BUNDESVERWALTUNG durchgeführt wird.

Diese wird aus Gründen der besseren Vergleichbarkeit in die Kostenschätzung einbezogen, jedoch mit einem niedrigeren Stundensatz von 100 € angesetzt, da diese Art der Erhebung keine Fähigkeiten auf dem Niveau des höheren Dienst voraussetzt. Zur weiteren Vereinfachung wird jedoch weiterhin die Jahresarbeitszeit des höheren Dienst herangezogen.

Aufgrund der mangelhaften Datenlage wird die Dateninventur

als Vollzeittätigkeit für ein VZÄ angenommen. Diese Annahme ist sehr wohlwollend, da dies bedeutet, dass ein VZÄ an jedem Arbeitstag für ein gesamtes Jahr entweder in Behörden vor Ort war oder die Inventur dokumentiert hat.

Der üblichen Schätzmethodik folgend ergibt sich die in Tabelle 5.6 dargestellte Kostenschätzung für den DATENATLAS.

Diese Kosten beinhalten keinerlei Aufwände innerhalb der BUNDESVERWALTUNG.

Tabelle 5.6: Kostenschätzung Datenatlas

Zeitaufwand (4+1 VZÄ; Jun. 2022 - Jul. 2025)	Betrag
3 Jahre (4 VZÄ, à 1.665 h)	19.980 h
1 Jahr (1 VZÄ Dateninventur)	1.665 h
Zwischensumme:	*21.645 h*
davon Entwicklung (80%)	15.984 h
davon Dateninventur (80%)	1.332 h
davon Projektmanagement (20%)	4.329 h
Stundensatz (4+1 VZÄ)	**Kosten**
Entwicklung: Branchendurchschnitt (102 €)	1.630.368 €
Dateninventur (100 €)	133.200 €
PM: Beratung und Management (120 €)	519.480 €
Gesamtkosten zum Juli 2025:	**2.283.048 €**

Fazit Wirtschaftlichkeitsbetrachtung

Da dem Autor des Gutachtens keine konkreten Erstellungskosten für den DATENATLAS vorliegen, obliegt es besser informierten Lesenden, zu bewerten, inwiefern die tatsächlichen Kosten in einem angemessenen Verhältnis zu den Kostenschätzungen liegen.

Tabelle 5.7 stellt die Kostenschätzungen der drei betrachteten Systeme gegenüber und weist zusätzlich die Entwicklungskosten pro Jahr aus, die sich aus der Division der Gesamtkosten mit dem Entwicklungszeitraum ergibt.

System	Technische Basis	Entwicklungs-start	Entwicklungs-zeitraum	Gesamtkosten	Kosten p.a.
OPEN.BYDATA	PIVEAU	2023	2,5 Jahre	1.202.784 €	481.114 €
CROSSASIA	FEDORA	2016	9,0 Jahre	3.590.754 €	398.973 €
DATENATLAS	*unbekannt*☆	2023	3,0 Jahre	2.283.048 €	761.016 €

☆ Die technische Basis des DATENATLAS ließ sich nicht ermitteln. Das Gutachten beinhaltet jedoch viele Indizien, die auf eine proprietäre Eigenentwicklung hinweisen.

Tabelle 5.7: Vergleich der Kostenschätzungen zum Juli 2025

Aus der Tabelle wird deutlich, dass aller Wahrscheinlichkeit nach ausschließlich der DATENATLAS eine proprietäre Eigenentwicklung darstellt. Neben den bereits angeführten Kritikpunkten bezüglich

der Nachnutzbarkeit bzw. *Digitalen Souveränität*[109] oder des Funktionsumfangs und der technischen Reife[110] zeigt sich, dass der Datenatlas die höchsten Entwicklungskosten pro Jahr aufweist.

Darauf folgt Open.ByData, welches ca. 63% dieser Kosten verursacht. Dahinter liegt CrossAsia. Die etwas niedrigeren Entwicklungskosten pro Jahr bei CrossAsia erscheinen insofern plausibel, da das System durch die lange Entwicklungsgeschichte und die Nachnutzung vieler in der Staatsbibliothek zu Berlin bereitgehaltenen Software-Systeme Synergie-Effekte besser nutzen kann.

Die deutlich höheren Kosten des Datenatlas lassen sich mit großer Wahrscheinlichkeit darauf zurückführen, dass die Bundesdruckerei hier auf eine komplette Eigenentwicklung gesetzt hat. Aus Gründen der Wirtschaftlichkeit ist das kaum nachvollziehbar, da vergleichbare *Repository-Systeme* seit langem als *Open Source* vorliegen und bereits erfolgreich in der Bundesverwaltung, u.a. in Form von GovData, oder den Ländern[111] im Produktiveinsatz sind.

Wie in Kapitel 4 und Abschnitt 5.4 dargelegt wurde, existieren Import-Funktionen aus GovData in den Datenatlas, so dass davon auszugehen ist, dass das Entwicklungsteam Kenntnisse über *Open Source*-Lösungen wie GovData auf Grundlage von CKAN spätestens in der Projektplanung erworben hat.

Der wirtschaftliche Totalverlust der im Datenatlas erfassten Daten kann durch den in Abschnitt 5.5 beschriebenen *vendor lock-in* nicht ausgeschlossen werden.

Wirtschaftlich nachvollziehbare Sachgründe gegen die Nachnutzung eines leistungsfähigen *Repository-Systems* sind aus Sicht des Autors nicht erkennbar.

Aus Einzelgesprächen konnte der Autor ermitteln, dass beispielsweise Rechte- und Rollen-Konzepte aufwändig implementiert werden mussten, was dafür spricht, dass kein etabliertes *Repository-System* genutzt wurde. Denn diese Funktionen sind derart basal, dass sie 1) von jedem *Repository-System* angeboten werden und 2) dort auch sicher implementiert sind.

Eine Entscheidung für ein etabliertes System hätte hier sowohl aus Wirtschaftlichkeits- als auch IT-Sicherheitsüberlegungen erfolgen müssen.

Hinzu kommt, dass der entwickelte Datenatlas trotz unverhältnismäßig hoher (geschätzter) Kosten funktional deutlich hinter die anderen Systeme zurückfällt und nur selten den *Stand der Technik* erreicht.

Wirtschaftlich und funktional bessere Entscheidungen haben hier Bayern (Open.ByData) und Brandenburg (Datenadler), GovData und die Zivilgesellschaft getroffen, wie u.a. in Abschnitt 5.2 ausführlich diskutiert wurde. Die dort genutzten Sys-

[109] Siehe Abschnitt 5.5.

[110] Siehe Abschnitt 5.2f..

[111] Siehe Abschnitt 5.2.

teme bauen alle auf bereits existierenden Lösungen auf, welche für die spezifischen Anwendungsfälle angepasst wurden.

Eine kursorische Prüfung dieser Systeme zeigte nicht einmal im Ansatz ähnliche Defizite auf, wie die im Rahmen der Kurzevaluierung über User Journeys in Kapitel 4 präsentierten, teils erheblichen, Mängel des DATENATLAS.

EIN ÄHNLICHES BILD zeichnet sich ab, wenn man die in Abschnitt 3.7 eingeführten sieben Teilaspekte der Wirtschaflichkeitsbetrachtung der VV-BHO (2025) heranzieht.

So ergibt sich ein klar erkennbarer *Zielkonfikt*: Während der Auftraggeber an der schnellen Bereitstellung einer Lösung auf dem *Stand der Technik* interessiert ist, ist es für den Dienstleister wesentlich attraktiver, eine eigene Lösung zu implementieren, um einen höheren Umsatz zu generieren – insbesondere wenn das vertragliche Rahmenwerk eine agile Entwicklung mit Abrechnung auf Stundenbasis – also einen *Dienstvertrag* – vorsieht.

Im Falle eines *Werkvertrags* liegt es hingegen im Interesse des Dienstleisters zu geringen Eigenkosten ein funktionstüchtiges Werk zu liefern. Wie im Rahmen des Gutachtens dargelegt wurde, wäre ein Werkvertrag aufgrund der gut erforschten Kernfunktionen der DATENATLAS angemessen und wirtschaftlich gewesen.

Bei der Recherche zum DATENATLAS konnte nur ermittelt werden, dass agil vorgegangen wurde[112]. Aus der Erfahrung des Gutachters tendiert die BUNDESVERWALTUNG jedoch häufig dazu, Dienstverträge auszuschreiben. Aufgrund der agilen Vorgehensweise erscheint dies auch in diesem Fall plausibel.

[112] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025

Wäre zuerst eine *Analyse der Ausgangslage und des Handlungsbedarfs*, wie sie §7(2.1) VV-BHO (2025) vorsieht, erfolgt, hätte eine Entscheidung anhand des Stands der Technik – der wie in Kapitel 3 illustriert, ohne Probleme hätte ermittelt werden können – aus Wirtschaftlichkeitsgründen zu der Nachnutzung einer bestehenden Lösung in Kombination mit einer Dateninventur führen müssen.

Demnach wäre auch eine Ausschreibung nach den Grundsätze der Vergabe nach §97(4) GWB (2024) in mindestens zwei Fachlosen zwingend gewesen, da der Aufteilung keinerlei technische oder wirtschaftliche Gründe entgegenstehen, d.h. ein Dienstvertrag für die Dateninventur, ein Werkvertrag für die Kernsoftware DATENATLAS und ein optionaler Dienstvertrag für eventuelle Anpassungen. Die letzten beiden Lose wären aus technischen Gründen zusammen zu vergeben, es sei denn man würde auf eine *Open-Source-Software* setzen.

Dass sich daraus *negative finanzielle Auswirkungen auf den Haushalt* – zumindest anhand der Grobschätzung – ergeben, wurde bereits in Tabelle 5.7 gezeigt.

Die zuständigen Stellen im BMF sollten dem hier die tatsächlich angefallenen Kosten gegenüberstellen.

Aus Platzgründen soll an dieser Stelle auf die Vorstellung der anderen vier Teilaspekte verzichtet werden, zumal deren Verletzung implizit aus den anderen Desiderata abgeleitet werden kann.

Aufgrund des allgemeinen Zustands des Datenatlas erscheint eine Diskussion der eigentlich zu implementierenden Erfolgskontrollen müßig, da diese die im Rahmen des Gutachtens aufgezeigten Mängeln hätten entdecken müssen.

Abschließend kann festgehalten werden, dass selbst die durch den Dienstleister an die Presse kommunizierten Ziele[113] kaum bis gar nicht erreicht wurden.

[113] Bundesdruckerei (2022, 2025)

6
Fazit und Ausblick

Das vorliegende Gutachten mit einem Umfang von mehr als 140 Seiten wurde mit einem Gesamtaufwand von ca. einer Arbeitswoche – ohne Rückgriff auf generative KI-Anwendungen – in den Randzeiten der Vollzeittätigkeit des Gutachters erstellt.

Hinzu kommt, dass für die Erhebung der Screenshots nur ca. 30 Minuten am Produktivsystem des Datenatlas zur Verfügung standen[1].

[1] Siehe Kapitel 4.

Bereits daraus wird deutlich, dass die vorliegenden Defizite und Desiderata leicht zu ermitteln waren und entsprechend einfach hätten abgestellt werden können.

Vorbemerkung

Die Erstellung des Gutachtens erfolgte *pro bono*, da der Autor den Aufbau von Datenkompetenzen, die nachhaltige Verwaltung und Kontextualisierung von Daten, z.B. in Form von *Linked (Open) Data*, und den Austausch dieser als essenziell für die Zukunftsfähigkeit der Öffentlichen Verwaltung – nicht nur mit Hinblick auf die bedarfsgetriebene Anwendung von Künstlicher Intelligenz – betrachtet.

Da der Autor weder an der Konzeption noch der Umsetzung des Datenatlas beteiligt war, wurde das Gutachten ohne interne Projektkenntnisse erstellt.

Es stützt sich ausschließlich auf die vorliegenden Screenshots (Stand Juli 2025), die Veröffentlichungen der Bundesdruckerei und die Aussagen einzelner Datenlabore sowie des Product Owners von GovData, der ebenfalls über keine tieferen, internen Projektkenntnisse verfügt.

Die vorliegende Analyse und Erhebung der Desiderata basiert im wesentlichen auf den Lehrinhalten der Vorlesung und Übung *Information Retrieval*, welche der Autor ab 2009 an der *Brandenburgischen Technischen Universität Cottbus-Senftenberg* am Lehrstuhl „Datenbanken und Informationssysteme“ im Rahmen des Bachelor-Studiums Informatik betreute.

Die Vorlesungsinhalte orientieren sich primär an dem frei verfügbaren Lehrbuch von Henrich (2008), welches unter CC-BY-NC-ND-Lizenz vorliegt[2].

[2] https://www.uni-bamberg.de/fileadmin/minf/Dateien/Publikationen/2008/henrich-ir1-1.2.pdf; Letzter Abruf: 21.07.2025

Zusätzlich wurden im wesentlichen etablierte Lehrbücher und hauptsächlich im Internet verfügbare Quellen herangezogen, welche sich ohne den Einsatz von *Künstlicher Intelligenz* oder den Zugriff auf Spezialbibliotheken ermitteln ließen.

Die einzige Ausnahme bildet die Kolumne von Zellhöfer (2023), welche hinter der Paywall des *Tagesspiegels* liegt.

FORMAL verfügt der Autor des Gutachtens als promovierter Informatiker und Informationswissenschaftler[3] über keine *verwaltungswissenschaftliche* Ausbildung außer seiner Erfahrungen aus praktischen Tätigkeiten über sieben Jahre in nachgeordneten Bundesbehörden. Dies sollte insbesondere bei rechtlichen Aussagen beachtet werden.

[3] Siehe Kapitel 1.

Mit dem Ruf auf die Professur für „Digitale Innovation in der öffentlichen Verwaltung“ im Jahr 2020 widmet er sich nun in Vollzeit der ganzheitlichen Verwaltungsdigitalisierung auf Landes- und Bundesebene.

DEM STAND DER TECHNIK, wie er in Kapitel 3 ausführlich dargestellt wurde, folgend, muss man die Entwicklungaufgabe des DATENATLAS als gut verstandenes und seit den 1980er häufig beschriebenes Problem bewerten.

Das Gesamtprojekt DATENATLAS zerfällt in zwei Teilgebiete: die Software-Entwicklung und die Dateninventur.

Diese zwei Teilprojekte wären nach §97(4) GWB (2024) in mindestens zwei Fachlosen zu vergeben[4].

[4] Siehe Abschnitt 5.8.

Hierbei ist die Dateninventur aufgrund der Vielzahl an Bundesbehörden als wesentlich komplexer als die Software-Entwicklung zu betrachten, welche sich an einer Vielzahl an existierenden Produkten orientieren kann, wie in Abschnitt 3.2 und Abschnitt 5.2 gezeigt wurde.

Damit liegt ein Mindestumfang an zu erwartenden Funktionalitäten für den DATENATLAS seit Projektbeginn vor.

Inwiefern der DATENATLAS diesen Mindestmaßstab erreicht, wird im nächsten Abschnitt diskutiert.

Im darauf folgenden Abschnitt 6.2 werden zwei ganzheitlichere Bewertungsansätze für das Gesamtprojekt herangezogen und im Anschluss daraus Handlungsempfehlungen abgeleitet.

6.1 *Bewertung der bereitgestellten Funktionen*

Nach Darstellung des Dienstleisters wurden agile Entwicklungsmethoden verwendet, um den DATENATLAS in den „MVP-Status“[5] zu überführen. MVP steht dabei für das *Minimum Viable Product*[6] – also das minimal brauchbare Produkt.

[5] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. `https://tinyurl.com/bdr-pm1`. Letzter Abruf: 21.07.2025

[6] Siehe Abschnitt 3.4.

Ein MVP dient typischerweise dazu, *schnell und ohne großen Personaleinsatz* ein *neues* Produkt mit einem minimalen Funktionsumfang zu entwickeln, um dessen Akzeptanz bei den potenziellen Nutzenden zu evaluieren. Anhand der gewonnenen Erkenntnisse

wird das Produkt weiterentwickelt, um die Nutzungsanforderungen bestmöglich umzusetzen.

Fraglich dabei ist, warum im Projektkontext Datenatlas *überhaupt* ein MVP entwickelt werden sollte, obwohl eine Vielzahl an empirisch gesicherten wissenschaftlichen Erkenntnissen zur *Informationsrecherche*[7] und diverse, praktische Lösungsansätze in Form nachnutzbarer Software-Systeme bereits zu Beginn des Entwicklungsprojekts vorlagen[8].

[7] Siehe Abschnitt 3.5.

[8] Siehe Abschnitt 5.2.

Zudem sind die *minimalen Use Cases der Bundesverwaltung* im Kontext des Datenatlas, wie in Kapitel 2 dargestellt, einfach zu ermitteln: einerseits die *Recherche von Metadaten* primär durch Laien-Nutzende und andererseits die *Erfassung von Metadaten* durch entsprechend ausgebildete Mitarbeitende.

Möchte man von einem MVP sprechen, so ist zu kritisieren, dass dieses Produkt gerade Mängel an seiner *Kernfunktionalität*, der *Informationsrecherche*, aufweist – also nicht die minimale Menge an Funktionen bereitstellt, damit das Produkt „viable" – also brauchbar – wird.

Im Prinzip unterstützt der Datenatlas im Rahmen der Recherche ausschließlich die *Known-Item-Search*[9], welche sich primär an Nutzende richtet, die ihr *Informationsbedürfnis* exakt spezifizieren können. Dass der Großteil der Mitarbeitenden der Bundesverwaltung unter diese Nutzendengruppe fällt, ist nicht vom aktuellen Forschungsstand gedeckt.

[9] Siehe Abschnitt 3.5.

Explorative Suchstrategien[10] sind im Datenatlas nur beim Datenimport[11] verfügbar, obwohl gerade bei diesem Szenario anzunehmen ist, dass entsprechend qualifizierte Mitarbeitende eher eine *Gerichtete Suche* nutzen würden.

[10] Siehe Abschnitt 3.5.

[11] Siehe Abschnitt 4.5.

Eine vom Dienstleister geplante „ressortübergreifende Suche und Exploration der Datenbestände"[12] ist zum aktuellen Zeitpunkt nicht im Datenatlas umgesetzt worden.

[12] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025

Zusätzlich zu der ausbaufähigen Unterstützung verschiedener *Informationssuchstrategien* ergeben sich weitere Mängel im Bereich der Usability[13], der Datenmodellierung und der Datenqualitätssicherung[14].

[13] Siehe Abschnitt 5.3.

[14] Siehe Abschnitt 5.4.

Selbst wenn sich die in Kapitel 4 beschriebenen Usability-Probleme beheben ließen, so bleiben die funktionalen Defizite im Rahmen der *Informationsrecherche* derart grundlegend, dass sich kein Anwendungsfall für den Datenatlas erkennen lässt, welcher auf breite Nutzendenakzeptanz treffen wird.

Hinzu kommt, dass eine Adressierung der Probleme im Bereich der *Informationsrecherche* zur Folge hätte, dass alle Funktionen, die unmittelbar mit dem Retrieval-Prozess verbunden sind, überarbeitet bzw. ausgetauscht werden müssten[15].

[15] Siehe Abschnitt 5.2.

Wäre von vornherein ein *Information-Retrieval-System*, wie in Abschnitt 5.2 beschrieben, als Kern eingesetzt worden, wäre eine

potenzielle Weiterentwicklung u.U. denkbar.

Inwiefern sich diese *wirtschaftlich* umsetzen lässt, wird in Abschnitt 6.3 thematisiert.

Vergleicht man den Datenatlas beispielsweise mit GovData wird deutlich, wie stark dieser funktional hinter das *Open Data-Portal* des Bundes oder vergleichbare Dienste[16] zurückfällt.

[16] Siehe Abschnitt 5.2.

Es ist davon auszugehen, dass GovData dem Dienstleister bekannt ist, da aus diesem u.a. Daten importiert werden können, und dieses Portal als funktionales Vorbild hätte dienen müssen.

In diesem Licht von einer Notwendigkeit der MVP-Entwicklung des Datenatlas zu sprechen, ist unangemessen, da alle notwendigen Funktionen zur Umsetzung der *minimalen Use Cases der Bundesverwaltung* bei der Nachnutzung einer der in Tabelle 5.2 genannten Systeme „out of the box“ zur Verfügung gestanden hätten.

Dass dieser Weg gangbar ist, lässt sich am Beispiel des Portals Open.ByData zeigen, dessen Team binnen vier Monaten einen MVP entwickelte[17], welcher den *Stand der Technik* erreicht und damit den Datenatlas hinter sich zurück lässt.

[17] Siehe Abschnitt 5.8.

Ob eine vergleichbare Entwicklungsgeschwindigkeit beim funktional schwächeren Datenatlas an den Tag gelegt wurde, entzieht sich dem Wissen des Autors.

Die Anfragemöglichkeiten des Datenatlas entsprechen bei Weitem nicht dem *Stand der Technik* des Jahres 1998 aus, wie in Tabelle 5.1 veranschaulicht wird.

Tatsächlich bietet der Datenatlas weniger Möglichkeiten der Anfrageformulierung als das bereits zu Beginn von Abschnitt 3.2 vorgestellte Beispiel eines OPAC (siehe Abb. 6.1).

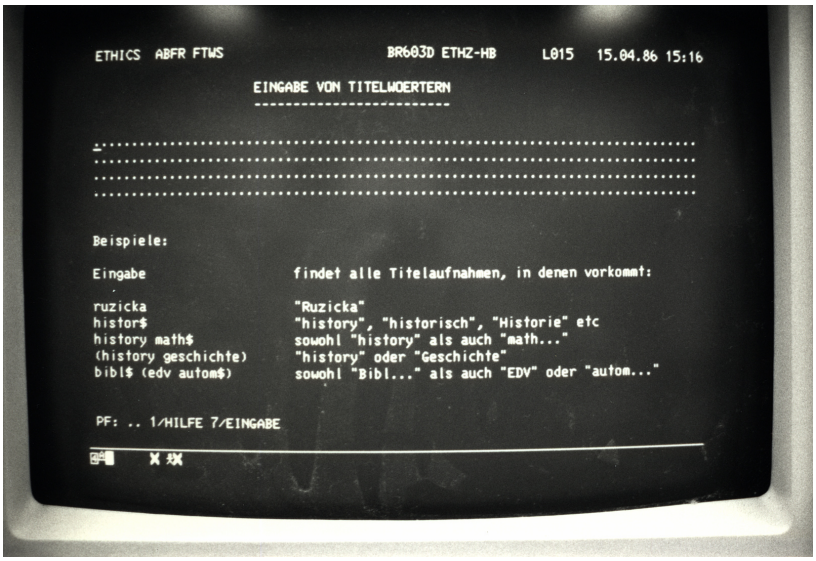


Abbildung 6.1: Titelwortabfrage im historischen digitalen Katalog der ETH Zürich (15.04.1986) CC BY SA ETH Zürich
DOI: 10.3932/ethz-a-000014637

Abbildung 6.1 macht deutlich, dass bereits im Frühjahr 1986 Funktionen wie die Phrasensuche, Wildcards und Boolesche Opera-

toren möglich waren – Anfragemöglichkeiten, die der DATENATLAS nicht unterstützt.

Folglich entspricht der Funktionsumfang des DATENATLAS in diesem Bereich nicht einmal dem des Jahres 1986.

Desiderata außerhalb des ursprünglichen Projektzuschnitts

Neben den funktionalen Defiziten des DATENATLAS existieren weitere Desiderata, welche klar außerhalb des Projektes liegen, jedoch zukünftig Beachtung finden müssen.

ES IST WENIG ZIELFÜHREND, eine Kernkomponente wie den (stark überarbeiteten) DATENATLAS allein für die interne Nutzung der Bundesebene bereitzustellen, da die Datenerfassung häufig auf Kommunal- und Landesebene erfolgt.

Zur Gewährleistung der *Interoperabilität* sollte *bundesweit* eine geeignete Lösung genutzt werden, welche den nahtlosen Übergang in die *Open-Data-Portale* des Bundes und der Länder, wie z.B. GOVDATA oder OPEN.BYDATA, ermöglicht.

IM RAHMEN DER STANDARDISIERUNG ist zu erwarten, dass der Gesetzgeber, wie in §6 OZG (2024) beschrieben, geeignete Architekturvorgaben, Qualitätsanforderungen und Interoperabilitätsstandards bis Ende 2026 vorlegt[18].

[18] Siehe Abschnitt 3.7.

Hier bleibt zu hoffen, dass in diesem Zuge auch verbindliche Anforderungen an die *Datenqualität* für *Open Data* und interne Verwaltungsdaten definiert werden.

GERADE MIT BLICK auf die *Langzeitarchivierung*[19] kritisch zu sehen ist die Konzeption des DATENATLAS als reines Metadaten-Portal. Um die Nachvollziehbarkeit des Daten-getriebenen Verwaltungshandels sicherzustellen, müssten eigentlich alle *Content-Daten* der Daten-bereitstellenden Behörden archiviert werden.

[19] Siehe Abschnitt 5.7.

Die Implementierung der Langzeitarchivierung lässt sich am effizientesten durch die Verwendung eines bereits bestehenden *Repository-Systems* erreichen.

6.2 Ganzheitliche Bewertung des Datenatlas

Aufgrund seiner besseren Zugänglichkeit im Vergleich mit den Normen der DIN EN ISO 9241-Familie und seiner ganzheitlicheren Bewertungsperspektive, soll der *Servicestandard* im Folgenden für eine erste Bewertung des DATENATLAS herangezogen werden.

Auch wenn der *Servicestandard* erst nach dem Projektbeginn des DATENATLAS veröffentlicht wurde, gleicht er doch in weiten Teilen der bereits seit 2020 verfügbaren Version, wie in Abschnitt 3.4 beschrieben wurde.

Der Datenatlas aus Sicht des Servicestandards

Der *Servicestandard* nennt 13 Kriterien, um Dienste verlässlicher, verständlicher und effizienter zu gestalten. Von den 13 Kriterien können fünf nicht im Rahmen dieses Gutachtens bewertet werden, wie in Abschnitt 3.4 begründet wurde.

Die Tabellen 6.1 und 6.2 stellen die übrigen acht Kriterien ihrem Umsetzungsgrad im DATENATLAS gegenüber, insofern dieser ermittelt werden konnte.

Kriterium	Umsetzungsgrad
1. Nutzende verstehen und Bedürfnisse erkennen	Die Erforschung von Suchprozessen wird seit langem wissenschaftlich betrieben, wie in Abschnitt 3.5 ausgeführt wurde. Obwohl diese Erkenntnisse klar motivieren, wann gerichtete oder explorative Suchstrategien unterstützt werden sollten, wurde dies nicht beachtet. Da die zugrundeliegenden empirischen Studien teils hunderte an Personen beobachteten, erscheint es unwahrscheinlich, dass die Mitarbeitenden der BUNDESVERWALTUNG hier grundsätzlich anders vorgehen würden. Inwiefern Nutzende im Rahmen der Entwicklung befragt wurden, konnte nicht ermittelt werden.
2. Problem beschreiben und Ziele bestimmen	Die Ziele des DATENATLAS werden durch die *Bundesdatenstrategie*★ vorgegeben, welche die Bundesdruckerei (2022, 2025) weiter präzisiert und selber festgelegt. Die Erreichung der letztgenannten Ziele wird in Abschnitt 6.2 thematisiert.
4. Lösungen entwickeln, testen, anpassen und Fachwissen einbinden	Aufgrund der teils erheblichen Abweichungen vom *Stand der Technik* (siehe Kapitel 5) scheint extern zur Verfügung stehendes Fachwissen nur in geringem Maß eingebunden worden zu sein. Ob die Datenlabore mit ihrer Expertise eingebunden wurden, wird in Abschnitt 6.3 ergründet.
5. Bestehendes wiederverwenden und Neues gemeinsam gestalten	Obwohl es eine Vielzahl an Lösungen gibt (siehe Abschnitt 5.2), wurden diese nicht nachgenutzt. Stattdessen wurde auf eine proprietäre Eigenentwicklung gesetzt. In welchem Umfang weitere Einrichtungen als potenzielle Kunden in die Anforderungsanalyse involviert wurden, entzieht sich der Kenntnis des Autors. Eine gemeinsame Planung mit verwandten Plattformen wie GOVDATA erfolgte – mit Ausnahme des Datenimports – nicht bzw. in sehr geringem Maße.

★ vgl. Bundesministerium für Digitales und Verkehr et al. (2023)

Tabelle 6.1: Umsetzung der Kriterien des Servicestandards I

Kriterium	Umsetzungsgrad
7. Offene Standards beachten und Schnittstellen bereitstellen	Nach Aussagen der Datenlabore stellt der Datenmodellkern des DATENATLAS eine Teilmenge von DCAT-AP (siehe Abschnitt 5.4) dar, wurde jedoch nicht offengelegt und entzieht sich damit einer automatisierten Validierung. Dadurch wird der automatische Austausch von Daten erschwert. Gängige Schnittstellen, wie sie vergleichbare Portale bieten, sind bis auf den Import aus GOVDATA nicht vorhanden. Aufgrund diverser Probleme mit dem verwendeten Datenmodell und der Datenqualität ist die Maschinenlesbarkeit von Daten kaum zu erwarten, wie in Abschnitt 5.4 begründet wurde.
10. Open Source nutzen und Code teilen	Der DATENATLAS liegt nicht als *Open Source* vor. Eine Nachnutzung bzw. Anpassung durch weitere Einrichtungen wird durch den entstandenen *vendor lock-in* erschwert. In der Folge mindert der DATENATLAS die *Digitale Souveränität* der BUNDESVERWALTUNG (siehe Abschnitt 5.5).
12. Wirkung messen und auf Ergebnissen aufbauen	Eine kontinuierliche Evaluierung wäre wichtig, um die Vollständigkeit und Qualität der Dateninventur messen zu können. Inwiefern diese erfolgt, ist unbekannt. Eine Möglichkeit der Wirkungsmessung im Bereich der Usability – der Inspektionstest – wurde in Abschnitt 4.7 erörtert. Diese Testmethode förderte erhebliche Mängel zutage. Da solch ein Test auch für den Dienstleister möglich gewesen wäre durch diesen aber vermutlich nicht erfolgte, bestehen Zweifel daran, inwiefern eine parallele Evaluierung zur Software-Entwicklung implementiert ist.
13. Rechtliche Hürden erkennen und Regelungen verbessern	Es ist anzunehmen, dass die Adressierung rechtlicher Hürden außerhalb des Projektauftrags liegt. Beispielhafte Desiderata aus diesem Bereich finden sich in Abschnitt 6.1.

Tabelle 6.2: Umsetzung der Kriterien des Servicestandards II

Der Datenatlas aus Sicht des Dienstleisters

Bereits in Kapitel 2 wurde die Pressemitteilung der BUNDESDRUCKEREI zitiert, welche seine Zielstellung *direkt* aus der *Bundesdatenstrategie* von 2023 ableitet (siehe unten):

> Wir erstellen einen Datenatlas Bundesverwaltung, der Daten aller Ministerien und ihrer Geschäftsbereiche auf Metadatenebene zeigt. Damit schaffen wir Transparenz über den vorhandenen Datenbestand. [...] Der Datenatlas nutzt und ergänzt bestehende Verwaltungsdatenübersichten wie die Verwaltungsdaten-Informationsplattform (VIP) des Statistischen Bundesamtes, die Registerlandkarte des Bundesverwaltungsamtes oder das Metadatenportal GovData zu offenen Daten von Bund, Ländern und Kommunen. Für den Datenatlas sind in den Ministerien die Datenlabore zuständig.[20]

[20] https://www.bundesdruckerei.de/de/innovation-hub/projekt-datenatlas; Letzter Abruf: 21.07.2025

Die Auslassung des Zitats ist interessant, da die *Bundesdatenstrategie* das *zu erreichende Ziel* auf Seite 11 sehr klar benennt:

> Ergänzend werden wir den Aufbau eines Datenpools der Bundesverwaltung für maschinenlesbare Daten vorantreiben. Datenatlas und Datenpool werden zur Grundlage für datengetriebene Prozesse und Entscheidungen in den Bundesbehörden, sie werden ein effektiveres Wissensmanagement ermöglichen und die Aufbereitung und Bereitstellung von Daten als Open Data unterstützen. Auf dieser Grundlage können Verwaltungen die Daten auch ressortübergreifend für ein effektives und zukunftsfähiges Verwaltungshandeln teilen.[21]

[21] Bundesministerium für Digitales und Verkehr, Bundesministerium für Wirtschaft und Klimaschutz, und Bundesministerium des Innern und für Heimat, Hrsg. *Fortschritt durch Datennutzung - Strategie für mehr und bessere Daten für neue, effektive und zukunftsweisende Datennutzung*. Die Bundesregierung, 2023. https://tinyurl.com/datenstrategiede

DIE VERFEHLUNG dieser weiteren Ziele durch den DATENATLAS wurde im Rahmen dieses Gutachtens begründet und um Desiderata ergänzt, welche die Entwicklung einer nutzbaren Lösung ermöglichen würden.

Ob dies auf Grundlage der technischen Basis des DATENATLAS sinnvoll ist, wird in Abschnitt 6.3 separat diskutiert.

Insbesondere die eigentlich durch die damalige Bundesregierung gewünschte Nutzung *maschinenlesbarer Daten* und die Unterstützung der Bereitstellung von *Open Data* – also der *Stand der Technik* – wurde nicht umgesetzt, wie in Abschnitt 5.4 präsentiert wurde.

Letztendlich führt dieser Mangel dazu, dass selbst bei der Nutzung des DATENATLAS kaum Kosten für die BUNDESVERWALTUNG eingespart würden, da der Verzicht auf *Linked (Open) Data-Technologien* oder *kontrollierte Vokabulare* zur Folge hat, dass Daten nicht automatisiert kontextualisiert werden können. Das wiederum macht manuelle Nacharbeiten nötig.

Letztendlich widerspricht die BUNDESDRUCKEREI damit ihrer eigenen Zielstellung „Verbindungen zwischen Elementen“[22] aufzeigen zu können: Denn genau das ist aufgrund des zugrundeliegenden Datenmodells kaum umsetzbar.

Auch die prinzipiell wertvolle Funktion, wonach „KI-Methoden [dabei] unterstützen [...], Informationen zu extrahieren“[23], wurde nicht implementiert. Dass die Bereitstellung einer solchen Funktion

[22] Bundesdruckerei. *Erstes Vollständiges Datenmodell Der Bundesverwaltung - Pressemitteilung*, 2022. https://tinyurl.com/bdr-pm3. Letzter Abruf: 21.07.2025

[23] Ebenda.

aufgrund verschiedenster Defizite unwahrscheinlich ist, wurde bereits in Abschnitt 5.6 diskutiert.

Folglich ist eine Indikatoren-gesteuerte Steuerung der BUNDESVERWALTUNG oder die teilautomatisierte Umsetzung von Berichtspflichten auf Grundlage der vorliegenden technischen Basis ebenso nicht gegeben[24].

[24] Siehe Abschnitt 3.6.

WELCHE ZIELE ERFOLGREICH umgesetzt wurden, postuliert eine Website der Bundesdruckerei (2025) in Form von fünf beantworteten Fragen zum DATENATLAS. Diese werden im Folgenden den Erkenntnissen dieses Gutachtens gegenübergestellt, um eine abschließende Bewertung treffen zu können:

> *Welche konkreten Vorteile bietet der Datenatlas Bund Verwaltungsmitarbeitenden?*
>
> Der Datenatlas beinhaltet eine zentrale, verständlich visualisierte Metadatenbank, die einen vereinheitlichten, transparenten und leicht zugänglichen Überblick über die vorhandenen Datenbestände in der Verwaltung ermöglicht. Mitarbeitende der Bundesverwaltung können so relevante Datenbestände leichter identifizieren.
>
> Zudem erhalten sie Zugriff auf detaillierte Informationen über die Datensätze sowie Kontaktdaten von Ansprechpersonen. So können sie die zuständigen Stellen schneller adressieren, wenn sie die Datenbestände nutzen möchten. Das steigert die Effizienz in der Datennutzung.

BEWERTUNG: Nach aktuellem Stand verfügt der DATENATLAS über keine über die tabellarische Darstellung herausgehende Visualisierung von Daten.

Durch die Mängel im Rahmen der Recherchemöglichkeiten[25] ist anhand wissenschaftlicher Erkenntnisse nicht damit zu rechnen, dass die Mehrzahl der Mitarbeitenden der BUNDESVERWALTUNG – also Laien-Nutzende – gewünschte Informationen *effizient* finden werden.

[25] Siehe Abschnitt 5.2f.

Die Entscheidung, *keine* explorativen Suchstrategien für einen Großteil der Nutzenden zu implementieren, steht im direkten Widerspruch zum eigenen Ziel, eine „ressortübergreifende Suche und Exploration der Datenbestände“[26] bereitzustellen.

Hinzu kommt, dass der DATENATLAS keinerlei *Relevanzbewertung* von Datensätzen[27] unterstützt. Daher kann nicht von einer Effizienzsteigerung im Bereich der Datennutzung ausgegangen werden.

[26] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. `https://tinyurl.com/bdr-pm1`. Letzter Abruf: 21.07.2025

[27] Siehe Abschnitt 5.2,

> *Welche Daten werden im Datenatlas Bund erfasst und welche Datenbestände katalogisiert?*
>
> Der Datenatlas enthält ausschließlich Metadaten, welche die Datenbestände fachlich, technisch und organisatorisch beschreiben. Wichtig: Der Zugriff auf die eigentlichen Nutzdaten erfolgt nicht über den Datenatlas selbst, sondern nur in Abstimmung mit den verantwortlichen Stellen. Welche Datenbestände im Datenatlas katalogisiert werden müssen, ist nicht festgelegt.

> Relevant sind grundsätzlich Datenbestände der Bundesverwaltung, die entweder aufgrund einer Rechtsgrundlage verarbeitet werden oder für politische Planungs- und Entscheidungsprozesse sowie die Regierungsarbeit wichtig sind. Zusätzlich können Datenbestände aufgenommen werden, die für die Aufgabenerfüllung anderer Organisationseinheiten von Interesse sind.

BEWERTUNG: Durch die Vagheit der Aufnahmekriterien wird die effiziente Datennutzung für die BUNDESVERWALTUNG erschwert, da für Nutzende nicht im Vorfeld erkennbar ist, ob der DATENATLAS für sie relevante Datensätze enthält.

Hier müsste zuerst eine Regelung erlassen werden, damit der DATENATLAS einen Großteil der Daten der BUNDESVERWALTUNG bereitstellen kann.

Es ist deshalb damit zu rechnen, dass die Nutzungserwartung an den DATENATLAS enttäuscht wird und sich die Lösung nicht durch interne Empfehlung innerhalb der BUNDESVERWALTUNG weiterverbreiten wird.

> *Wie wird mit den bereits bestehenden kleinen Datenkatalogen der Ressorts verfahren?*
>
> Grundsätzlich soll es im Rahmen des Projekts keine doppelte Erhebung von Metadaten geben. Allerdings kann es sinnvoll sein, bereits bestehende Informationen zu nutzen.
>
> Entscheidend ist dabei, ob diese aktuell genug sind, ob das Format passt und ob der erwartete Mehrwert den Aufwand rechtfertigt, der mit der Anbindung an die Schnittstelle des Datenatlas Bund einhergeht.

BEWERTUNG: Im Rahmen des Gutachtens war es nicht möglich zu bestimmen, in welchem Umfang und in welcher Qualität Dateninventuren in der BUNDESVERWALTUNG durchgeführt wurden.

> *Wie wird der Datenatlas in der Bundesverwaltung eingeführt?*
>
> Der Datenatlas wurde ressortweise eingeführt, beginnend mit dem Bundesministerium der Finanzen (BMF) und Bundesministerium des Innern und für Heimat (BMI). Jedes Ressort ist technisch mit einem Zugang zum Datenatlas ausgestattet. Das Vorgehen basiert auf den Erfahrungen aus dem Pilotprojekt im Bundesministerium der Finanzen (BMF).
>
> Innerhalb jedes Ressorts gliedert sich die Implementierung in drei Schritte: die Initialisierung, die Datenerhebung und den Einsatz der Anwendung. Diese Phasen können teilweise parallel ablaufen. Langfristig soll in der Bundesverwaltung ein Bewusstsein für relevante Datenbestände und deren Katalogisierung im Datenatlas geschaffen werden, damit neue Datenbestände selbstständig erfasst werden können.

BEWERTUNG: Aufgrund der aufgezeigten Mängel des DATENATLAS ist fraglich, ob das Ziel der Nutzendenakzeptanz zum jetzigen Zeitpunkt erreicht werden kann.

Nach dem *Stand der Technik* wäre zu vermuten, dass die ressortweise Einführung von Evaluierungen begleitet wird und die

Datenlabore als zuständige Kompetenzzentren dabei aktiv durch den Dienstleister eingebunden worden sind.

Aufgrund der Defizite im Bereich der *Usability*[28] ist davon auszugehen, dass ein solches Vorgehen aktuell – wenn überhaupt – nur sporadisch verfolgt wird.

[28] Siehe Abschnitt 5.3.

Da im Rahmen der Datenerhebung anzunehmen ist, dass neue Anforderungen an den Datenatlas entdeckt werden, ist es unverständlich, dass keine Anpassungen der Software vorgesehen sind – was dem angeblich genutzten agilen Ansatz[29] widerspricht.

[29] Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025

Es bleibt offen, was mit Datenbeständen geschieht, welche sich nicht im Datenmodell des Datenatlas abbilden lassen.

Das heißt im Extremfall, dass in der Bundesverwaltung existierende, qualitativ hochwertige Datensätze durch die Aufnahme in den Datenatlas qualitativ gemindert werden, wie in Abschnitt 5.4 am Beispiel des GovData-Imports aufgezeigt wurde.

In der Folge ist damit zu rechnen, dass dadurch primär auf interne Recherchewerkzeuge oder Domänenwissen gesetzt wird.

> *Wie können Interessierte in der Verwaltung den Datenatlas Bund jetzt nutzen? Wo kann man auf den Datenatlas zugreifen?*
>
> Der Datenatlas Bund ist für freigeschaltete Ministerien und Behörden über das Internet und das Netz des Bundes erreichbar. Mitarbeitende von Bundesbehörden können sich selbst im Datenatlas registrieren, um ihn anschließend als Web-Anwendung zu nutzen.

Bewertung: Es ist für die direkte Bundesverwaltung möglich auf den Datenatlas zuzugreifen.

Im Rahmen eines Austauschs mit einem Mitglied des Bundestags wurde die Vermutung geäußert, dass das Parlament zum aktuellen Zeitpunkt keinen direkten Zugriff auf die Anwendung hat.

Ob dieser über die Bundesdruckerei auf Wunsch ermöglicht wird, wurde nicht ermittelt.

Zusammenfassung

Dass weder die selbst gesteckten Ziele für den Datenatlas erreicht noch die acht Kriterien des *Servicestandards* umgesetzt werden, wird aus oben genannten Bewertungsansätzen deutlich.

Dabei ist es besonders überraschend, dass die Bundesdruckerei selbst die *Gerichtete Suche* und das *Browsing* als wichtige Kernfunktionen benennt, jedoch nicht umsetzt, wie in Abschnitt 5.2 gezeigt werden konnte.

Der Verzicht auf die Nachnutzung funktional reiferer Software-Systeme durch den Dienstleister führt letztendlich nicht nur zu einer Minderung des Funktionsangebots, die im vorigen Abschnitt 6.1 beschrieben wurde, sondern auch zu deutlich höheren Kosten, wie in Abschnitt 5.8 dargestellt wurde.

Eine Nachnutzung der vorgestellten Systeme hätte von vornherein u.a. wichtige Funktionen wie die *Gerichtete Suche* und das *Browsing* beinhaltet, da diese seit langem *Stand der Technik* sind.

Die Bindung an einen Dienstleister steht der Erreichung der *Digitalen Souveränität* diametral gegenüber.

Durch den entstehenden *vendor lock-in* und die Bereitstellung einer *Closed-Source-Software* begibt sich die Bundesverwaltung in ein langfristiges Abhängigkeitsverhältnis zur Bundesdruckerei.

Wenn die Nutzung des Datenatlas trotz der aufgezeigten Mängel verpflichtend würde, führte dies unmittelbar zu der Bildung eines weiteren, wenig interoperablen Datensilos.

Momentan existieren weder eine effektive Datenqualitätssicherung noch Export-Möglichkeiten im Datenatlas, so dass im Falle eines Wechselwunsches zu einer anderen Plattform, erhebliche Aufwände für die Datenmigration anfallen würden.

Hier zeigen sich die Vorteile einer offenen Plattform wie GovData, welches aktuell von CKAN nach Piveau wechselt[30] und dabei sowohl auf Datenstandards als auch Standardschnittstellen setzen kann.

[30] Siehe Abschnitt 5.2

Extrinsisch motvierter Weiterentwicklungsdruck, wie er durch §6 OZG (2024) zu erwarten ist, wird durch die Bindung an den Dienstleister potenziell weniger wirtschaftlich umsetzbar sein als bei einem offenen System mit einem Konsortialmodell (siehe Abschnitt 6.3).

Inwiefern ein Versagen im Bereich der Projektsteuerung vorliegt, lässt sich aus der Außenperspektive schwer einschätzen.

Zwar sind dem Autor des Gutachtens Meinungsäußerungen von Mitarbeitenden der Datenlabore am Rande von Tagungen bekannt, die Teilaspekten des Gutachtens entsprechen, jedoch ist nicht bekannt, inwiefern diese Bedenken in die Breite getragen wurden oder durch die Projektsteuerung gehört wurden.

Selbst im Falle von Problemen der Projektsteuerung wäre der Dienstleister in der Pflicht gewesen, auf den *Stand der Technik* hinzuweisen.

Anerkennend zu erwähnen ist der hohe angefallene Aufwand der Bundesdruckerei für die Abstimmung mit den Datenlaboren und den Bundeseinrichtungen, die im Rahmen der Dateninventur besucht wurden.

Funktional gesehen rechtfertigt dieser Aufwand und die hohe Komplexität in einem Fachlos jedoch nicht das Gesamtergebnis.

Als konstruktives Ergebnis des Projekts bleibt eigentlich nur eine Dateninventur, vorausgesetzt diese ist repräsentativ und umfassend durchgeführt worden.

6.3 *Handlungsempfehlungen*

Die folgende Handlungsempfehlung basiert auf der in Abschnitt 5.8 dargelegten Einschätzung, dass die Umsetzung des DATENATLAS im Einklang mit §97(4) GWB (2024) in mindestens zwei unabhängigen Fachlosen erfolgt sein müsste, da die Dateninventur und die Software-Entwicklung voneinander klar abgrenzbar sind.

AUFGRUND DER EKLATANTEN MÄNGEL ist das Software-Entwicklungsprojekt DATENATLAS mit *sofortiger Wirkung* zu stoppen, um nicht weitere Mittel in eine technisch und konzeptionell wenig überzeugende Lösung zu investieren, welche kaum den *Stand der Technik* erreicht.

Sollte wider Erwarten für das Los „Software-Entwicklung" ein Werkvertrag mit der BUNDESDRUCKEREI ausgehandelt worden sein, so darf die Abnahme aufgrund der präsentierten Sachmängel nicht erfolgen – auch wenn die konkrete Leistungsbeschreibung dem Autor unbekannt ist und nur aus den Pressemitteilungen der BUNDESDRUCKEREI abgeleitet werden konnte.

Im Falle eines Dienstvertrags ist dieser sofort zu beenden.

NEBEN DEM UMSTAND, dass eine proprietäre Lösung durch den entstehenden *vendor lock-in* nachhaltig die *Digitale Souveränität* der BUNDESVERWALTUNG schwächt[31], bieten typische *Open-Source-Lösungen* wesentlich mehr Funktionen[32] und wären damit allein aus Gründen der Wirtschaftlichkeit zu bevorzugen gewesen.

[31] Siehe Abschnitt 5.5.

[32] Siehe Abschnitt 5.2.

WEGEN DER SCHWERWIEGENDEN Defizite des DATENATLAS geht der Autor des Gutachtens davon aus, dass Mitarbeitende der Datenlabore ähnliche Argumente *zumindest* Projekt-intern im Rahmen der Entwicklung vorgetragen haben.

Bei Kompetenzzentren für die Daten-getriebene Verwaltungsmodernisierung ist davon auszugehen, dass diese primär „Information Professionals" beschäftigen, die im Rahmen der Bestenauslese besetzt wurden und nicht allein über eine juristische Qualifikation verfügen.

Der Autor geht davon aus, dass im Projektverlauf seitens der Datenlabore oder einzelner Mitarbeitender ein Projektabbruch gefordert worden sein dürfte.

Es ist ferner davon auszugehen, dass seitens der Datenlabore u.a. auf die in Tabelle 5.2 vorgestellten Systeme oder vergleichbare Systeme unter *Open-Source-Lizenz* hingewiesen wurde – zumal im näheren Umfeld, d.h. bei GOVDATA, OPEN.BYDATA oder DATENADLER ähnliche Entwicklungen, z.T. lange vor Beginn des Projektes, zu beobachten waren.

Auch der mindestens zu erwartende Funktionsumfang eines Recherchewerkzeugs[33] ist in diesem Kreis, wie auch bei allen Hochschulabsolventinnen und Hochschulabsolventen, als bekannt vorauszusetzen.

[33] Siehe Abschnitt 3.2.

Das BMF muss prüfen, ob solche Initiativen existierten und warum diese Initiativen nicht gehört wurden.

AUS FURCHT vor Konsequenzen haben alle Mitarbeitenden der verschiedenen Datenlabore darauf bestanden, anonym zu bleiben.

Sollten einzelne oder mehrere Individuen dieses Personenkreises aufgrund ihrer fachlichen Expertise ähnliche Bedenken, wie die in diesem Gutachten aufgezeigten, artikuliert haben, so müssten diese entsprechend belobigt werden, da sie einen wirtschaftlichen Schaden von der BUNDESREPUBLIK hätten abwenden können.

EINE WEITERENTWICKLUNG oder ein Refactoring des DATENATLAS würde diverse Kernfunktionen der Software betreffen, wie bereits in Abschnitt 5.2 gezeigt werden konnte, und lässt sich – auch aufgrund der bereits existierenden und funktional besseren *Open-Source-Lösungen* – kaum wirtschaftlich darstellen.

Dies gilt auch für den Weiterbetrieb durch das BMDS oder das ITZBUND, wobei hier auch noch die IT-Sicherheit betreffende Szenarien zu prüfen wären (u.a. den Datenimport aus GOVDATA).

EINE INTEGRATION der vorliegenden Software in den *Deutschland-Stack* verbietet sich sowohl aus Überlegungen, welche die *Digitale Souveränität* (siehe unten) betreffen, als auch aus den im Gutachten erläuterten Mängeln des DATENATLAS.

Diese Aussage gilt ebenso für die Nachnutzung der Software durch weitere Behörden.

Sollte der Einsatz des DATENATLAS in anderen Einrichtungen aktuell noch diskutiert werden, sollte diesen Behörden das vorliegende Gutachten zur Kenntnisnahme gegeben werden.

EINE WIRTSCHAFTLICHKEITSKONTROLLE des DATENATLAS im Rahmen der Erfolgskontrolle nach VV-BHO (2025) muss durch das BMF umgehend erfolgen.

Zwar zeigt sich anhand der Kostenschätzung, die in Abschnitt 5.8 vorgenommen wurde, dass der DATENATLAS die höchsten Entwicklungskosten pro Jahr aufweisen dürfte, jedoch existieren keine öffentlich zugänglichen Angaben zu den realen Kosten.

Liegen diese Kosten *wesentlich* höher als die Schätzung, muss kritisch geprüft werden, ob im Vorfeld eine umfassende Wirtschaftlichkeitsuntersuchung nach §7(2.1) VV-BHO (2025)[34] erfolgt ist und auf Grundlage welcher Vertragsform – also als Werk- oder Dienstvertrag – eine Vergabe durchgeführt wurde.

[34] Siehe Abschnitt 3.7.

Dass sowohl die *Zielerreichungskontrolle* als auch die *Wirkungskontrolle* ein negatives Ergebnis haben werden, wurde im gesamten Gutachten begründet.

LETZTENDLICH ist eine Neuentwicklung des DATENATLAS unumgänglich, da aufgrund seiner leicht zu entdeckenden Defizite erhebliche Zweifel an der Eignung des Dienstleisters für die Entwicklung

von Software-Lösungen der *Informationsrecherche* bestehen.

Das vorliegende Gutachten hat in Kapitel 5 diverse Verbesserungsmöglichkeiten und Umsetzungsmöglichkeiten vorgestellt, die Dienstleister für ihre Arbeit am DATENATLAS heranziehen könnten. Dass der aktuelle Dienstleister in der Lage ist, damit ein Recherche-Werkzeug auf dem *Stand der Technik* zu implementieren, scheint ausgeschlossen.

Aufgrund der beschriebenen Probleme wird deutlich, dass der aktuelle Dienstleister über andere als die für die konkrete Umsetzung des DATENATLASY nötigen Kompetenzen auf den Gebieten *Information Retrieval* und *Linked (Open) Data* verfügt, welche in diesem Gutachten aus Platzgründen stark verkürzt dargestellt wurden.

Es bleibt zu vermuten, dass der Dienstleister zuerst in einen internen Kompetenzaufbau investieren müsste, um das Projekt erfolgreich umsetzen zu können.

Diese Aussage wird dadurch begründet, dass sich die angeführte Literatur mit ihren Aussagen zu großen Teilen auch in gängigen Lehrbüchern findet – also den *Stand der Technik* formt.

Hinzu kommen die einschlägigen Normen, welche eigentlich bekannt hätten sein müssen.

ANSTELLE EINER EIGENENTWICKLUNG sollte bei der Neuentwicklung – schon allein aus Gründen der Wirtschaftlichkeit – auf eine bestehende *Open-Source-Lösung* wie FEDORA oder PIVEAU und den Aufbau einer Entwicklungsgemeinschaft gesetzt werden.

Für die Verwendung von FEDORA spricht die Möglichkeit, einem bereits existierenden Konsortium beitreten zu können, welches die offene Weiterentwicklung garantiert, jedoch eher einen Fokus auf Forschungseinrichtungen und Bibliotheken hat.

Auch wenn bereits verschiedene Bundesländer (Bayern, Brandenburg) auf PIVEAU als Basis setzen und GOVDATA mittelfristig dorthin migrieren wird, besteht bei dieser Lösung ein höheres Risiko eines *vendor lock-ins*[35] als bei FEDORA.

[35] Siehe Abschnitt 5.5.

Dieses Risiko ließe sich durch die Etablierung eines Konsortialmodells – ähnlich dem Vorbild von FEDORA – minimieren, um eine langfristige Weiterentwicklung und die *Interoperabilität* der deutschen Datenplattformen sicherzustellen.

Ähnliche Modelle existieren bereits als Kooperation zwischen dem Bund und den Ländern.

Beispielhaft genannt werden kann hier ePayBL[36], welches dem BMF als federführendes Ministerium bei der Entwicklung des DATENATLAS bekannt gewesen sein muss, da es die Fachaufsicht über die GENERALZOLLDIREKTION hat, welche den Bund in dieser Entwicklungsgemeinschaft vertritt.

[36] `https://www.epaybl.de/`; Letzter Abruf: 01.08.2025

Eine solche Lösung, welche den in Abschnitt 5.5 skizzierten Anforderungen an die *Digitale Souveränität* genügt, hätte das Potential in den *Deutschland-Stack* aufgenommen zu werden.

In Erwartung der Standardisierung durch den Gesetzgeber nach §6 OZG (2024)[37], ist ebenfalls damit zu rechnen, dass sich die noch

[37] Siehe Abschnitt 3.7.

zu benennenden Architekturvorgaben und Qualitätsanforderungen nur in einem Konsortialmodell *wirtschaftlich* umsetzen lassen.

ALS ORIENTIERUNGHILFE während der Neuentwicklung sollten die FAIR-Prinzipien[38] dienen, welche aktuell nicht durch den DATENATLAS umgesetzt werden.

Die FAIR-Prinzipien stellen Anforderungen an die Bereitstellung und Speicherung von Daten dar und lauten wie folgt:

- *Findable* -– Auffindbarkeit,
- *Accessible* -– Zugänglichkeit,
- *Interoperable* -– Interoperabilität,
- *Reusable* -– Wiederverwendbarkeit.

Diese entsprechen weitgehend den Zielen der *Bundesdatenstrategie*, die mit dem DATENATLAS hätten umgesetzt werden sollen.

Die vier das Akronym bildende Punkte werden jeweils weiter auformuliert und umfassen z.B. Aspekte wie die Maschinenlesbarkeit von Daten.

Aus Platzgründen soll hier jedoch auf die offizielle Website der FAIR-Initiative[39] verwiesen werden, welche weitere Realisierungshinweise gibt.

[38] Mark D. Wilkinson, Michel Dumontier, und Aalbersberg et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016. ISSN 2052-4463. DOI: 10.1038/sdata.2016.18. https://www.nature.com/articles/sdata201618. Letzter Abruf: 01.08.2025

[39] https://www.go-fair.org/fair-principles/; Letzter Abruf: 01.08.2025

DIE DATENINVENTUR als zweites Fachlos sollte weiterverfolgt werden, wenn geeignete Evaluierungsmaßnahmen für den Umfang und die Qualität dieser implementiert sind.

Eine umfassende Dateninventur in der BUNDESVERWALTUNG stellt die Basis für weitere Recherche- und Datenplattformen dar, welche – wie oben detailliert dargestellt – auf *Open-Source-Software* basieren sollten.

Die Aufnahmekriterien für Datensätze in den DATENATLAS sind bisher sehr vage. Hier sollte der Gesetzgeber klare Vorgaben machen, um eine möglichst weite Abdeckung zu erreichen.

Transparenzhinweis

Aus Transparenzgründen sei darauf hingewiesen, dass der Autor des Gutachtens im Jahr 2023 an einem nicht bewilligten Drittmittelantrag beteiligt war, der auf einer engen Kooperation mit der FRAUNHOFER-GESELLSCHAFT und die Nutzung von PIVEAU aufgebaut hätte.

Außerdem betreute er als Zweitgutachter die Masterarbeit des ehemaligen Product Owners von OPEN.BYDATA, bevor dieser seine Tätigkeit dort aufnahm.

Zuvor war der Autor ca. sieben Jahre intensiv an der STAATSBIBLIOTHEK ZU BERLIN mit der Konzeption, Einführung und Weiterentwicklung von u.a. auf FEDORA basierenden *Repository-Systemen* befasst.

Im Rahmen dieser Tätigkeit veranlasste er den kostenpflichtigen Beitritt der Staatsbibliothek zu Berlin in die Fedora-Community[40].

[40] https://fedorarepository.org/meet-our-members/; Letzter Abruf: 01.08.2025

Literaturverzeichnis

Bill Albert und Tom Tullis. *Measuring the User Experience: Collecting, Analyzing, and Presenting UX Metrics*. Morgan Kaufmann, an imprint of Elsevier, Cambridge,MA, 3. Auflage, 2023. ISBN 978-0-12-818080-8.

Ethem Alpaydın. *Maschinelles Lernen*. Oldenbourg, 2008. ISBN 978-3-486-58114-0.

AVV DatA. *Allgemeine Verwaltungsvorschrift über den Austausch von Daten im Bereich der Lebensmittelsicherheit und des Verbraucherschutzes (AVV Datenaustausch)*, 2010. https://www.verwaltungsvorschriften-im-internet.de/bsvwvbund_15122010_321221010032.htm. Letzter Abruf: 25.07.2025.

Ricardo Baeza-Yates und Berthier Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search*. Pearson Addison-Wesley [u.a.], Harlow, 2. Auflage, 2011.

J. Marcia Bates. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, 13(5):407–424, 1989.

J. N. Belkin, G. P. Marchetti, und C. Cool. BRAQUE: Design of an Interface to Support User Interaction in Information Retrieval. *Inf. Process. Manage.*, 29(3):325–344, 1993. http://dx.doi.org/10.1016/0306-4573(93)90059-M.

Nicholas Belkin. Interaction with Texts: Information Retrieval as Information-Seeking Behavior. In *Information Retrieval*, Seiten 55–66, 1993.

Nicholas Belkin. Intelligent Information Retrieval: Whose Intelligence? In *ISI '96: Proceedings of the Fifth International Symposium for Information Science*, Seiten 25–31, 1996.

Tim Berners-Lee. *RFC 1630: Universal Resource Identifiers in WWW*, 1994.

Tim Berners-Lee, James Hendler, und Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.

BHO. *Bundeshaushaltsordnung*, 2024. https://www.gesetze-im-internet.de/bho/. Letzter Abruf: 26.07.2025.

Dania Bilal. Children's Use of the Yahooligans! Web search Engine: Cognitive, Physical, and Affective Behaviors on Fact-based Search Tasks. *J. Am. Soc. Inf. Sci.*, 51(7):646–665, 2000. http://dx.doi.org/10.1002/(SICI)1097-4571(2000)51:7<646::AID-ASI7>3.0.CO;2-A.

BITV 2.0. *Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung)*, 2023.

Lina Bruns, Benjamin Dittwald, und Fritz Meiners. *Leitfaden für Qualitativ Hochwertige Daten und Metadaten*, 2019. https://www.fokus.fraunhofer.de/content/dam/fokus/dokumente/dps/flyer/NQDM_Leitfaden_2019.pdf. Letzter Abruf: 01.08.2025.

Bundesdruckerei. *Erstes Vollständiges Datenmodell Der Bundesverwaltung - Pressemitteilung*, 2022. https://tinyurl.com/bdr-pm3. Letzter Abruf: 21.07.2025.

Bundesdruckerei. *Datenatlas Bund - Der Souveräne Datenkatalog für die Bundesverwaltung*, 2025. https://tinyurl.com/bdr-pm1. Letzter Abruf: 21.07.2025.

Bundesministerium für Digitales und Verkehr, Bundesministerium für Wirtschaft und Klimaschutz, und Bundesministerium des Innern und für Heimat, Hrsg. *Fortschritt durch Datennutzung - Strategie für mehr und bessere Daten für neue, effektive und zukunftsweisende Datennutzung*. Die Bundesregierung, 2023. https://tinyurl.com/datenstrategiede.

Bundesrat. *Allgemeine Verwaltungsvorschrift über die Übermittlung von Daten aus der amtlichen Lebensmittel- und Veterinärüberwachung sowie dem Lebensmittel-Monitoring (AVV Datenübermittlung - AVV-DÜb) - 834/98*, 1998. https://dserver.bundestag.de/brd/1998/D834+98.pdf. Letzter Abruf: 01.08.2025. Drucksache 834/98.

Bundesverwaltungsamt. *S-O-S-Methode Für Großprojekte*, 2021. https://www.bva.bund.de/SharedDocs/Downloads/DE/Behoerden/Beratung/GrossPM/Handbuch/S-O-S_Handbuch.pdf?__blob=publicationFile&v=2. Letzter Abruf: 18.08.2025.

F. E. Codd. A Relational Model of Data for Large Shared Data Banks. *Commun. ACM*, 13(6):377–387, 1970. http://doi.acm.org/10.1145/362384.362685.

Alain Colmerauer und Philippe Roussel. The Birth of Prolog. *SIGPLAN Not.*, 28(3):37–52, March 1993. ISSN 0362-1340. DOI: 10.1145/155360.155362. https://doi.org/10.1145/155360.155362.

Alan Cooper. *The Inmates Are Running the Asylum*. Macmillan Publishing Co., Inc., Indianapolis, IN, USA, 1999.

Alan Cooper, Robert Reimann, und Dave Cronin. *About Face 3: The Essentials of Interaction Design*. Wiley, Indianapolis, Ind., 2007.

Bruce W. Croft und H. R. Thompson. I3R: A New Approach to the Design of Document Retrieval Systems. *J. Am. Soc. Inf. Sci.*, 38(6): 389–404, 1987. http://dx.doi.org/10.1002/(SICI)1097-4571(198711)38:6<389::AID-ASI1>3.0.CO;2-4.

Bruce W. Croft, Donald Metzler, und Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Pearson, Boston, Mass., international edition Auflage, 2009.

J. C. Date. A Formal Definition of the Relational Model. *SIGMOD Rec.*, 13(1):18–29, 1982. http://doi.acm.org/10.1145/984514.984515.

Deutscher Bundestag - Wissenschaftliche Dienste. *Grundsätze der Aktenführung in der Bundesverwaltung*, 2023. https://tinyurl.com/veraktung. Letzter Abruf: 26.07.2025.

DIN EN ISO 9241-210. *Ergonomie der Mensch-System-Interaktion - Teil 210: Menschzentrierte Gestaltung Interaktiver Systeme (ISO 9241-210:2019)*, 2020.

DIN EN ISO 9241-220. *Ergonomie der Mensch-System-Interaktion - Teil 220: Prozesse zur Ermöglichung, Durchführung und Bewertung Menschzentrierter Gestaltung für Interaktive Systeme in Hersteller- und Betreiberorganisationen (ISO 9241-220:2019)*, 2020.

DIN SPEC 66336. *Qualitätsanforderungen für Onlineservices und -Portale der Öffentlichen Verwaltung (Servicestandard)*, 2025. https://servicestandard.gov.de/din-spec-66336/. Letzter Abruf: 28.07.2025.

Sándor Dominich. *The Modern Algebra of Information Retrieval*. Springer-11645 / Dig. Serial]. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2008.

EGovG. *Gesetz zur Förderung der elektronischen Verwaltung (E-Government-Gesetz)*, 2024. https://www.gesetze-im-internet.de/egovg/. Letzter Abruf: 25.07.2025.

D. Ellis. A Behavioural Model for Information Retrieval System Design. *J. Inf. Sci.*, 15(4-5):237–248, 1989. http://dx.doi.org/10.1177/016555158901500406.

David Ellis und Merete Haugan. Modelling the Information Seeking Patterns of Engineers and Research Scientists in an Industrial Environment. *Journal of Documentation*, 53(4):384–403, 1997. http://www.ingentaconnect.com/content/mcb/278/1997/00000053/00000004/art00003.

R. Fagin und E. L. Wimmers. A Formula for Incorporating Weights into Scoring Rules. *Special Issue of Theoretical Computer Science*, 230 (2):309–338, January 2000. DOI: 10.1016/S0304-3975(99)00224-8.

M. Flickner, S. H. Sawhney, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, und P. Yanker. Query by Image and Video Content: The QBIC System. *IEEE Computer*, 28(9):23–32, 1995.

A. Edward Fox, S. Betrabet, M. Koushik, und W. Lee. Extended Boolean Models. In B. W. Frakes und R. Baeza-Yates, Hrsg., *Information Retrieval: Data Structures and Algorithms*, Seiten 393–418. Prentice Hall, 1992. ISBN 0-13-463837-9.

Brian R. Gaines. The Technology of Interaction—Dialogue Programming Rules. *International Journal of Man-Machine Studies*, 14(1):133–150, 1981. ISSN 0020-7373. DOI: 10.1016/S0020-7373(81)80037-5.

Thomas Geis und Knut Polkehn. *Praxiswissen User Requirements: Nutzungsqualität systematisch, nachhaltig und agil in die Produktentwicklung integrieren : Aus- und Weiterbildung zum UXQB® Certified Professional for Usability and User Experience - Advanced Level "User Requirements Engineering"*. dpunkt.verlag, 1. Auflage, 2018. ISBN 978-3-86490-527-8.

Thomas Geis und Guido Tesch. *Basiswissen Usability und User Experience: Aus- und Weiterbildung zum UXQB® Certified Professional for Usability and User Experience (CPUX) - Foundation Level (CPUX-F)*. dpunkt.verlag, Heidelberg, 1. Auflage, 2019. ISBN 978-3-86490-599-5.

GWB. *Gesetz gegen Wettbewerbsbeschränkungen*, 2024. https://www.gesetze-im-internet.de/gwb/BJNR252110998.html.

Aurélien Géron. *Praxiseinstieg Machine Learning mit Scikit-Learn und TensorFlow: Konzepte, Tools und Techniken für intelligente Systeme*. Animals. O'Reilly, 2018. ISBN 978-3-96009-061-8 978-3-96010-114-7.

C. Hanson. *Opinion – Libraries are Software*, 2015. https://www.codyh.com/writing/software.html. Letzter Abruf: 27.05.2027.

A. Marti Hearst. *Search User Interfaces*. Cambridge Univ. Press, Cambridge, 2009.

Andreas Henrich. *Information Retrieval 1 - Grundlagen, Modelle und Anwendungen*. Otto-Friedrich-Universität Bamberg - Lehrstuhl für Medieninformatik, 1.2 Auflage, 2008.

Peter Ingwersen. *Information Retrieval Interaction*. Taylor Graham, London, 1992. http://www.gbv.de/dms/hbz/toc/ht004327073.PDF.

Peter Ingwersen. Cognitive Perspectives of Information Retrieval Interaction: Elements of a Cognitive IR Theory. *Journal of Documentation*, 52(1):3–50, 1996. ISSN 0022-0418. DOI: 10.1108/eb026960.

ISO 14721. *Space Data System Practices - Reference Model for an Open Archival Information System (OAIS)*, 2025.

ISO 25964-1. *Information and Documentation - Thesauri and Interoperability with Other Vocabularies - Part 1: Thesauri for Information Retrieval*, 2011.

IT-Planungsrat. *Ergebnisprotokoll der 26. Sitzung des IT-Planungsrats (28.06.2018)*, 2018. https://t1p.de/95bke. Letzter Abruf: 20.07.2025.

Diane Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends Inf. Retr.*, 3:1–224, 2009. http://portal.acm.org/citation.cfm?id=1618301.1618302.

Kompetenzzentrum Open Data. *Leitfaden Metadaten - Version 2.0*, Bundesverwaltungsamt, 2023. https://tinyurl.com/leitfaden-md. Letzter Abruf: 24.07.25.

Thomas Kudraß und Thomas Brinkhoff, Hrsg. *Taschenbuch Datenbanken*. Fachbuchverl. Leipzig im Carl-Hanser-Verlag, 2007. ISBN 978-3-446-40944-6.

C. C. Kuhlthau. Inside the Search Process: Information Seeking from the User's Perspective. *Journal of the American Society for Information Science*, 42(5):361–371, 1991.

Daniel Lambach und Kai Oppermann. Narratives of Digital Sovereignty in German Political Discourse. *Governance*, 36(3):693–709, 2023. ISSN 0952-1895, 1468-0491. DOI: 10.1111/gove.12690.

Vivian Larrea, Milene Selbach Silveira, und Tiago Da Silva. The use of prototypes as a tool in Agile software development. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, Seiten 842–849. ACM, 2024. ISBN 979-8-4007-0243-3. DOI: 10.1145/3605098.3636064. https://dl.acm.org/doi/10.1145/3605098.3636064.

Ho Joon Lee. Properties of Extended Boolean Models in Information Retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Seiten 182–190. Springer-Verlag New York, Inc., 1994. ISBN 0-387-19889-.

Lexikon-Redaktion des Gabler Verlages, Hrsg. *Gabler Kompakt-Lexikon Wirtschaft*. Gabler Kompakt-Lexikon Wirtschaft. Springer Gabler, 11. Auflage, 2013. ISBN 978-3-658-00008-0.

Gary Marchionini, Gary Geisler, und Ben Brunk. Agileviews: A Human-Centered Framework for Interfaces to Information Spaces. In *Proceedings of the Annual Conference of the American Society for Information Science*, Seiten 271–280, 2000.

Gurpreet Singh Matharu, Anju Mishra, Harmeet Singh, und Priyanka Upadhyay. Empirical Study of Agile Software Development Methodologies: A Comparative Analysis. *ACM SIGSOFT Software Engineering Notes*, 40(1):1–6, 2015. ISSN 0163-5948. DOI: 10.1145/2693208.2693233. https://www.cse.unr.edu/~dascalus/Paper_MOUNICA.pdf.

Peter Morville und Jeffery Callender. *Search Patterns: Design for Discovery*. Safari Tech Books Online. O'Reilly, Sebastopol, Calif., 1st ed. Auflage, 2010. http://proquest.safaribooksonline.com/9781449380205/http://www.gbv.eblib.com/patron/FullRecord.aspx?p=536695.

Jakob Nielsen. *Why You Only Need to Test with 5 Users*, 2000. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/. Letzter Abruf: 27.07.2025.

Jakob Nielsen und Thomas K. Landauer. A Mathematical Model of the Finding of Usability Problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, Seiten 206–213. ACM, New York, NY, USA, January 1993. ISBN 0-89791-575-5. DOI: 10.1145/169059.169166.

Henrik Nottelmann und Norbert Fuhr. From Uncertain Inference to Probability of Relevance for Advanced IR Applications. In Fabrizio Sebastiani, Hrsg., *Advances in Information Retrieval*, volume 2633 of *Lecture Notes in Computer Science*, Seiten 79–79. Springer Berlin / Heidelberg, 2003. http://dx.doi.org/10.1007/3-540-36618-0_17.

Open Data Support. *Einführung in das Metadaten-Management*, Europäische Kommission, 2013. https://data.europa.eu/sites/default/files/d2.1.2_training_module_1.4_introduction_to_metadata_management_de_edp.pdf. Letzter Abruf: 24.07.25.

OZG. *Gesetz zur Verbesserung des Onlinezugangs zu Verwaltungsleistungen (Onlinezugangsgesetz)*, 2024. https://tinyurl.com/ozg1-rev. Letzter Abruf: 26.07.2025.

OZGÄndG. *Gesetz zur Änderung des Onlinezugangsgesetzes sowie weiterer Vorschriften zur Digitalisierung der Verwaltung*, 2024. https://tinyurl.com/ozgaendg2. Letzter Abruf: 13.04.2025.

Caroline Paulick-Thiel und Henrike Arlt. *Öffentliches Gestalten: Handbuch für innovatives Arbeiten in der Verwaltung*. Technologiestiftung Berlin, 1. Auflage, 2020. ISBN 978-3-00-065930-0. https://t1p.de/b3ndy.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, und Francesco Osborne. Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review*, 56(11):13071–13102, 2023. ISSN 0269-2821, 1573-7462. DOI: 10.1007/s10462-023-10465-9.

Jennifer Preece, Yvonne Rogers, und Helen Sharp. *Interaction design: Beyond human-computer interaction*. Wiley, New York, NY, 2002.

Harald Reiterer, Gabriela Mußler, M. Thomas Mann, und Siegfried Handschuh. INSYDER - An Information Assistant for Business Intelligence. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, Seiten 112–119. ACM, 2000. ISBN 1-58113-226-3. http://doi.acm.org/10.1145/345508.345559.

Rektorenkonferenz d. HS f. d. öffentl. Dienst. *Digitale Grundkompetenzen an den Verwaltungsstudiengängen der Hochschulen für den öffentlichen Dienst – Orientierungen und Empfehlungen zur Umsetzung des Positionspapiers Verwaltungsstudiengänge der IMK vom 06.12.2024*, 2025. https://tinyurl.com/ovdigital. Letzter Abruf: 09.06.2025.

E. S. Robertson und Karen Spärck Jones. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976. http://dx.doi.org/10.1002/asi.4630270302.

E. Stephen Robertson. The Probability Ranking Principle in IR. In *Journal of Documentation*, volume 4, Seiten 281–286. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1977. ISBN 1-55860-454-5. http://portal.acm.org/citation.cfm?id=275537.275701.

Tony Russell-Rose und Tyler Tate. *Designing the Search Experience: The Information Architecture of Discovery*. Morgan Kaufmann, Amsterdam, Boston, Heidelberg, 2013.

Gerard Salton und Christopher Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988. http://dx.doi.org/10.1016/0306-4573(88)90021-0.

Gerard Salton, A. Edward Fox, und Harry Wu. Extended Boolean Information Retrieval. *Commun. ACM*, 26(11):1022–1036, 1983.

Eve Sauvage, Sabrina Campano, Lydia Ouali, und Cyril Grouin. Does the Structure of Textual Content Have an Impact on Language Models for Automatic Summarization? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, Seiten 280–285. Association for Computational Linguistics, 2024. DOI: 10.18653/v1/2024.acl-srw.25.

Ingo Schmitt. *Ähnlichkeitssuche in Multimedia-Datenbanken: Retrieval, Suchalgorithmen und Anfragebehandlung*. Oldenbourg, München, 2006.

Elwood Claude Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 1948.

Ben Shneiderman und Catherine Plaisant. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson, Boston, 4. Auflage, January 2005. ISBN 0-321-19786-0.

Dagobert Soergel. *Organizing Information: Principles of Data Base and Retrieval Systems*. Academic Press Professional, Inc., San Diego, CA, USA, 1985.

Ian Sommerville. *Software Engineering*. Informatik. Pearson Studium, 6. [2. Nachdr.] Auflage, 2003. ISBN 3-8273-7001-9.

Staatsbibliothek zu Berlin. *SBB StaBiKat - Online-Hilfe*, 2001. https://t1p.de/k2xhc. Letzter Abruf: 22.07.2025.

Toni Steimle und Dieter Wallach. *Collaborative UX Design: Lean UX Und Design Thinking: Teambasierte Entwicklung Menschzentrierter Produkte*. dpunkt.verlag, 1. Auflage, 2018. ISBN 978-3-86490-532-2.

Lisa Stubert, Klemens Maget, Max Bruno Eckert, und Hans Hack. *Linked Open Data in der Praxis – Vernetzte Verwaltungsdaten am Beispiel der Berliner Organigramme*, 2025. https://t1p.de/p9n6b. Letzter Abruf: 27.07.2025.

G. Arlene Taylor. *Introduction to cataloging and classification*. Library and Information Science Text Series. Libraries Unlimited, Westport, Conn., 10th ed. Auflage, 2006. http://www.loc.gov/catdir/toc/ecip0612/2006011701.html/http://www.gbv.de/dms/bowker/toc/9781591582304.pdf.

Cornelis Joost "Keith" van Rijsbergen. *Information Retrieval*. Butterworths, London, 2 Auflage, 1979.

Cornelis Joost "Keith" van Rijsbergen. A Non-classical Logic for Information Retrieval. In *The Computer Journal*, volume 29(6), Seiten 481–485. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1986. ISBN 1-55860-454-5. http://portal.acm.org/citation.cfm?id=275537.275699.

VgV. *Verordnung über die Vergabe öffentlicher Aufträge (Vergabeverordnung)*, 2016. https://www.verwaltungsvorschriften-im-internet.de/bsvwvbund_14032001_DokNr20110981762.htm. Letzter Abruf: 26.07.2025.

VV-BHO. *Allgemeine Verwaltungsvorschriften zur Bundeshaushaltsordnung*, 2025. https://tinyurl.com/vv-bho. Letzter Abruf: 26.07.2025.

G. W. Waller und H. Donald Kraft. A Mathematical Model of a Weighted Boolean Retrieval System. *Information Processing and Management*, 15(5):235–245, 1979.

Richard Y. Wang und Diane M. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4):5–33, 1996. ISSN 07421222. http://www.jstor.org/stable/40398176. Letzter Abruf: 2025-07-24.

W. Ryen White und Resa Roth. *Exploratory Search: Beyond the Query-response Paradigm*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, San Rafael, 2009.

Mark D. Wilkinson, Michel Dumontier, und Aalbersberg et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016. ISSN 2052-4463. DOI: 10.1038/sdata.2016.18. https://www.nature.com/articles/sdata201618. Letzter Abruf: 01.08.2025.

R. R. Yager. On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making. *IEEE Trans. on Systems, Man, and Cybernetics*, 18(1):183–190, 1988.

A. Lotfi Zadeh. Fuzzy Logic. *IEEE Computer*, 21(4):83–93, 1988.

Lotfi A. Zadeh. Fuzzy Sets. *Information and Control*, 3(8):338–353, January 1965. DOI: 10.1016/S0019-9958(65)90241-X.

David Zellhöfer. *A Preference-based Relevance Feedback Approach for Polyrepresentative Multimedia Retrieval*. Dissertationsschrift, BTU Cottbus - Senftenberg, 2015.

David Zellhöfer. Agilität und Weltkulturerbe: Erfahrungen mit sieben Jahren agilen Methoden an der Staatsbibliothek zu Berlin – Eine Fallstudie I/II. *ABI Technik*, 41(3):194–201, 2021. ISSN 2191-4664, 0720-6763. DOI: 10.1515/abitech-2021-0032. https://www.degruyter.com/document/doi/10.1515/abitech-2021-0032/html. Letzter Abruf: 16.04.2025.

David Zellhöfer. Methodenvermittlung des Software Engineerings als Hebel für die Digitale Transformation der Öffentlichen Verwaltung. In Veronika Thurner, Barne Kleinen, Juliane Siegeris, und Debora Weber-Wulff, Hrsg., *Software Engineering im Unterricht der Hochschulen*, volume P-321 of *LNI*, Seiten 137–148. Gesellschaft für Informatik, Bonn, 2022. DOI: 10.18420/SEUH2022_14.

David Zellhöfer. Der Begriff „Leuchtturm“ ist in der Regel ein Warnsignal. *Der Tagesspiegel (23.05.2023)*, Seite B 24, 2023. https://tinyurl.com/kolumne-tsp. Letzter Abruf: 24.07.2025.

David Zellhöfer, Oliver Schöner, und Gerrit Gragert. *Building Library Information Systems in Times of Vanishing Developer Resources – A Fedora-4-based Approach*. 40th European Library Automation Group Systems Seminar, Fachvortrag, 2019.

Jin Zhang. *Visualization for Information Retrieval*, volume 23 of *Springer-11645 / Dig. Serial]*. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2008. http://dx.doi.org/10.1007/978-3-540-75148-9.

H.-J. Zimmermann. *Fuzzy Set Theory - And Its Applications*. Kluwer Academic Publishers, Norwell, MA, USA, 3 Auflage, 1996.

M. M. Zloof. Query By Example. In *Proc. of AFIPS National Computer Conference*, volume 44, Seiten 431–438. AFIPS Press, 1975.

Anhang

A.1 Übersetzung des englischsprachigen Zitats (S. 28)

> Da die Bibliothek zur Software geworden ist, ist es nicht länger denkbar, dass unsere Dienstleistungen getrennt von unserer Software existieren. Unsere besten Möglichkeiten für Unterstützung, Beratung und Anleitung liegen innerhalb unserer Software.
>
> Unsere Nutzenden sind täglich erfolgreich bei der Informationssuche im Internet, beim Einkaufen, [...].
>
> In keinem dieser Fälle erwarten sie, dass für ihr Vorhaben wichtige Dienste ohne die damit verbundene jeweiligen Software existieren könnnen. Vielmehr ist die Software selbst der Dienst.
>
> *Übersetzt nach Hanson (2015); siehe Seite 28.*

A.2 Information Retrieval in PostgreSQL – Implementierungsskizze

Da die Vorteile des *Information Retrievals* gerade bei der Suche nach natürlichsprachigen Texten in der Fachwelt bekannt sind, versuchen auch die Hersteller von *relationalen Datenbankmanagement-Systemen* (RDBMS) diese Funktionen nachzubilden, obwohl sie eigentlich außerhalb des zugrundeliegenden Booleschen Modells liegen.

Texte und Zeichenketten lassen sich seit langem in den Tabellen von RDBMS ablegen. Hierfür stehen spezifische Datentypen wie Char, Varchar, Text oder CLOB zur Verfügung.

Anfragen erfolgen, wie in Abschnitt 5.2 beschrieben, mittels des Like-Operators unter Einbeziehung von Wildcards, um ein exaktes Erkennungsmuster spezifizieren zu können.

Hinzu kommt der Similar To-Operator, welcher reguläre Ausdrücke verarbeitet.

Die Verwendung der beiden genannten Operatoren führt jedoch nicht dazu, dass ein *Best Matching*, wie es beim *Information Retrieval* üblich ist, durchgeführt wird.

Auch Methoden wie die Stammformreduktion, um semantisch ähnliche Begriffe wie „Haus“ und „Häuser“ unter einem Begriff zusammenzufassen, stehen hier nicht zur Verfügung.

Zur Umsetzung einer zeitgemäßen Textsuche verfolgen die RDBMS-Anbieter unterschiedliche Wege.

Das im Folgenden skizzierte Beispiel basiert auf PostgreSQL[41] und setzt die Version 8.3 voraus, die 2008 erschienen ist.

[41] https://www.postgresql.org/docs/18/textsearch-controls.html; Letzter Abruf: 01.08.2025

Datenbasis

Gegeben sei die nachfolgende Tabelle `ir_test`, in deren Spalte `title` im weiteren Verlauf gesucht werden soll:

```
id|title
-----------------------------------------------------
 1|i want out - helloween
 2|hello world
 3|hello hello
 4|hello [is it me you are looking for?] - lionel richie
 5|caesar - the oh hellos
```

Textsuche mittels Standard-SQL

Mittels Standard-SQL kann nach allen Zeilen gesucht werden, welche die Buchstabenfolge `llo` enthält.

```
1  SELECT title FROM ir_test WHERE title SIMILAR TO '%(llo)%';
```

Alternativ können alle Zeilen angefragt werden, welche das Wort „hello“ enthalten:

```
1         SELECT title FROM ir_test WHERE title LIKE '%hello%';
```

Als Rückgabe erhält man in beiden Fällen:

```
title
-----------------------------------------------------
i want out - helloween
hello world
hello hello
hello [is it me you are looking for?] - lionel richie
caesar - the oh hellos
```

Vorbereitungen zur modernen Textsuche

Möchte man nun eine moderne Textsuche umsetzen, muss der Tabelle eine weitere Spalte hinzugefügt werden, welche den Datentyp `tsvector` erhält. Dieser Datentyp nimmt Lexeme auf, die einer Stammformreduktion (siehe unten) unterzogen wurden, und hält deren Position im Ausgangstext fest.

Im Beispiel wird Englisch als Sprache des Texts in Spalte `title` festgelegt und die neue Spalte `textsearch` mit den Ergebnissen der PostgreSQL-Funktion `to_tsvector()` gefüllt. Diese Funktion berechnet die jeweiligen `tsvector` und entfernt englische Stoppworte (siehe unten) sowie Interpunktionszeichen.

```
1  ALTER TABLE ir_test ADD COLUMN textsearch tsvector;
2
3  UPDATE ir_test SET textsearch=to_tsvector('english',title);
```

Um auf der neuen Spalte eine Textsuche durchführen zu können, ist es notwendig, einen speziellen Index zu erstellen. Hierbei handelt es sich um den sogenannten *Generalized Inverted Index* (GIN).

Der Indexerstellung folgt die Ausgabe der entstandenen Tabelle.

```
1
2  CREATE INDEX textsearch_idx ON ir_test USING GIN (textsearch);
3
4  SELECT * FROM ir_test;
```

In der unten stehenden Tabelle sieht man in Zeile 5, dass die Stammformreduktion bei „hellos“ aktiv wurde und den Begriff auf das Lexem „hello“ abgebildet hat. Der Term „helloween“ wird gesondert behandelt.

Außerdem wurden Stoppworte wie „i“ und „out“ entfernt, da sie sehr häufig in englischen Texten vorkommen.

```
id|title                                                |textsearch
--|-----------------------------------------------------|---------------------------
 1|i want out - helloween                               |'helloween':4 'want':2
 2|hello world                                          |'hello':1 'world':2
 3|hello hello                                          |'hello':1,2
 4|hello [is it me you are looking for?] - lionel richie |'hello':1 'lionel':9 [...]
 5|caesar - the oh hellos                               |'caesar':1 'hello':4 'oh':3
```

Moderne Textsuche

Um in einer Spalte im `tsvector`-Format suchen zu können, implementiert PostgreSQL den Matching-Operator `@@`.

Dieser erwartet als Operand ebenfalls eine `tsvector`-Repäsentation der Anfrage „hello“. Diese wird erneut mittels der bereits bekannten Funktion `to_tsvector()` erzeugt.

```
1  SELECT title FROM ir_test WHERE textsearch @@ to_tsquery('hello');

   title
   -----------------------------------------------------
   hello world
   hello hello
   hello [is it me you are looking for?] - lionel richie
   caesar - the oh hellos
```

Im Ergebnis wird deutlich, dass nun alle Titel enthalten sind, die „hello“ oder ein semantisches Äquivalent enthalten: „i want out - helloween“ ist folglich nicht mehr Teil der Rückgabe.

Da das relationale Datenbankmodell von sich aus keine Ordnung kennt, wird jedoch eine ungeordnete Multimenge zurückgegeben.

Relevanzsortierung

Wie bereits ausgeführt wurde, ist eine Relevanzsortierung im Sinne des *Best Matchings* notwendig, damit Anwenderinnen und Anwender besonders relevante Treffer schnell entdecken können.

Hierzu ist es nötig, eine weitere berechnete Spalte `rank` mittels der Funktion `ts_rank_cd()` zu erstellen, welche danach zur Sortierung herangezogen werden kann:

```
1  SELECT title, ts_rank_cd(textsearch, query) AS rank FROM ir_test, to_tsquery('hello') query
2          WHERE query @@ textsearch ORDER BY rank DESC;
```

Erwartungsgemäß wird nun die Zeile, welche den Suchbegriff am häufigsten enthält, oben platziert.

```
title                                               |rank
----------------------------------------------------|----
hello hello                                         | 0.2
```

```
hello world                                          | 0.1
hello [is it me you are looking for?] - lionel richie | 0.1
caesar - the oh hellos                               | 0.1
```

Fazit

Das Beispiel zeigt auf, dass der Umgang mit Texten und der Textsuche in RDBMS möglich, aber unelegant ist.

Hier eignen sich *Information-Retrieval-Systeme* wesentlich besser, da diese häufig auch noch Funktionen zur Hervorhebung von Suchtreffern oder weitere Anfragemöglichkeiten bereitstellen, die weit über die hier gezeigte Schlagwort-basierte Suche hinausgehen.

A.3 Übersetzung der Fedora-Entwicklungsziele (S. 99)

> *Persistente Identifer (Identifiers)* Bereitstellung von persistenten Identifiern; eindeutige Namen für alle Ressourcen ohne Bezug auf die Maschinenadresse
>
> *Beziehungen (Relationships)* Unterstützung von Beziehungen zwischen Objekten [vgl. *Linked (Open) Data*]
>
> *Gezähmte Inhalte (Tame Content)* Normalisierung heterogener Inhalte und Metadaten, basierend auf einem erweiterbaren Objektmodell
>
> *Integriertes Management (Integrated Management)* Effiziente Verwaltung durch Repository-Administratoren nicht nur von Daten und Metadaten in einem Repository, sondern auch der sonstigen Programme, Dienste und Werkzeuge, welche die Präsentation dieser Daten und Metadaten ermöglichen
>
> *Interoperabler Zugriff (Interoperable Access)* Sicherstellung von Interoperabilität mittels eines Standardprotokolls zum Austausch von Informationen über Objekte und für den Zugriff auf Objektinhalte, sowie zum Auffinden und zur Ausführung flexibler Operationen auf digitalen Objekten
>
> *Skalierbarkeit (Scalability)* Unterstützung von mehr als 10 Millionen Objekten
>
> *Sicherheit (Security)* Zurverfügungstellung flexibler Authentifizierung und Security-Policy-Durchsetzung
>
> *Langzeitarchivierung (Preservation)* Bereitstellung von Funktionen zur Unterstützung Langzeitarchivierung, einschließlich textbasierter Serialisierung von Objekten und Versionierung von Inhalten
>
> *Nachnutzung von Inhalten (Content Recon*[42]*)* Wiederverwendung von Objekten, einschließlich der Möglichkeit, dass Objektinhalte in beliebig vielen Kontexten innerhalb eines Repositories wiederverwendet werden; Umnutzung von Objekten, die dynamische Inhaltstransformationen ermöglichen, um neue Präsentationsanforderungen zu erfüllen

Übersetzt nach `https://fedorarepository.org/core-attributes-of-fedora-repository-enable-complex-modeling-of-data-and-objects-for-re-use-in-a-wide-variety-of-instances/`; *Letzter Abruf: 01.08.2025; siehe Seite 99.*

A.4 Automatisierter Datenabruf mittels SPARQL

Abschnitt 3.6 führte, neben den Ideen des *Semantic Web* bzw. *Linked (Open) Data*, das Konzept des *Knowledge Graphs* ein.

Die Wikidata[43] stellt einen frei verfügbaren *Knowledge Graph* dar, welcher durch Dritte, z.B. über einen SPARQL-Endpoint[44], angefragt werden kann.

Da die Daten der Wikidata unter einer offenen Lizenz vorliegen und damit frei im Sinne des Urheberrechts nachgenutzt werden können, wäre die Wikidata ein typischer Partner eines Recherchesystems, welches auf den Prinzipien von *Linked (Open) Data* basiert.

[43] https://www.wikidata.org

[44] https://query.wikidata.org/

Beispielhaft wurde im Hauptteil des Gutachtens Arbeitsschritt 4.1 Informationsrecherche VI – Schlagwort-basierte Suche: Verfeinerung der Filterung vorgestellt, bei dem es möglich ist, ungültige Kombinationen aus dem Ministerium mit Fachaufsicht und den fachlich zuständigen Behörden zu wählen.

Würden die Daten im Datenatlas Graph-basiert vorliegen, könnte ein solches Fehlbedienungsrisiko einfach ausgeschlossen werden, wenn nach der Selektion eines Ministeriums nur noch der Teilgraph selektierbar wird, welcher Einrichtungen beinhaltet, die diesem tatsächlich untergeordnet sind.

Dazu soll im Folgenden ein entsprechender Lösungsansatz skizziert werden, der so – trotz seiner Einfachheit – aktuell nicht im Datenatlas umgesetzt werden kann.

Aufgrund der unzureichenden SPARQL-Formulierungspraxis des Autors, wandte dieser sich mit der folgenden Bitte um Unterstützung bei der Formulierung einer SPARQL-Anfrage für Wikidata am 29. Juli 2025, 17:31 Uhr, auf dem Social-Media-Dienst Mastodon an die Öffentlichkeit:

> [...] The query should list all German federal ministries and their associated federal or otherwise supervised federal agencies etc. The list should make visible which federal ministry is „in charge“ of each entity.[45]

[45] https://openbiblio.social/@david/114937237762430186; Letzter Abruf: 29.07.2025

Bereits 15 Minuten später antworte der Nutzer Daniel Baránek (daelba@sciences.social) mit der folgenden SPARQL-Anfrage:[46]

[46] Die Anfrage ist direkt unter https://w.wiki/Et3g; Letzter Abruf: 29.07.2025 ausführbar.

```
SELECT ?item ?itemLabel ?min ?minLabel WHERE {
        SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],mul,en". }
        ?item (wdt:P112|wdt:P749) ?min.
        ?min wdt:P31 wd:Q896375
}
```

Abbildung A.2 bildet einen Ausschnitt des Anfrage-Ergebnisses in tabellarischer Form ab.

Hierbei beinhaltet die Spalte *itemLabel* die nachgeordneten Behörden und die Spalte *minLabel* die Bundesministerien mit Fachaufsicht. Zusätzlich sind die *Persistent Identifier* der jeweiligen Entitäten in *item* und *min* ausgewiesen.

item	itemLabel	min	minLabel
wd:Q1005427	Bundesamt für Bevölkerungsschutz und Katastrophenhilfe	wd:Q502698	Bundesministerium des Innern
wd:Q1005429	Bundesamt für Bauwesen und Raumordnung	wd:Q502698	Bundesministerium des Innern
wd:Q1005454	Bundesamt für Kartographie und Geodäsie	wd:Q502698	Bundesministerium des Innern
wd:Q1005470	Bundesamt für Migration und Flüchtlinge	wd:Q502698	Bundesministerium des Innern
wd:Q1005522	Bundesamt für zentrale Dienste und offene Vermögensfragen	wd:Q502698	Bundesministerium des Innern
wd:Q1005549	Bundesanstalt für den Digitalfunk der Behörden und Organisationen mit Sicherheitsaufgaben	wd:Q502698	Bundesministerium des Innern
wd:Q1005650	Bundesausgleichsamt	wd:Q502698	Bundesministerium des Innern

Abbildung A.2: Tabellarische Darstellung der SPARQL-Anfrage

Eine visuelle Aufarbeitung den resultierenden Ausschnitts des *Knowledge Graphs* der WIKIDATA findet sich in Abb. A.3.

Aus illustrativen Gründen ist hier das BUNDESMINISTERIUM FÜR DIGITALES UND STAATSMODERNISIERUNG hervorgehoben, um die Aktualität des Graphen zu unterstreichen.

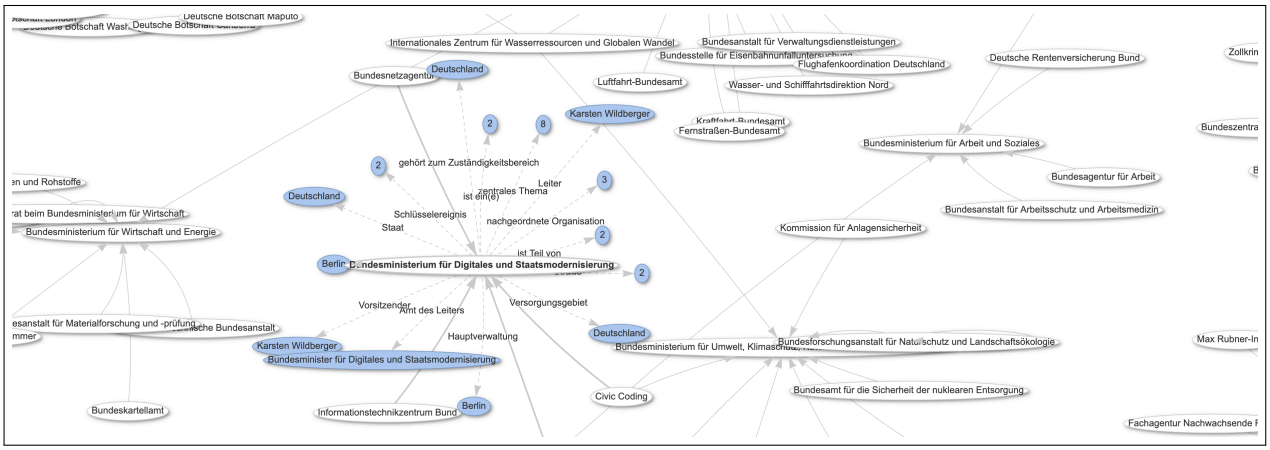


Abbildung A.3: Graph der SPARQL-Anfrage (Ausschnitt)

Am gleichen Tag um 19:04 Uhr stellte der Nutzer Jan Ainali (ainali@social.coop) eine weitere Verfeinerung der Anfrage zur Verfügung[47].

Die oben genannten SPARQL-Anfragen könnten direkt in einer auf *Linked (Open) Data-Technologien* basierenden Plattform nachgenutzt werden, um das oben beschriebene Usability-Problem zu umgehen.

[47] Die Anfrage ist direkt unter `https://w.wiki/Et6B`; Letzter Abruf: 29.07.2025 ausführbar.

A.5 Anfrage an die Bundesdruckerei

Anfrage vom 31.07.2025 an `datenatlas@bdr.de`:

> Liebe Kolleg*innen bei der Bundesdruckerei,
>
> ich bin Professor für Digitale Innovation in der öffentlichen Verwaltung und arbeite in dieser Rolle gerade meine Folien zum Thema Datenmanagement und Datenkompetenzen für die Ausbildung des Verwaltungsnachwuchses auf Landes- und Bundesebene. Dabei stelle ich diverse Systeme vor, die bereits existieren bzw. für den Einsatz vorgesehen sind.
>
> Zusätzlich skizziere ich immer einige wenige technische Infrastruktur-Aspekte etc., damit die ÖV-Studierenden (keine Techniker!) erkennen

können, dass es viele, komplexe Ansätze gibt und moderne Software-Entwicklung viele verschiedene Komponenten enthält. Leider konnte ich den Datenatlas nicht auf openCode, Github oder Gitlab finden und mir die Fragen selbst beantworten.

Könnten Sie vielleicht, ohne große ins Detail zu gehen, ein paar Parameter nennen bzw. diese Anfrage ans Architektur-Team weitergeben? Um das Ausfüllen zu erleichtern nenne ich gleich ein paar Beispiele für die Aspekte, die ich immer aufliste, und bedanke mich für Ihre Hilfe. Es geht wirklich nur um die Namen/Produkte und nicht die Konzepte.

1. Geplante Lizenz: z.B. Closed oder Open Source
2. Frontend: z.B. Java Server Pages, Custom jQuery, React etc.
3. Middleware: Jakarta/J2EE, node.js, .NET, SAP etc
4. Persistenz/Suche: PostgreSQL, Elasticsearch, Oracle, Redis, SAP etc.
5. Log-Analyse: Kibana/Elasticsearch, Solr, Splunk
6. Betriebssystem: Linux, Windows, macOS
7. Hauptsächlich genutzte Programmiersprache/Frameworks: PHP, Java, JavaScript, Ruby etc.

Vielen Dank für Ihre Unterstützung.

Viele Grüße nach Kreuzberg!

Prof. Dr.-Ing. David Zellhöfer

Professor für Digitale Innovation in der öffentlichen Verwaltung; Fachbereich 3, Allg. Verwaltung

[...]

Antwort vom 19.08.2025 seitens `datenatlas_pmo@bdr.de`

Sehr geehrter Herr Prof. Zellhöfer,

wir danken Ihnen für Ihr – anhaltendes - Interesse am Datenatlas. Viele Informationen zum Projekt finden Sie auf der Website Datenatlas Bund: Basis datengetriebener Verwaltung[48]. Für darüber hinaus gehende Informationen stehen wir Ihnen nicht zur Verfügung.

Mit freundlichen Grüßen

[Aus Datenschutzgründen entfernt]

Ihr Datenatlas-PMO

Bundesdruckerei GmbH [...]

[48] `https://www.bundesdruckerei.de/de/innovation-hub/projekt-datenatlas`

A.6 Generierungsbeispiel RDF-Graph

Abbildung A.4 wurde durch den *Online RDF Graph Visualizer*[49] unter Beibehaltung der Standard-Einstellungen erzeugt.

Als Grundlage der Visualisierung wurde die GND-ID 2117465-9, angegeben. Hierbei handelt es sich um einen *Persistent Identifier*, welcher das Bundesministerium des Innern bezeichnet.

[49] `https://issemantic.net/rdf-visualizer`; Letzter Abruf: 01.08.2025

Die auf der Website verwendeten Daten[50] liegen imRDF/Turtle-Format vor und wurden automatisch detektiert.

Die Überblendung am unteren Bildrand wurde nachträglich aus ästhetischen Gründen ergänzt.

[50] https://d-nb.info/gnd/2117465-9/about/lds

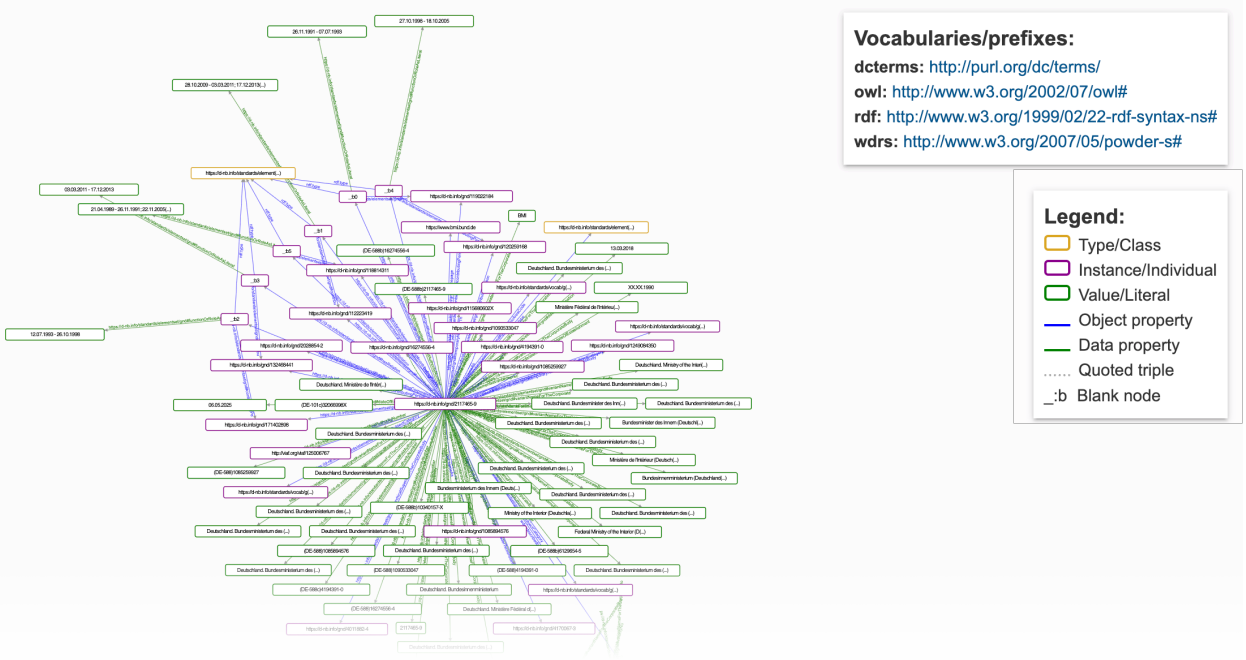


Abbildung A.4: RDF-Graph am Beispiel des Bundesministeriums des Innern

Index