

H T
W I
G N

Hochschule Konstanz
Department of Computer Science

Submitted by
Jonas Mayer
Student Number 305630

jonas.mayer@htwg-konstanz.de

B

C



Bachelor Thesis

Datenanalyse PEP

S



Konstanz, TODO

Bachelor Thesis

Datenanalyse PEP

by

Jonas Mayer

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science
in Applied Computer Science

at the Hochschule Konstanz University of Applied Sciences,

Student Number: 305630

Date of Submission: TODO

Supervisor: **Prof. Dr. Doris Bohnet**
Second Examiner:

An electronic version of this thesis is available at <https://github.com/jonez187/bachelorarbeit--htwg-latex>.

Abstract

Hier Abstract schreiben

Inhaltsverzeichnis

Abstract	iii
1 Einleitung	1
2 Theoretische Grundlagen	3
2.1 PEP-Ecopassport	3
2.1.1 PEP-Standard	3
2.1.2 Aufbau typischer PEP-Dokumente	4
2.2 Datenextraktion aus PDF-Dokumenten	7
2.2.1 Das Portable Document Format (PDF)	7
2.2.2 Herausforderungen bei der automatisierten Extraktion	8
2.2.3 Extraktionsansätze	9
2.2.4 Zielformat JSON	13
2.2.5 Informationsextraktion ins Zielformat	13
2.3 Statistische Grundlagen	15
2.3.1 Deskriptive und explorative Statistik	15
2.3.2 Explorative Datenanalyse (EDA) und Visualisierungen	16
2.3.3 Automatisierung, Reproduzierbarkeit und Datenqualität	17
2.3.4 Lineare Regression	17
3 Pipeline und Datenbasis (Methodik)	19
3.1 Überblick der Pipeline	19
3.1.1 Datenerhebung und PEP-Recherche	20
3.1.2 PDF-Parsing und Extraktion	22
3.1.3 Normalisierung und Datenbereinigung	24
3.2 Datenbasis	24
3.2.1 Variablenstruktur	24
3.2.2 Qualitätssicherung	24
4 Analyse der erarbeiteten Daten	25
4.1 Deskriptive Annäherung an die PEP-Daten	25
4.1.1 Vollständigkeit der Werte	25
4.1.2 Überblick der <i>Input</i> -Variablen	26
4.1.3 Überblick der Umweltindikatoren	29

1

Einleitung

Hier Einleitung schreiben

TESTSTETS

2

Theoretische Grundlagen

Hier werden die theoretischen Grundlagen für die vorliegende Arbeit gelegt. Ausgangspunkt ist die Beobachtung, dass Smart-Home-/IoT-Produkte Umweltwirkungen nicht nur in der Nutzungsphase (Stromverbrauch), sondern ebenso durch Materialzusammensetzung, Fertigung, Distribution und Entsorgung verursachen. Für die standardisierte Berichterstattung solcher Wirkungen existieren deklarative Formate wie die PEP Eco-passports, die Indikatoren entlang des Lebenszyklus ausweisen. Damit diese Angaben für quantitative Analysen nutzbar werden, sind konsistente Begriffe, Einheiten und Moduldefinitionen ebenso erforderlich wie ein Verständnis zentraler statistischer Verfahren zur Muster- und Zusammenhangsanalyse. Dieses Kapitel führt daher zunächst in Struktur und Inhalte von PEP-Deklarationen ein und skizziert anschließend die methodischen Bausteine (u. a. Lineare Regression und PCA), die in den folgenden Kapiteln zur Reduktion von Variablen, zur Erklärung von Indikatorvarianz und zur Ableitung einer praxistauglichen Heuristik für Produkte ohne PEP eingesetzt werden.

2.1. PEP-Ecopassport

Was ist PEP-Ecopassport, was steht drin, was ist interessant für mich

2.1.1. PEP-Standard

Der *PEP Ecopassport®* ist ein international anerkanntes Programm zur Erstellung standardisierter Umweltpunktdeklarationen für elektrische, elektronische sowie Heizungs-

, Lüftungs-, Klima- und Kälteprodukte (HVAC). Träger des Programms ist die *P.E.P. Association*, eine gemeinnützige Organisation, deren Ziel es ist, ein gemeinsames und verlässliches Referenzsystem für Umweltinformationen dieser Produktkategorien bereitzustellen. Das Programm versteht sich als Branchenspezialisierung innerhalb des Rahmens der *Environmental Product Declarations (EPD)* gemäß ISO 14025 und der Lebenszyklusnormen nach ISO 14040, und basiert somit auf international festgelegten Normen. [Ass24]

Ein *PEP Ecopassport* ist somit eine *Typ III-Umweltdeklaration* im Sinne der ISO 14025. Diese Deklarationen basieren auf quantitativen Ergebnissen einer Lebenszyklusanalyse (*Life Cycle Assessment, LCA*) und dienen der vergleichenden Bewertung von Produkten mit identischer Funktion. Die Datenerhebung und Berechnung erfolgt nach vordefinierten Parametern, die in sogenannten *Product Category Rules (PCR)* und bei Bedarf in *Product Specific Rules (PSR)* festgelegt sind. Jede PEP-Deklaration unterliegt einer unabhängigen Überprüfung der angewandten Methodik und der zugrunde liegenden LCA-Daten. [Has+13]

Das Programm zielt auf Transparenz und Vergleichbarkeit ab. Hersteller erhalten ein einheitliches Verfahren, um ökologische Leistungskennwerte ihrer Produkte objektiv und nachvollziehbar zu kommunizieren. Für Anwender, Beschaffer und Energieberater stellen die PEP-Daten eine verlässliche Grundlage für ökologische Bewertungen und Beschaffungsentscheidungen dar.

Die Teilnahme am PEP-Programm ist freiwillig, gewinnt jedoch in der Praxis an Bedeutung, da Umweltproduktdeklarationen zunehmend als Nachweis oder Auswahlkriterium in Ausschreibungen und Produktbewertungen herangezogen werden. Eine gesetzliche Verpflichtung zur Erstellung besteht bislang nur in Einzelfällen, beispielsweise in Frankreich, wenn ein Hersteller aktiv mit Umweltvorteilen wirbt.

Das PEP-Programm unterscheidet sich klar von unternehmensbezogenen Treibhausgas-Bilanzierungen: Es erfasst ausschließlich produktspezifische Umweltwirkungen entlang des Lebenszyklus und folgt dabei den methodischen Vorgaben der ISO 14040-Reihe. Für umfassende *GHG-Assessments* auf Organisationsebene sind PEP-Daten daher nicht geeignet. [Ass24]

2.1.2. Aufbau typischer PEP-Dokumente

Ein vollständiges PEP umfasst typischerweise etwa zehn Seiten und gliedert sich in mehrere inhaltlich definierte Abschnitte.

Titel- und Metadatenblatt Das Deckblatt enthält grundlegende Angaben zum Produkt (Name, Version, Sprache, Hersteller), zum Veröffentlichungs- und Revisionsdatum

sowie zum Status der Erklärung (z. B. *in review* oder *verified*). Darüber hinaus sind Kontaktinformationen, Firmenadresse und Registrierungsnummer enthalten.

Allgemeine Produktinformationen Dieser Abschnitt beschreibt die funktionale Einheit (*functional unit*), in welcher auch der Stromverbrauch dargestellt ist. Weiterhin werden Referenzlebensdauer, hier meist 10 bis 20 Jahre, die Produktfunktion, Anwendungsbereiche und gegebenenfalls weitere Varianten aufgeführt.

Materialzusammensetzung Die Zusammensetzung des Produkts wird tabellarisch nach Hauptgruppen ausgewiesen, z. B. Kunststoffe, Metalle und weitere Materialien (Papier/Karton, Elektronik, Sonstiges). Diese Angaben ermöglichen eine spätere Aggregation der Stoffanteile in harmonisierten Datenstrukturen.

Szenarien und Lebenszyklusphasen PEP-Dokumente sind entlang der Phasen des Produktlebenszyklus strukturiert, die den Vorgaben der EN 15804 entsprechen:

- **Herstellung (A1–A3):** Produktion und Vormaterialien, modelliert mit landesspezifischem Strommix (z. B. italienischer Grid Mix).
- **Distribution (A4):** Transport vom Werk zum Markt; häufig standardisierte Annahmen (z. B. 1 000 km Schiff, 3 300 km Lkw).
- **Installation (A5):** Montage, meist nur Verpackungsabfälle berücksichtigt.
- **Nutzungsphase (B):** Betrieb des Geräts mit angegebenem Energieverbrauch, z. B. 126 kWh über 20 Jahre, basierend auf europäischem Netzstrommix.
- **End-of-Life (C1–C4):** Entsorgungsszenario gemäß PCR-Vorgaben (Recycling-, Deponie-, Transportanteile).
- **Optionale Phase (D):** Rückgewinnung und Wiederverwendung außerhalb des Systemgrenzenmodells.

In der weiteren Datenaufbereitung werden diese Phasen zu den Kategorien *manufacturing*, *distribution*, *installation*, *use* und *end_of_life* zusammengefasst.

Energiemodelle Zusätzlich werden die verwendete Energiemodelle angegeben (z. B. *France Grid Mix*). Die Genauigkeit dieser Angaben variiert deutlich zwischen den Dokumenten. In einigen Fällen ist jeder Produktlebenszyklusphase ein spezifisches Land zugeordnet, während andere PEPs für alle Phasen einen einheitlichen europäischen Strommix angeben.

Umweltindikatoren Die Umweltwirkungen werden für jede Lebenszyklusphase sowie als Gesamtwert angegeben. Die für diese Arbeit relevanten Indikatoren sind in der Tabelle 2.1 aufgeführt.

Tabelle 2.1: Umweltindikatoren

Indikator	Beschreibung
Acidification	Versauerung von Böden und Gewässern durch säurebildende Emissionen
Climate Change (Fossil)	Treibhauspotenzial durch fossile CO ₂ -Emissionen
Climate Change (Land Use and Land Use Change)	Treibhauspotenzial infolge von Landnutzungsänderungen (LULUC)
Climate Change (Total)	Gesamtes Treibhauspotenzial aus allen Quellen
Eutrophication (Freshwater)	Nährstoffanreicherung in Binnengewässern
Eutrophication (Marine)	Nährstoffanreicherung in marinen Ökosystemen
Eutrophication (Terrestrial)	Nährstoffanreicherung in terrestrischen Ökosystemen
Hazardous Waste Disposed	Entsorgung gefährlicher Abfälle
Ozone Depletion	Abbau der stratosphärischen Ozonschicht durch FCKW-Emissionen
Photochemical Ozone Formation (Human Health)	Bildung von bodennahem Ozon (Sommersmog)
Radioactive Waste Disposed	Entsorgung radioaktiver Abfälle
Resource Use (Fossils)	Nutzung fossiler Energieressourcen
Resource Use (Minerals and Metals)	Verbrauch abiotischer Ressourcen (Metalle und Mineralien)
Water Use	Entnahme und Verbrauch von Frischwasser

Verifikations- und Anhangsangaben Im abschließenden Teil werden die angewendeten Regelwerke und Datenquellen genannt, z. B. *PCR-ed3-EN-2015_04_02* und *PSR-0005-ed2-EN-2016_03_29*, die eingesetzte Software (z. B. SimaPro 9.3 mit Ecoinvent 3.8) sowie die Verifizierungsstelle und deren Akkreditierungsnummer. Darüber hinaus enthält dieser Abschnitt Angaben zum *Materialaufbau* und zum verwendeten *Energiemodell*. Die Materialzusammensetzung wird in der Regel als prozentuale Massenanteile nach Hauptgruppen (Kunststoffe, Metalle, Elektronik, Sonstiges) dargestellt, teils in Tabellenform, teils grafisch als Kreisdiagramm. Das Energiemodell beschreibt die angenommenen Strommixe und Spannungsniveaus je Lebenszyklusphase, beispielsweise den nationalen Grid Mix für die Herstellung und den europäischen Durchschnittsmix für die Nutzungsphase.

Obwohl der inhaltliche Mindestumfang und die zu berichtenden Umweltindikatoren durch die zugrundeliegenden ISO- und PCR-Vorgaben festgelegt sind, besteht keine feste formale Struktur. Das Layout, die grafische Aufbereitung und die Anordnung der Tabellen können je nach Hersteller, Software und Version variieren. So enthalten einige PEPs tabellarische Aufstellungen sämtlicher Indikatoren, während andere ergänzend

oder teilweise ausschließlich Diagramme und grafische Vergleichsdarstellungen beinhalten.

2.2. Datenextraktion aus PDF-Dokumenten

Da die PEP-Ecopassport-Umweltdaten ausschließlich in PDF-Dateien veröffentlicht werden, besteht der erste Schritt darin, die Informationen zu extrahieren. Dieser Prozess ist nicht trivial und bildet die Grundlage für die weitere Verarbeitung, Strukturierung und Analyse der Umweltindikatoren.

2.2.1. Das Portable Document Format (PDF)

Das *Portable Document Format (PDF)* ist eines der beliebtesten elektronischen Dokumentenformate. Das PDF-Format ist primär ein *layoutbasiertes Format*. Es wurde entwickelt, um das Erscheinungsbild der Originaldokumente plattform- und anwendungsübergreifend zu bewahren. [LB95] Das Format beschreibt Objekte auf einer niedrigen Strukturebene und legt die *Positionen und Schriftarten der einzelnen Zeichen* fest, aus denen der sichtbare Text zusammengesetzt ist. Zu den beschriebenen Objekten gehören:

- Gruppen von Zeichen (Textobjekte)
- Linien, Kurven und Bilder
- Stilattribute wie Schriftart, Farbe, Strichführung, Füllung und geometrische Formen.

[BK17]

Obwohl PDF die visuelle Darstellung eines Dokuments zuverlässig bewahrt, fehlt den meisten Dateien eine explizite logische Struktur auf höherer Ebene. Die folgenden semantischen Einheiten sind im Format *nicht direkt enthalten* und werden nur durch die oben genannte niedrige Strukturebene zusammengesetzt:

- logische Komponenten wie Wörter, Textzeilen, Absätze, Tabellen oder Abbildungen [CF04]
- Informationen über die *semantischen Rollen* des Textes (z. B. Haupttext, Fußnote oder Bildunterschrift), [BK17]
- eine eindeutige Lese- und Wortreihenfolge, insbesondere bei mehrspaltigen Layouts oder eingebetteten Elementen. [BK17]

Hinzuzufügen ist, dass PDF-Dokumente mit semantischen Informationen *getaggt* werden können. In der Praxis sind diese zusätzlichen Informationen selten gegeben. Die für diese Arbeit relevanten PEP-Ecopassport-PDFs sind alle nicht getaggt. [BK17]

Das Fehlen dieser semantischen Informationen erschwert die Wiederverwendung, Bearbeitung oder Modifikation des Layouts und Inhalts erheblich. [CZ17] Die automatische Extraktion dieser Metadaten und Textinhalte ist daher eine zentrale, aber fehleranfällige Aufgabe, da es keine allgemein verbindlichen Standards für die Strukturierung solcher Informationen in PDF-Dokumenten gibt. [Lip+13]

2.2.2. Herausforderungen bei der automatisierten Extraktion

Die Rekonstruktion des Textflusses und der semantischen Einheiten aus den Positionen einzelner Zeichen ist komplex.

1. Wortidentifikation Die korrekte Bestimmung von Wortgrenzen ist nicht trivial:

- *Abstände*: Die Abstände zwischen Zeichen können innerhalb einer Zeile variieren, sodass keine feste Regel existiert, um Wortgrenzen ausschließlich anhand der Zeichenpositionen zu bestimmen. [BK17]
- *Silbentrennung*: In mehrspaltigen Layouts getrennte Wörter müssen korrekt wieder zusammengeführt werden. [BK17]
- *Ligaturen*: Zeichenkombinationen wie „fl“ oder „fi“ werden im PDF oft als einzelnes Zeichen gespeichert und müssen beim Extrahieren in mehrere Zeichen aufgelöst werden. [Lip+13]
- *Diatkritische Zeichen*: Buchstaben mit Diakritika (z. B. à, á) können als zwei separate Zeichen gespeichert sein und müssen beim Parsing zu einem Zeichen zusammengeführt werden. [BK17]

2. Lesereihenfolge (Reading Order) Die korrekte Lesereihenfolge ist entscheidend für die Verständlichkeit des Textes und der weiterführenden Interpretation. [BK17] In mehrspaltigen Layouts sind Textzeilen im PDF häufig in einer verschränkten Reihenfolge gespeichert. Ohne Korrekturmechanismen führt dies zu unleserlichem, inhaltlich falsch zusammengesetztem Text. [LB95]

3. Absatzgrenzen (Paragraph Boundaries) Die Erkennung von Absatzanfängen und -enden ist besonders schwierig:

- *Unterbrechungen*: Text, der zu einem Absatz gehört, kann durch Formeln, Tabellen oder Abbildungen unterbrochen und später auf derselben Seite fortgesetzt werden.
- *Seitenumbrüche*: Absätze können am Seiten- oder Spaltenende abgeschnitten und auf der folgenden Seite fortgeführt werden, ohne dass dies im PDF strukturell kenntlich gemacht wird.

[BK17]

4. Technologische und Layout-Herausforderungen

- *Überlagerungen (Overlays)*: In grafisch komplexen Dokumenten können Text- und Bildelemente überlappen, etwa wenn Beschriftungen in Abbildungen eingebettet sind. Dies erschwert die korrekte Segmentierung. [CF04]
- *Segmentierungsfehler*: Bei Tabellen, Karten oder Diagrammen kann Text aus unterschiedlichen logischen Einheiten fälschlicherweise in dieselbe Gruppe aggregiert werden. [CF04]
- *Type-3-Fonts*: Manche Zeichen (insbesondere Ligaturen und Sonderzeichen) werden im PDF nicht als Textobjekte, sondern als Vektorgrafiken gespeichert. Solche Elemente sind mit herkömmlicher Textextraktion nicht identifizierbar und erfordern erweiterte, teils OCR- oder ML-basierte Verfahren. [BK17]

2.2.3. Extraktionsansätze

Da diese Probleme weit verbreitet und bekannt sind, ergeben sich mehrere Extraktionsansätze, die das Ziel PDF-Dateien in ein strukturiertes Format zu bringen, um sie anschließend weiter zu analysieren.

Klassische Verfahren (regelbasierte Parser) Ein Beispiel für ein klassisches, regelbasiertes Extraktionsstool ist die Open-Source-Bibliothek *pdfplumber*. Sie ist vollständig in Python implementiert und baut auf der weit verbreiteten Bibliothek *pdfminer.six* auf. Das Werkzeug wurde speziell für die Text- und Tabellenextraktion aus PDF-Dokumenten entwickelt und gilt als eine der benutzerfreundlichsten Lösungen im Python-Ökosystem. [AA25]

pdfplumber konvertiert beim Einlesen einer PDF-Datei deren Inhalt in ein analysierbares Python-Objekt, das sämtliche Seiteninformationen wie Text, Linien, Rechtecke

und Bilder enthält. Jede Seite wird dabei als Sammlung geometrischer Objekte behandelt, deren Koordinaten und Stilattribute (z. B. Schriftart, Farbe, Position) präzise erfasst sind. [YCZ25] Diese Informationen werden über Objektlisten verfügbar gemacht. Für die Erkennung und Extraktion von Tabellen nutzt das Tool einfache visuelle Heuristiken: Standardmäßig werden horizontale und vertikale Linien einer Seite als potentielle Zellgrenzen interpretiert. [YCZ25] Über Parameter wie `table_settings` oder `snap_tolerance` lässt sich die Erkennung an unterschiedliche Layouts anpassen. So können beispielsweise die Toleranzen für Linienabstände verändert werden, um verschobene Spalten oder leere Zellen zu korrigieren. [YCZ25]

Der regelbasierte Ansatz von *pdfplumber* führt bei klar strukturierten, editierbaren PDF-Dokumenten zu guten Ergebnissen. In Studien zur Leistungsbewertung von Extraktions-tools erzielte es in Domänen wie juristischen oder technischen Dokumenten hohe F1-Scores (beispielsweise 0,98 im Bereich *Law*). [AA25]

Die Grenzen des Werkzeugs zeigen sich jedoch vor allem bei komplexen oder unregelmäßig formatierten PDF-Dateien. Insbesondere wissenschaftliche Dokumente mit mathematischen Ausdrücken, eingebetteten Formeln oder verschachtelten Tabellen führen zu deutlichen Leistungseinbußen. In der Kategorie *Scientific* sank der F1-Score auf 0,76, was vor allem auf unvollständige Tabellenerkennung und fehlerhafte Segmentierung zurückzuführen ist. Auch bei Patenten oder Dokumenten mit grafischen Strukturen (z. B. chemischen Formeln oder Bauzeichnungen) stößt der regelbasierte Ansatz an seine Grenzen. [AA25]

Aus Sicht der zuvor beschriebenen Herausforderungen adressiert *pdfplumber* also Probleme auf der Ebene der Zeichen- und Worterkennung weitgehend. Die Wiederherstellung der Lesereihenfolge erfolgt allerdings rein geometrisch, ohne semantisches Verständnis, wodurch Textpassagen aus mehrspaltigen Layouts häufig in falscher Reihenfolge extrahiert werden. Absatzgrenzen, semantische Rollen (z. B. Überschriften, Fließtext, Bildunterschriften) und komplexe Tabellenstrukturen erkennt *pdfplumber* nicht zuverlässig.

Damit steht *pdfplumber* exemplarisch für klassische Extraktions-tools, die ohne maschinelles Lernen oder tiefere Dokumentenverständnis-Modelle arbeiten und deshalb bei komplexen, visuell strukturierten Dokumenten wie den PEP-Ecopassports an ihre methodischen Grenzen stoßen.

Erweiterte Verfahren (z.B. Docling) Ein modernes, KI-gestütztes Gegenbeispiel zu klassischen, regelbasierten Extraktions-tools ist das Open-Source-Toolkit *Docling*. Es wurde mit dem Ziel entwickelt, PDF-Dokumente und andere Formate in eine maschinell verarbeitbare, reich strukturierte Repräsentation zu überführen. Im Gegensatz zu Werkzeugen wie *pdfplumber*, die auf geometrischen Heuristiken basieren, kombiniert

Docling klassische Parsing-Verfahren mit tiefen neuronalen Modellen für Layout- und Strukturerkennung. [Aue+24] Das Toolkit ist vollständig in Python implementiert, modular aufgebaut und kann lokal ohne Cloud-Anbindung ausgeführt werden, was es insbesondere für den Einsatz in sensiblen Datenumgebungen geeignet macht. [Aue+25]

Technisch basiert *Docling* auf einer linearen Verarbeitungs-Pipeline, die mehrere spezialisierte Komponenten kombiniert. Nach dem initialen Parsen durch ein PDF-Backend (z. B. *qpdf* oder *pypdfium*) werden für jede Seite Bitmap-Abbilder erzeugt, auf denen KI-Modelle für Layout- und Strukturerkennung ausgeführt werden. [Aue+24] Das zugrunde liegende Layout-Analysemmodell *DocLayNet* identifiziert auf Basis eines trainierten Objektdetektors verschiedene Seitenelemente und deren Begrenzungsrahmen – etwa Absätze, Überschriften, Listen, Abbildungen oder Tabellen. [Aue+24] Diese visuellen Einheiten werden mit den extrahierten Text-Tokens verknüpft und zu konsistenten Dokumentstrukturen zusammengeführt. Für erkannte Tabellenobjekte kommt anschließend das Vision-Transformer-Modell *TableFormer* zum Einsatz, das die logische Zeilen- und Spaltenstruktur einer Tabelle rekonstruiert und die Zellen semantisch klassifiziert (z. B. Kopf- oder Körperzellen). Für gescannte oder bildbasierte Dokumente steht optional eine OCR-Komponente auf Basis von *EasyOCR* zur Verfügung. [Aue+25]

Das Herzstück von *Docling* bildet das Datenmodell *DoclingDocument*, eine vereinheitlichte interne Repräsentation, die sämtliche Inhalte eines Dokuments (Text, Tabellen, Bilder, Layoutinformationen, Hierarchieebenen und Metadaten) in strukturierter Form abbildet. Nach Abschluss aller Erkennungsschritte werden die Ergebnisse zu einem vollständigen *DoclingDocument* zusammengeführt und können in verschiedenen Formaten exportiert werden, darunter JSON, Markdown und HTML. Im Post-Processing ergänzt ein sprachsensitives Modell weitere Merkmale wie die Korrektur der Lesereihenfolge, die automatische Spracherkennung und die Extraktion zentraler Metadaten (Titel, Autoren, Referenzen). [Aue+24]

Durch diese Architektur adressiert *Docling* mehrere der in Abschnitt 2.3 beschriebenen Extraktionsprobleme, die klassische Tools nur unzureichend lösen können. Es rekonstruiert eine konsistente Lesereihenfolge auch bei mehrspaltigen Layouts, erkennt logische Dokumentstrukturen und kann Tabellen semantisch interpretieren, anstatt sie rein geometrisch zu segmentieren. [Aue+25] Darüber hinaus bietet es eine robuste Metadaten- und Inhaltsklassifizierung, die zwischen Fließtext, Überschriften, Listen, Bildunterschriften und Formeln unterscheidet. Die erzeugten Ausgaben sind reich strukturiert und dienen als Grundlage für weiterführende Analysen oder Datenpipelines, etwa zur Wissensextraktion, semantischen Suche oder automatisierten Inhaltsklassifikation. [Aue+24]

Im Vergleich zu klassischen, regelbasierten Parsern wie *pdfplumber* kombiniert *Docling* geometrische und visuelle Merkmale mit semantischem Verständnis. Dadurch

ermöglicht es eine qualitativ hochwertige, KI-gestützte Dokumentenkonvertierung, die sowohl schnelle als auch stabile Ergebnisse liefert und für komplexe Dokumente wie PEP-Ecopassports einen erheblichen Qualitätsgewinn in der Extraktion bietet. [Aue+24]

Tabelle 2.2 fasst die Extraktionsfähigkeiten der beiden Ansätze zusammen.

Tabelle 2.2: Vergleich der Extraktionsfähigkeiten von *pdfplumber* und *Docling*

Aspekt	<i>pdfplumber</i>	<i>Docling</i>
Zeichen- und Worterkennung	gut – präzise Koordinatenanalyse für editierbare PDFs	gut – kombiniert geometrische und visuelle Merkmale
Lesereihenfolge (Reading Order)	schlecht – keine Korrektur v.a. bei mehrspaltigem Layout	gut – erkennt Spalten und Lesefluss kontext-sensitiv
Absatz- und Textstruktur	teilweise – heuristisch aus Zeilenabständen abgeleitet	gut – erkennt Absätze, Überschriften und Listen
Tabellenerkennung	teilweise – zuverlässig bei klaren Linien, sonst fehlerhaft	gut – rekonstruiert Tabellen semantisch mit KI-Modell
Grafiken und eingebettete Objekte	nicht – keine Analyse oder Erkennung	teilweise – erkennt Abbildungen und Beschriftungen
Metadatenextraktion	nicht – keine Unterstützung	gut – extrahiert Titel, Autoren, Referenzen
Nicht-textuelle Inhalte (OCR)	nicht – nur editierbare PDFs	gut – optionale OCR für gescannte Dokumente
Komplexe Layouts (mehrspaltig, technisch)	schlecht – häufige Fehl-segmentierung	gut – robuste Layout-Analyse durch <i>DocLayoutNet</i>
Semantische Rollen (z. B. Caption, Footnote)	nicht – keine Klassifi-zierung	gut – unterscheidet semantische Dokument-elemente

2.2.4. Zielformat JSON

Die aus PDF-Dokumenten extrahierten Inhalte, beispielsweise in Markdown- oder Textform, bieten trotz ihrer besseren Lesbarkeit keine strukturierte Grundlage für eine automatisierte Datenanalyse. Weder die mit *pdfplumber* gewonnenen Textsegmente noch die von *Docling* erzeugten Markdown-Dateien enthalten eine einheitliche logische Struktur, die eine konsistente Zuordnung von Umweltindikatoren, Materialien oder Metadaten über verschiedene PEPs hinweg erlaubt. Für weiterführende Analysen ist daher ein fest definiertes, maschinenlesbares Zielformat erforderlich, das alle relevanten Inhalte in klar benannten Feldern abbildet.

Das in dieser Arbeit verwendete textbasierte Austauschformat *JavaScript Object Notation (JSON)* hat sich als De-facto-Standard für den strukturierten Datenaustausch etabliert und ist sowohl für Menschen gut lesbar als auch für Maschinen leicht zu verarbeiten [Pez+16]. Obwohl es historisch aus der JavaScript-Syntax hervorgegangen ist, wird JSON sprachunabhängig in nahezu allen modernen Programmiersprachen eingesetzt [Pez+16]. JSON kombiniert einfache Datentypen (*string*, *number*, *boolean*, *null*) mit komplexen Strukturen wie *objects* (Schlüssel–Wert-Paare) und *arrays* (geordnete Listen), wodurch hierarchische, verschachtelte Informationen kompakt und eindeutig dargestellt werden können [Pez+16].

In dieser Arbeit dient JSON als einheitliches Zielformat für die harmonisierte Speicherung der extrahierten PEP-Ecopassport-Daten. Das Format ermöglicht eine konsistente, maschinenlesbare Repräsentation komplexer Strukturen wie Umweltindikatoren, Materialkompositionen und Energieverbrauchsmodellen und lässt sich nahtlos in nachgelagerte Analyseumgebungen (z. B. Python, R oder Datenbanken) integrieren [Pez+16]. Damit bildet JSON die Grundlage für eine standardisierte und reproduzierbare Datenanalyse.

2.2.5. Informationsextraktion ins Zielformat

Die Informationsextraktion (Information Extraction, IE) dient dazu, implizite semantische Strukturen in Texten explizit zu machen. Sie umfasst die Identifikation und Klassifikation semantisch definierter Entitäten, Relationen und Ereignisse in unstrukturierten Texten. [Gri15] Im Kontext dieser Arbeit beschreibt IE den Prozess, aus den mit *pdfplumber* oder *Docling* gewonnenen Textdaten die relevanten PEP-Inhalte (z. B. Umweltindikatoren, Materialien und Metadaten) in eine strukturierte Form zu überführen. Grundsätzlich lassen sich zwei methodische Ansätze unterscheiden: klassische, regel- oder modellbasierte Pipeline-Systeme und moderne, auf großen Sprachmodellen (LLMs) basierende Verfahren.

Regelbasierte Pipeline-Ansätze Traditionelle IE-Systeme folgen einer mehrstufigen Verarbeitungspipeline. Typischerweise werden dabei in aufeinanderfolgenden Schritten benannte Entitäten erkannt, syntaktische Strukturen analysiert, Koreferenzen aufgelöst und schließlich Relationen zwischen Entitäten extrahiert. Solche Systeme nutzen überwiegend probabilistische Sequenzmodelle wie Hidden Markov Models (HMMs), Conditional Random Fields (CRFs) oder Feature-basierte Klassifikatoren. [Gri15] Der Vorteil liegt in der hohen Präzision und der Nachvollziehbarkeit einzelner Verarbeitungsschritte und ihrer Deterministik. Ihre Schwächen zeigen sich jedoch bei komplexen oder stark heterogenen Textformaten, wie sie in PDF-Dokumenten vorkommen: Fehler in einer frühen Pipeline-Stufe können sich fortpflanzen (Fehlersummierung) und die Erstellung regelbasierter Komponenten ist zeit- und ressourcenintensiv, vor allem bei großen Unterschieden in der Struktur des inputs, wie es hier gegeben ist. [Gri15]

LLM-basierte Ansätze Eine neuere Möglichkeit stellen große Sprachmodelle (LLMs) dar, um Informationsextraktion als semantisches Verständnisproblem zu formulieren. LLMs können Textpassagen kontextsensitiv interpretieren und strukturierte Ausgaben, etwa in JSON-Form, direkt generieren. [Nad+24] Sie sind in der Lage, Entitäten, Relationen und numerische Werte inhaltlich zuzuordnen, ohne dass ein manuelles Regelwerk oder ein domänen spezifisch annotiertes Trainingskorpus erforderlich ist. [Mor+25] Zudem ermöglichen sie die Extraktion aus komplexen Layouts, indem sie zuvor durch Tools wie *Docling* generierte Markdown- oder Textsegmente semantisch analysieren. Damit entfällt die sequentielle Verarbeitung einzelner Pipeline-Stufen. [Nad+24] Der Hauptnachteil besteht in der geringeren Deterministik und möglichen Halluzinationen (falsch generierten Werten), die durch präzises Prompt-Design und Validierungsschritte minimiert, aber nicht vollständig ausgeschlossen werden können. [Mor+25]

Vergleich Während klassische Pipelines durch ihre strukturelle Klarheit und Nachvollziehbarkeit überzeugen, stoßen sie bei unstrukturierten PDF-Dokumenten mit komplexem Layout schnell an Grenzen. LLM-basierte Verfahren bieten eine deutlich höhere Flexibilität und können heterogene Dokumentinhalte semantisch erfassen und strukturieren, erfordern jedoch eine sorgfältige Kontrolle der Modellantworten. Zudem führen sie zu einer geringeren Transparenz. Für die vorliegende Arbeit ist die Kombination beider Ansätze zentral: *Docling* dient der robusten Layout- und Strukturerkennung, während ein LLM die semantische Extraktion und Überführung in das harmonisierte JSON-Format übernimmt.

2.3. Statistische Grundlagen

Die in dieser Arbeit verwendeten statistischen Verfahren bilden die methodische Grundlage zur Analyse und Modellierung der aus PEP Ecopassports extrahierten Daten. Dazu werden zunächst *deskriptive und explorative* Verfahren eingesetzt, um Strukturen, Streuungen und Ausreißer in den Daten sichtbar zu machen. Die *Hauptkomponentenanalyse* wird dafür verwendet die wichtigsten Merkmale zu identifizieren. Darauf aufbauend wird die *lineare Regression* als einfaches, interpretierbares Modell genutzt, um heuristische Beziehungen zwischen Einflussgrößen und den resultierenden Umweltindikatoren zu identifizieren. Diese Kombination ermöglicht eine robuste, nachvollziehbare und datengetriebene Einschätzung ökologischer Wirkzusammenhänge im Datensatz.

2.3.1. Deskriptive und explorative Statistik

Die deskriptive und explorative Statistik bilden die Grundlage der quantitativen Datenanalyse. Beide dienen der Zusammenfassung, Beschreibung und Visualisierung von Datensätzen, um zentrale Merkmale einer Verteilung zu charakterisieren und potentielle Muster oder Auffälligkeiten zu erkennen. [FM09] Der Schwerpunkt liegt nicht auf Hypothesentests, sondern auf dem Verständnis der vorhandenen Daten. [Dim+19] In dieser Arbeit werden die Verfahren auf aus PEP Ecopassports extrahierte Kennwerte angewendet.

Deskriptive Statistik Die deskriptive Statistik umfasst numerische und grafische Verfahren zur Beschreibung (i) der *zentralen Tendenz* und (ii) der *Streuung* von Daten. [FM09] Ziel ist die Verdichtung großer Datenmengen auf wenige aussagekräftige Kennzahlen. Zu den typischen Lagemaßen gehören **Mittelwert**, **Median** und **Modalwert**. Der Mittelwert beschreibt die durchschnittliche Ausprägung, während der Median die geordnete Verteilung in zwei gleich große Hälften teilt. Der Median gilt als *robustes Lagemaß*, da er – im Gegensatz zum Mittelwert – wenig durch Ausreißer beeinflusst wird. [Dim+19] Für die Streuung werden Standardabweichung, Spannweite und insbesondere der **Interquartilsabstand (IQR)** verwendet. Der IQR beschreibt die mittleren 50 % der Daten und ist ein robustes Maß, das gegenüber Extremwerten stabil bleibt. Für ordinale Merkmale ist der Median das geeignete Lagemaß; der IQR, ergänzt um Minimum und Maximum, quantifiziert die Streuung. [FM09]

Verteilungsformen und Schiefe Ein zentrales Merkmal numerischer Daten ist die Form ihrer Verteilung. In symmetrischen Verteilungen fallen Mittelwert, Median und Modus zusammen. Bei *rechtsschiefen* Verteilungen liegen einzelne hohe Werte weit über dem zentralen Bereich, sodass der Mittelwert größer als der Median ist; bei *links-schiefen* Verteilungen gilt das umgekehrte Muster. [KSY18] Schiefe beeinflusst die Interpretation von Lage- und Streumaßen und motiviert den Einsatz robuster Kennwerte wie Median und IQR. In der explorativen Praxis werden zudem log-transformierte Werte betrachtet, um stark asymmetrische Verteilungen zu symmetrisieren und visuell leichter interpretierbar zu machen.[MJ10]

Log-Transformation und methodische Alternativen Wenn eine Verteilung deutlich von der Normalität abweicht, bestehen drei grundlegende Optionen: (i) eine regelbasierte *Ausreißerprüfung* mit dokumentierter Entfernung, [FM09] (ii) eine *Log-Transformation* zur Annäherung an Symmetrie und zur besseren Vergleichbarkeit in Visualisierungen [MJ10], oder (iii) die Anwendung *nicht-parametrischer* Verfahren, die keine Normalverteilung voraussetzen [FM09]. Parametrische Methoden setzen in der Regel Normalität voraus; die Entscheidung erfolgt im Rahmen der explorativen Analyse.

2.3.2. Explorative Datenanalyse (EDA) und Visualisierungen

Die explorative Datenanalyse ergänzt die deskriptive Statistik durch strukturentdeckende Verfahren. Sie dient der visuellen Erkundung und Bewertung von Mustern, Ausreißern oder Zusammenhängen zwischen Variablen, ohne dass zuvor Hypothesen formuliert werden müssen. Zentrale Visualisierungen sind Histogramme und Boxplots. [KSY18]

Histogramme Histogramme stellen Häufigkeitsverteilungen kontinuierlicher Merkmale über Klassen dar. Sie erlauben Rückschlüsse auf Symmetrie, Schiefe und Mehrgipfligkeit und dienen zur Prüfung von Verteilungsannahmen. [MJ10]

Boxplots Boxplots visualisieren Median (Q_2), Quartile (Q_1, Q_3) und potenzielle Ausreißer. Die sogenannten *Whisker* markieren üblicherweise den Bereich bis zum 1,5-fachen Interquartilsabstand; Werte außerhalb gelten als potenzielle Ausreißer. [MJ10] Diese Darstellungsform ermöglicht die Beurteilung von Streuung, Schiefe und Extremwerten und eignet sich für den Vergleich mehrerer Merkmale. [KSY18]

2.3.3. Automatisierung, Reproduzierbarkeit und Datenqualität

Eine *skriptbasierte*, reproduzierbare Umsetzung (z. B. in R) gewährleistet konsistente Analysen und Nachvollziehbarkeit aller Berechnungsschritte. Im Rahmen der explorativen Datenanalyse unterstützt sie die Datenqualitätsbewertung durch: (i) Erkennung von Datenfehlern, Ausreißern und fehlenden Werten, (ii) Überwachung einfacher Profilierungsmaße (z. B. Anteil fehlender oder eindeutiger Werte) zur Bewertung von Vollständigkeit und Eindeutigkeit, sowie (iii) Aufdeckung semantischer Inkonsistenzen und Formatabweichungen. Diese systematische, nachvollziehbare Vorgehensweise bildet die Grundlage für die Qualitätssicherung entlang der Pipeline $PDF \rightarrow JSON$ und schafft Transparenz über den Zustand der analysierten Daten.

2.3.4. Lineare Regression

Die lineare Regression dient in dieser Arbeit als methodische Grundlage zur Modellierung der Umweltwirkungen von Produkten auf Basis quantitativer Einflussgrößen. Ziel ist es, Zusammenhänge zwischen erklärenden Variablen wie *Produktgewicht*, *Materialzusammensetzung*, *Stromverbrauch* und *verwendetem Energiemix* und den resultierenden *Umweltindikatoren* zu quantifizieren und zur Abschätzung unbekannter Werte nutzbar zu machen.

Modellstruktur Das Regressionsmodell beschreibt den linearen Zusammenhang zwischen einer abhängigen Variable y (z. B. einem Umweltindikator) und mehreren unabhängigen Variablen x_1, x_2, \dots, x_k (z. B. Gewicht, Stromverbrauch, Materialanteile):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon.$$

Dabei ist β_0 der Achsenabschnitt, β_i die Regressionskoeffizienten der jeweiligen Einflussgrößen und ε ein zufälliger Fehlerterm, der unerklärte Varianzanteile abbildet. Die Koeffizienten β_i quantifizieren die Richtung und Stärke des Einflusses einzelner Variablen auf den Zielindikator. [MPV22]

Zentrale Annahmen Für die lineare Regression gelten folgende Grundannahmen:

- **Linearität:** Die Beziehung zwischen abhängiger und unabhängigen Variablen ist näherungsweise linear.
- **Erwartungswert Null:** Die Fehlerterme haben einen Erwartungswert von null, $E(\varepsilon) = 0$.

- **Homoskedastizität:** Die Varianz der Fehler ist konstant und unabhängig von den Regressoren.
- **Unabhängigkeit:** Die Fehlerterme sind voneinander unkorreliert.

[Su2012] Diese Annahmen sichern die Unverzerrtheit und Effizienz der Parameterschätzungen. Für explorative Anwendungen, wie sie in dieser Arbeit verfolgt werden, steht jedoch die Strukturentdeckung im Vordergrund. Moderate Abweichungen von den Idealannahmen sind daher akzeptabel, sofern sie dokumentiert werden.

Modellinterpretation Die Koeffizienten β_i geben an, wie stark sich der Zielindikator y im Mittel verändert, wenn sich die Einflussgröße x_i um eine Einheit ändert, während alle anderen Variablen konstant bleiben. Das *Bestimmtheitsmaß* R^2 beschreibt den Anteil der Varianz des Zielindikators, der durch die erklärenden Variablen erklärt wird, und dient als zentrales Maß der Modellgüte. [MPV22] Zur Bewertung der Modellangemessenheit werden Residuenanalysen eingesetzt, um Abweichungen von Linearität oder Homoskedastizität sichtbar zu machen.

Anwendungsrahmen In dieser Arbeit wird die multiple lineare Regression verwendet, um Heuristiken zur Abschätzung der Umweltwirkungen von Elektro- und Elektronikprodukten zu entwickeln. Das Modell dient der quantitativen Erfassung von Zusammenhängen zwischen Produktmerkmalen und Umweltindikatoren und daraus schließend der möglichst präzisen Prognose der Umweltindikatoren anhand der Input-Variablen. Damit bildet die lineare Regression eine nachvollziehbare, statistisch fundierte Basis für die Entwicklung eines vereinfachten Bewertungsmodells innerhalb der PEP-Datenanalyse.

3

Pipeline und Datenbasis (Methodik)

Dieses Kapitel beschreibt den Aufbau der Datenpipeline, die Extraktion der relevanten Variablen aus PEP-Ecopassport-Dokumenten sowie die Struktur und Aufbereitung der resultierenden Datenbasis.

3.1. Überblick der Pipeline

Ziel der entwickelten Pipeline ist die automatisierte Extraktion strukturierter Daten aus PEP-Ecopassport-Dokumenten im PDF-Format. Die PEPs bilden die zentrale Quelle für produktbezogene Umweltinformationen, enthalten jedoch uneinheitlich formatierte Tabellen und Textblöcke, die eine direkte Auswertung erschweren.

Die Pipeline wandelt die heterogenen PDF-Dokumente in ein einheitliches, maschinenlesbares Datenformat um. Als Input dienen die PEP-PDFs, als Output entsteht eine strukturierte CSV-Datei, die sämtliche relevanten Variablen zu Produkt, Materialien, Energieverbrauch und Umweltindikatoren enthält. Der Prozess umfasst mehrere aufeinanderfolgende Schritte:

- **Erfassung und Download:** Automatisierte Recherche und Speicherung der verfügbaren PEP-Dokumente aus der öffentlichen PEP-Datenbank.
- **Extraktion:** Umwandlung der PDF-Dateien in Rohtext und Tabelleninhalte mittels Dokumentenparser; Layout- und Tabellenstrukturen werden erkannt.
- **Interpretation:** Zuordnung der erkannten Inhalte zu definierten Variablen mithilfe regelbasierter und LLM-gestützter Methoden.

- **Normalisierung:** Harmonisierung von Einheiten, Materialnamen und Energiemodellen zur Sicherstellung der Vergleichbarkeit.
- **Export:** Zusammenführung aller Informationen in eine flache, analysierbare CSV-Datei als Grundlage der nachfolgenden statistischen Auswertung.

Abbildung 3.1 zeigt den groben schematischen Aufbau des Gesamtprozesses von der Rohdatenerfassung bis zur strukturierten Datenbasis. Der Teil der Normalisierung und Datenbereinigung wird in 3.1.3 detaillierter dargestellt.

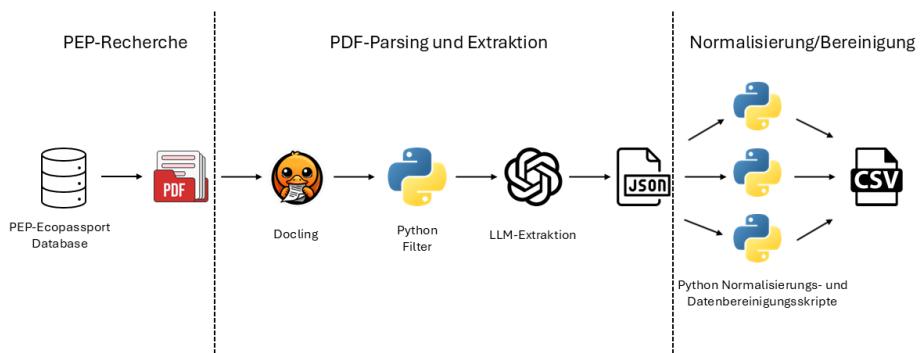


Abbildung 3.1: Schematischer Aufbau der Pipeline: von der PEP-Erfassung bis zur strukturierten Datenbasis.

3.1.1. Datenerhebung und PEP-Recherche

Ziel der Datenerhebung war die Identifikation und Extraktion von PEP-Ecopassport-Dokumenten, die sich auf Geräte der Gebäudeautomatisierung oder IoT Komponenten beziehen. Die PEP-Datenbank bildet dabei die Quelle der Untersuchung. Für jedes gefundene Dokument wurden Produktinformationen, Material- und Energiedaten sowie Metadaten in strukturierter Form erfasst.

Zielsetzung Im ersten Schritt der Arbeit sollte eine Datenbasis zu IoT-Produkten in der PEP-Datenbank aufgebaut werden. Dazu wurden die Dokumente automatisiert recherchiert, die zugehörigen PDFs heruntergeladen, analysiert und nach Relevanz für den Smart-Home- bzw. IoT-Bereich klassifiziert. Das Ergebnis wurde in CSV-Formaten gespeichert und diente als Grundlage für die weitergehende Analyse.

Keyword-basierte Suche Zur Ermittlung der verfügbaren PEP-Dokumente wurde zunächst die Suchfunktion der PEP-Datenbank analysiert. Über die Browser-Entwicklertools konnten die zugrunde liegenden Netzwerkanfragen identifiziert werden. Insbesondere der Endpoint `/xhr/searchPep` lieferte HTML-Snippets mit Produktnamen und Links. Diese wurden mithilfe von JavaScript iterativ abgefragt, geparst und in einer CSV-Datei gespeichert. Die Implementierung verwendete `fetch`-Requests mit kopierten FormData-Parametern und CSRF-Tokens. Für die Paginierung wurde die Gesamtanzahl der Treffer aus dem Response genutzt, um alle Ergebnisse in Schleifen abzurufen. Um gezielt IoT-nahe Produkte zu erfassen, wurden die Abfragen mit spezifischen Suchbegriffen in den FormData-Parametern erweitert (z. B. `controller`, `sensor`, `gateway`, `wifi`, `knx`, `zigbee`, `cloud`). Dadurch konnten 184 PEP-Einträge, welche potentiell Gebäudeautomatisierungsgeräte beschreiben identifiziert und als CSV exportiert werden. Anschließend erfolgte eine manuelle Prüfung und Klassifikation der Ergebnisse, da die Suchbegriffe im Produkttitel nicht ausnahmslos auf Komponenten der Gebäudeautomation schließen können. Zudem wendet die Suchfunktion der PEP-Plattform nicht alle Filter korrekt an und es treten Überschneidungen zwischen den Seiten auf.

Klassifikation der Produkte Die ermittelten Produkte wurden in einer Excel-Datei manuell kategorisiert. Neben dem Produktnamen und der URL enthielt die Datei eine Spalte *IoT-Einschätzung* mit vordefinierten Auswahloptionen. Farbcodierungen erleichterten die visuelle Trennung zwischen den Gruppen. Zur Validierung wurden zusätzlich die jeweiligen PEP-PDFs gelesen und, falls erforderlich, weitere Produktinformationen aus Herstellerportalen herangezogen. So konnten 102 Smart-Home-relevante Geräte eindeutig als IoT oder IoT-nah eingestuft werden.

Kategorisierung Die Zuordnung erfolgte nach funktionalen Kriterien:

- **Gebäudeautomatisierung:** Geräte mit Konnektivität (z. B. ZigBee, WiFi, KNX) oder Cloud-Anbindung, wie Gateways, smarte Sensoren oder Steuerungen.
- **eher ja:** Komponenten mit indirekter IoT-Relevanz, etwa Erweiterungsmodulen. (Teilweise aussortiert)
- **eher nein:** Elektronik mit digitaler Funktion, jedoch ohne Vernetzung. (Aussortiert)
- **kein IoT:** Produkte ohne Kommunikationsfähigkeit (z. B. Kabel, Trafos, LED-Panels). (Aussortiert)

Manuelle Ergänzungen Zusätzlich zur halbautomatisierten Suche wurden IoT-relevante Unternehmen gezielt identifiziert (z.B. ABB, Siemens, Schneider Electric, Legrand,

Somfy, Daikin, Bosch, Honeywell). Deren PEP-Dokumente wurden manuell durchsucht und ergänzt. Dadurch erweiterte sich der Datensatz um weitere 145 PEPs. Insgesamt umfasst die erstellte Datenbasis 247 PEP-Dokumente, die anschließend in der Parsing-Pipeline verarbeitet und vereinheitlicht wurden.

3.1.2. PDF-Parsing und Extraktion

Die Extraktion strukturierter Daten aus PEP-PDFs stellte den technisch anspruchsvollsten Teil der Arbeit dar. Ziel war es, aus den heterogenen Dokumenten eine konsistente, maschinenlesbare Repräsentation der Umweltindikatoren, Materialanteile und Metadaten zu erzeugen. Die finale Lösung kombiniert eine robuste Layoutanalyse mit Docling und eine LLM-basierte, schemagesteuerte Inhaltsinterpretation.

Entwicklung und Vorläufer Zu Beginn wurde eine auf `pdfplumber` basierende Pipeline eingesetzt, die mit Regex- und Textheuristiken Tabellen und Materiallisten erkannte. Obwohl dieser Ansatz für einzelne PDFs funktionierte, erwies sich die Übertragbarkeit als unzureichend. Grund dafür waren stark varierende Layouts, Tabellen als Rasterbilder und uneinheitliche Bezeichnungs- und Einheitenformate. Bereits kleine Änderungen im Tabellenkopf führten zu fehlerhaften Zuordnungen von Indikatornamen oder Spalten. Die Vielzahl individueller Ausnahmen entwickelte sich zur sogenannten „*Exception-Hölle*“, in der jedes zusätzliche Sonderlayout neue Regressionen verursachte. Eine vollständige Generalisierung des `pdfplumber`-Parsers war im Rahmen der Arbeit nicht realistisch umsetzbar.

Diese Limitierungen führten zur Entwicklung einer neuen, modularen Pipeline, die auf dem Open-Source-Framework *Docling* von IBM basiert. Docling erlaubt die strukturierte Segmentierung von PDF-Inhalten in Absätze, Tabellen, Listen und Bilder und exportiert diese in Markdown oder JSON. Dadurch konnte die textuelle Logik vom Layout entkoppelt und die Zuverlässigkeit der Downstream-Verarbeitung deutlich verbessert werden.

Methodisches Konzept Die Pipeline trennt klar zwischen Layoutanalyse und Inhaltsinterpretation:

- **Docling-Konvertierung:** PDF-Dateien werden in eine Markdown-Struktur überführt. OCR und Bildbeschreibung sind deaktiviert, um Laufzeit und Speicherverbrauch zu reduzieren. Tabellen- und Abschnittsgrenzen bleiben erhalten.
- **LLM-basierte Extraktion:** Der konvertierte Text wird in Abschnitten an ein Sprachmodell übergeben, das definierte Variablen extrahiert und im JSON-Format zurückgibt.

Die Promptstruktur erzwingt strikte Datentypen und klare Feldbezeichnungen.

- **Merging:** Ergebnisse der Indikatorenextraktion und der Metadatenextraktion werden über den PDF-Dateinamen zusammengeführt; parserbasierte Werte haben Vorrang vor LLM-Schätzungen.

Wahl des Modells und der Schnittstelle Für die semantische Extraktion wurde *GPT-5* verwendet, angesprochen über die *Responses-API*. Diese neue Schnittstelle unterstützt native strukturierte Ausgaben und optional eine Schema-Validierung. Der Aufruf erfolgt im `response_format=json_object-` oder `json_schema`-Modus, wodurch fehlerhafte JSON-Formate praktisch ausgeschlossen sind. Gegenüber dem früher verwendeten *gpt-3.5-turbo* zeigte sich *GPT-5* deutlich stabiler in der Erkennung von numerischen Werten, Einheiten und Modulzuordnungen (A1–A3, A4, A5, B*, C*, D). Zudem reduziert sich der Post-Processing-Aufwand erheblich, da keine nachträgliche JSON-Reparatur erforderlich ist.

Die Kombination aus Docling und *GPT-5* führte somit zu einem reproduzierbaren und skalierbaren Verfahren, das auch bei komplexen Layouts konsistente Ergebnisse liefert.

Extraktionslogik

- **Indikatoren:** Matching über Name und Einheit auf Basis der EF 3.1-Labels (z. B. kg CO₂ eq, kg Sb eq, MJ). Header-Kontext wird mitgeparst, B-Phasen werden nicht aggregiert.
- **Plausibilität:** Flatline-Filter (identische Modulwerte), Prüfungen von *Total* vs. Modulsumme, Toleranz für negative D-Werte.
- **Materialien:** Parser-Ergebnisse aus Docling-Tabellen sind führend. Das LLM ergänzt nur, überschreibt jedoch keine verlässlichen Prozentwerte.
- **Normalisierung:** Zahlen ohne Einheitenzeichen; Einheiten ausschließlich im Feld `unit`. Keine heuristische Imputation oder Raten.

Robustheit und Grenzen Die neue Pipeline konnte die Anzahl fehlerhafter oder unvollständiger Einträge deutlich reduzieren. Fallback-Mechanismen greifen bei fehlerhaften Tabellen automatisch auf den Fließtext zurück, wodurch auch reine Text-PEPs ausgewertet werden können. Für PDFs mit reinen Rastertabellen bleibt jedoch eine Einschränkung bestehen, da ohne OCR keine Inhaltsextraktion möglich ist. Trotz dieser Grenzen erlaubt der Ansatz eine hohe Reproduzierbarkeit und Transparenz: identische Eingaben erzeugen identische Ausgaben. Die Pipeline priorisiert Nachvollziehbarkeit und strukturelle Konsistenz gegenüber maximaler Vollständigkeit.

3.1.3. Normalisierung und Datenbereinigung

Mapping von Materialnamen, Energiemodellen, Einheitenharmonisierung, Umgang mit fehlenden Werten, Ausreißerprüfung.

3.2. Datenbasis

Quantitative Übersicht über die gewonnene Datenbasis: Anzahl der Dokumente, Variablen, Struktur, Coverage.

3.2.1. Variablenstruktur

Beschreibung der Hauptgruppen (Input, Output, Metadaten).

3.2.2. Qualitätssicherung

Verfahren zur Plausibilitätsprüfung (z. B. visuelle Kontrolle, Stichproben, Filter).

4

Analyse der erarbeiteten Daten

Auf Grundlage der in ?? beschriebenen Datenbasis wird die Analyse durchgeführt.

4.1. Deskriptive Annäherung an die PEP-Daten

Wie im Kapitel ?? angesprochen ist es sinnvoll die erarbeiteten Daten deskriptiv anzuschauen, um unter anderem auch die Datenqualität zu gewährleisten.

4.1.1. Vollständigkeit der Werte

Zur Bewertung der Datenvollständigkeit wurde der Anteil fehlender Werte pro Variable berechnet und in einem Balkendiagramm dargestellt (Abb. 4.1). Die Missingness umfasst sowohl Nullwerte aus der Datenpipeline als auch Indikatoren, die in den PEPs nicht berichtet werden.

Die Analyse zeigt deutliche Unterschiede zwischen den Indikatoren: Für *Wasserknappheit* fehlen rund 78 % der Werte, während mehrere weitere Indikatoren wie *Eutrophierung marines Gewässer*, *Klimawandel (fossil, total)* und *Eutrophierung terrestrisch* Fehlstände von etwa 30 % aufweisen. Andere Kernfelder, darunter *Gesamtgewicht* und *Stromverbrauch*, sind vergleichsweise vollständig.

Die fehlenden Werte werden in der weiteren Analyse berücksichtigt. Der Indikator *Wasserknappheit* wird aufgrund der hohen Ausfallrate von der Auswertung ausgeschlossen, da keine belastbaren statistischen Aussagen getroffen werden können.

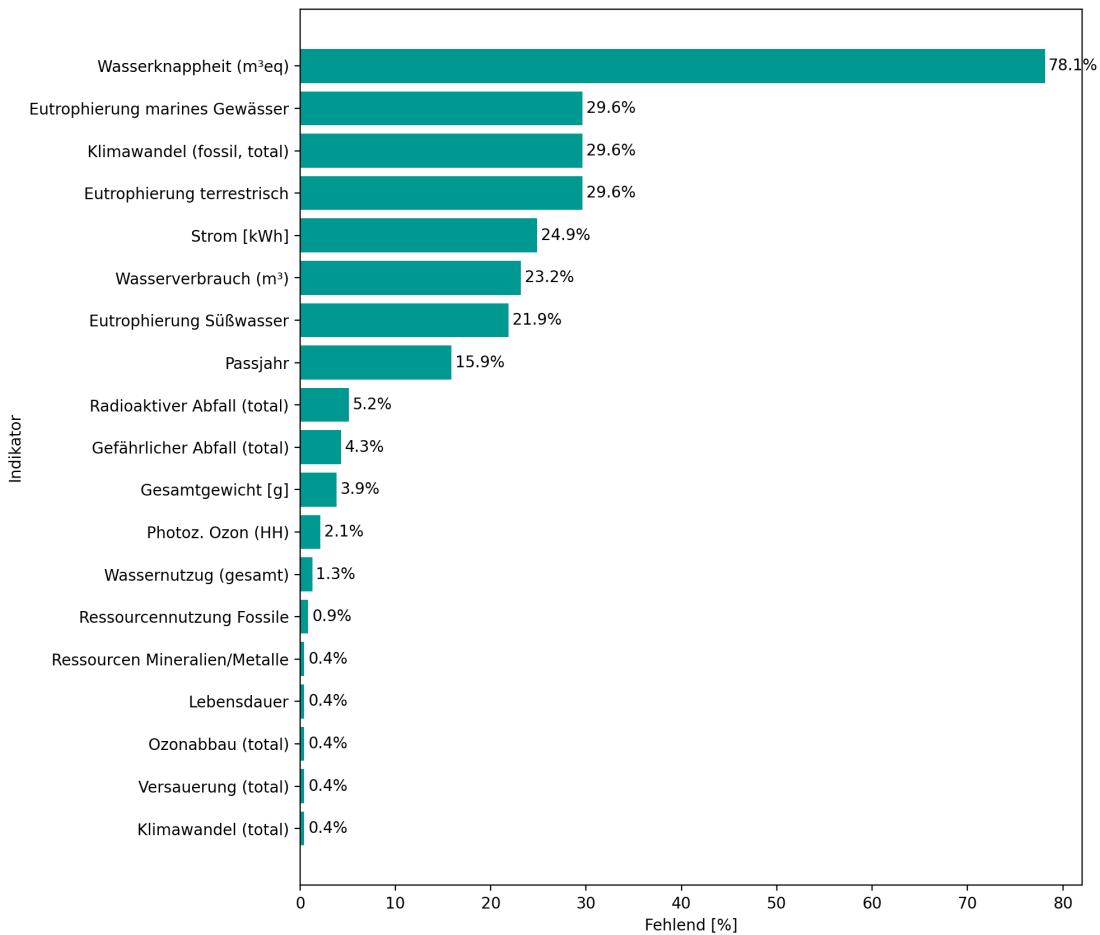


Abbildung 4.1: Anteil fehlender Werte pro Umweltindikator (%; N = 233 Produkte).

Zusätzlich wurden die Erscheinungsjahre der PEP-Ecopassports untersucht. Die Veröffentlichungen reichen von 2020 bis 2025 und verteilen sich wie in Abb. 4.2 dargestellt. Der deutliche Anstieg ab 2022 zeigt die zunehmende Etablierung des Formats und eine stärkere Datenverfügbarkeit in den letzten Jahren.

4.1.2. Überblick der *Input*-Variablen

Für die Input-Variablen werden robuste Kennzahlen (**Median**, **IQR**) berichtet und durch den **Mittelwert** ergänzt, um die Wirkung der Schiefe (v. a. Rechtsschiefe) zu verdeutlichen. Ausreißer werden nicht entfernt; ihre Einflüsse spiegeln sich im Mittelwert wider.

Die in Tab. 4.1 dargestellten Basisvariablen zeigen deutlich **rechtsschiefe Verteilungen** mit großen Interquartilsabständen (IQR). Beim *Gesamtgewicht* reicht die Spannweite von 0.04 kg bis über 13 000 kg, was die starke Heterogenität der betrachteten Produkte verdeutlicht. Das kleinste Produkt ist ein leichtes elektronisches Gerät, ein

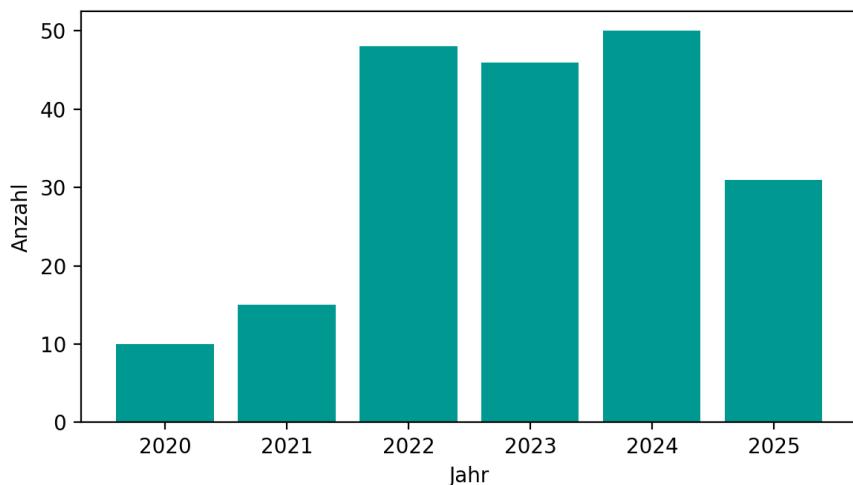


Abbildung 4.2: Erscheinungsjahre der analysierten PEP-Dokumente ($N = 233$ Produkte).

Variable	Einheit	Min	Median	Max	IQR	Mittelwert
Gesamtgewicht	kg	0.0395	2.178	13022.6	125.210	278.023
Stromverbrauch	kWh	0.026	326.511	8203569.5	86147.1	228061.654

Tabelle 4.1: Robuste deskriptive Kennzahlen der Basisvariablen.

Connected dimmer mit Bluetooth interface ([PEP-Link](#)), während das größte Produkt, ein *Flüssigkeitskühler mit drehzahlgeregelter Schraubenverdichter und Greenspeed™-Technologie* ([PEP-Link](#)), mehr als 13 t erreicht. Der Mittelwert liegt mit 278 kg weit über dem Median (2.18 kg), was die ausgeprägte Rechtsschiefe bestätigt.

Auch der *Stromverbrauch* weist eine extreme Streuung auf (ca. 86147 kWh), mit Werten zwischen 0.026 kWh und über 8.2e6 kWh. Damit ist das kleinste Produkt nahezu stromlos im Betrieb, während das größte Produkt eine mehrjährige oder großtechnische Nutzung abbildet. Der Mittelwert (228000 kWh) übersteigt den Median (327 kWh) um mehrere Größenordnungen, was die starke Rechtsverschiebung der Verteilung verdeutlicht.

Die Zusammenhänge zwischen Stromverbrauch und Umweltauswirkungen hängen maßgeblich von der Art der Stromerzeugung ab. Da sich die Strommixe regional unterscheiden, variieren auch die resultierenden Emissionen je nach Herkunftsland des Energiebezugs.

Im Datensatz zeigt sich, dass der Großteil der verwendeten Energiemodelle auf allgemeine europäische Strommixe (EU27) und Frankreich entfällt. Besonders in den Phasen Nutzung und End-of-Life ist der Anteil europäischer Modelle deutlich höher.

Dies liegt vermutlich daran, dass die Produkte häufig europaweit vertrieben und verwendet werden. Daher ist es schwierig, den tatsächlichen Energiebezug eines spezifischen Landes realistisch abzubilden, weshalb in der Regel ein repräsentativer europäischer Durchschnitt angenommen wird.

Der hohe Anteil von Frankreich ist auf eine große Anzahl an PEP-Dokumenten aus Frankreich zurückzuführen. Von dort stammt die Association P.E.P und das Format ist dort am meisten etabliert.

Auch in der Herstellungsphase dominiert ein europäischer Energiemix, ergänzt durch einzelne Modelle aus Deutschland und China, was auf internationale Produktionsketten hinweist. Insgesamt verdeutlicht die Verteilung, dass die meisten PEP-Deklarationen von europäischen Strommixen ausgehen, wodurch die berechneten Umweltauswirkungen tendenziell niedrigere fossile Anteile aufweisen, als es bei stärker kohleabhängigen Regionen (z. B. China) der Fall wäre.

Aufgrund der europäischen Prägung des Datensatzes ist die Aussagekraft der anschließenden Regression für Produkte außerhalb Europas eingeschränkt. Entsprechende Auswertungen werden mit erhöhter Unsicherheit und geringerer Datenqualität verbunden sein.

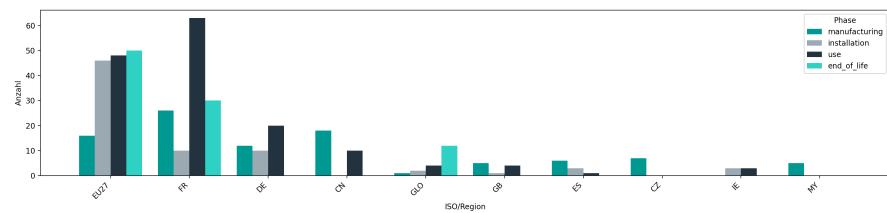


Abbildung 4.3: Verteilung der verwendeten Energiemodelle (ISO-Regionen) über die Lebenszyklusphasen

Eine weitere Variable, die die Umweltindikatoren stark beeinflussen, und damit in der zu entwickelnden Heuristik eine Rolle spielen muss, ist die Zusammensetzung des Produkts aus den verschiedenen Materialien. In der Tabelle 4.2 wird aufgeführt, aus welchen Materialien das durchschnittliche PEP-Produkt aus dem Datensatz besteht (Mittelwert). N gibt die Anzahl der Produkte an, in welchen das aufgeführte Material auftaucht. Wie in den meisten PEP-Dokumenten werden die modularen Materialien in die Gruppen *Metalle*, *Plastik* und *andere* gegliedert.

Die in Tabelle 4.2 dargestellten Materialanteile zeigen eine sehr heterogene Zusammensetzung der untersuchten Produkte. Mit durchschnittlich rund 26 % ist *Stahl* das mengenmäßig dominierende Einzelmaterial, gefolgt von *Papier* (15.7 %), welches vor allem für Verpackungen verwendet wird, und *Polycarbonat (PC)* (8.2 %). Die hohe Varianz der Währung Metalle in nahezu allen PEPs vertreten sind, treten bestimmte Kunststoffe und Spezialmaterialien (z. B. PMMA, PBT, PPS) nur in wenigen Fällen auf. Die Kategorie *Andere* enthält zahlreiche kleinvolumige Komponenten, deren summier-

Tabelle 4.2: Durchschnittliche Materialanteile nach Hauptkategorien (Mittelwert in %).

(a) Metalle			(b) Kunststoffe			(c) Andere		
Material	Mittelwert	n	Material	Mittelwert	n	Material	Mittelwert	n
Stahl	26.46	199	Polycarbonat (PC)	8.23	113	Papier	15.73	197
Aluminium	6.13	140	ABS	2.84	105	Elektronik	3.69	102
Kupfer	5.41	161	Polyamid (PA)	2.37	118	Holz	3.12	81
Messing	0.86	82	PVC	2.08	86	Glas	2.90	67
Zamak	0.45	15	PS	1.13	62	PCBA	1.79	24
Nickel	0.10	12	PP	0.80	75	PCB	1.27	49
Zinn	0.06	20	Gummi	0.69	70	Kabel	0.35	38
Zink	0.05	9	PMMA	0.66	19	Kältemittel	0.35	53
Bronze	0.01	5	Epoxidharz	0.60	42	Ferrit	0.28	38
Neodym	0.01	5	Polyesterharz	0.57	23	Elektromotoren	0.27	9
Hartlot	0.00	7	PE	0.44	68	Lack / Farbe	0.15	39
			PU	0.40	40	Tinte	0.08	15
			PBT	0.23	17	Silizium	0.08	9
			PET	0.13	25	Batterie	0.08	10
			POM	0.09	16	Thionylchlorid	0.08	5
			TBBPA	0.07	9	Öl	0.07	8
			HIPS	0.06	4	Mineralwolle	0.06	13
			Silikon	0.04	6	Bitumen	0.04	7
			EPDM	0.02	5	Titandioxid	0.04	14
			PPS	0.02	5	Quarz	0.02	7
			Sonstige	0.03	–	Flussmittel	0.02	6
						Filz	0.01	11
						Aluminiumoxid	0.01	5
						Haftkleber	0.01	4
						Sonstige	0.48	–

ter Anteil jedoch nicht vernachlässigbar ist. Insgesamt spiegelt sich in der Verteilung die Diversität der erfassten Produktgruppen wider.

4.1.3. Überblick der Umweltindikatoren

Die Umweltindikatoren bilden die Output-Variablen, auf deren Basis später die Heuristik entwickelt wird. Eine deskriptive Betrachtung verdeutlicht bereits die Verteilungsstruktur der Daten.

Wie Tabelle 4.3 zeigt, weisen alle Umweltindikatoren deutlich **rechtsschiefe Verteilungen** auf: Der Mittelwert liegt bei allen Größen um ein Vielfaches über dem Median. Besonders ausgeprägt ist dies bei *Climate change (total)*, *Resource use (fossils)*, *Water use* und *Hazardous waste disposed*, bei denen einzelne Extremwerte die Verteilungen dominieren. Dagegen zeigen *Ozone depletion* und *Resource use (minerals/metals)* geringere Abstände zwischen Mittelwert und Median, bleiben aber ebenfalls schief. Insgesamt bestätigt sich eine stark heterogene Datenbasis mit wenigen Produkten, die sehr hohe Umweltwirkungen aufweisen.

Indikator (total)	Min	Median	Max	IQR	Mean	Einheit
Acidification	0.000,017	0.4295	3650	10.31	110.78	kg SO ₂
Climate change (total)	0.0031	86.75	1,040,000	1979.43	22,740.20	kg CO ₂
Eutrophication (freshwater)	0.000,001	0.0266	236	0.314	2.624	kg P e
Hazardous waste disposed	0	39.3	489,000	596.69	6438.71	kg
Ozone depletion	0	0.000,007	0.192	0.000,286	0.003,23	kg CF ₂
Photochemical ozone formation (HH)	0.000,002	0.181	1410	3.126	40.60	kg C ₂ H ₆
Resource use (fossils)	0.0326	1620	106,000,000	95,076	1,583,968.96	MJ
Resource use (minerals/metals)	0.000,001	0.003,92	5.87	0.0495	0.2279	kg Sb
Radioactive waste disposed	0	0.0656	3260	0.5381	22.56	kg
Water use	0.000,093	42.4	5,770,000	383.85	98,836.32	m ³

Tabelle 4.3: Gesamtindikatoren (Total) mit Median/IQR und Mittelwert (gerundet auf zwei Nachkommastellen).

Literatur

- [LB95] William S. Lovegrove und David F. Brailsford. "Document analysis of PDF files: methods, results and implications". In: (1995). URL: <https://nottinghamrepository.worktribe.com/output/1024553>.
- [CF04] Hui Chao und Jian Fan. "Layout and Content Extraction for PDF Documents". In: (2004). DOI: 10.1007/978-3-540-28640-0_20. URL: https://link.springer.com/chapter/10.1007/978-3-540-28640-0_20.
- [FM09] Murray J. Fisher und Andrea P. Marshall. "Understanding descriptive statistics". In: (2009). DOI: <10.1016/j.aucc.2008.11.003>. URL: <https://www.sciencedirect.com/science/article/abs/pii/S1036731408001732>.
- [MJ10] Gill Marshall und Leon Jonker. "An introduction to descriptive statistics: A review and practical guide". In: (2010). DOI: <10.1016/j.radi.2010.01.001>. URL: <https://www.sciencedirect.com/science/article/pii/S1078817410000027>.
- [Has+13] Mehrdad Hassanzadeh u. a. "Environmental declaration in compliance with ISO 14025 thanks to a collaborative program of electrical and electronic industry: The PEP ecopassport program". In: (2013). DOI: <10.1049/cp.2013.0577>. URL: <https://ieeexplore.ieee.org/abstract/document/6683180>.
- [Lip+13] Mario Lipinski u. a. "Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents". In: (2013). DOI: <10.1145/2467696.2467753>. URL: <https://dl.acm.org/doi/abs/10.1145/2467696.2467753>.
- [Gri15] Ralph Grishman. "Information Extraction". In: (2015). DOI: <10.1109/MIS.2015.68>. URL: <https://ieeexplore.ieee.org/abstract/document/7243219>.
- [Pez+16] Felipe Pezoa u. a. "Foundations of JSON Schema". In: (2016). DOI: <10.1145/2872427.2883029>. URL: <https://dl.acm.org/doi/abs/10.1145/2872427.2883029>.
- [BK17] Hannah Bast und Claudius Korzen. "A Benchmark and Evaluation for Text Extraction from PDF". In: (2017). DOI: <10.1109/JCDL.2017.7991564>. URL: <https://ieeexplore.ieee.org/abstract/document/7991564>.

- [CZ17] Andreiwid Sheffer Corrêa und Pär-Ola Zander. "Unleashing Tabular Content to Open Data: A Survey on PDF Table Extraction Methods and Tools". In: (2017). DOI: [10.1145/3085228.3085278](https://doi.org/10.1145/3085228.3085278). URL: <https://dl.acm.org/doi/abs/10.1145/3085228.3085278>.
- [KSY18] Parampreet Kaur, Jill Stoltzfus und Vikas Yellapu. "Descriptive statistics". In: (2018). DOI: [10.4103/IJAM.IJAM_7_18](https://doi.org/10.4103/IJAM.IJAM_7_18). URL: https://journals.lww.com/ijam/fulltext/2018/04010/Descriptive_statistics.7.aspx.
- [Dim+19] Gabrijela Dimić u. a. "Descriptive Statistical Analysis in the Process of Educational Data Mining". In: (2019). DOI: [10.1109/TELSIKS46999.2019.9002177](https://doi.org/10.1109/TELSIKS46999.2019.9002177). URL: <https://ieeexplore.ieee.org/document/9002177>.
- [MPV22] Douglas C. Montgomery, Elizabeth A. Peck und G. Geoffrey Vining. "Introduction to Linear Regression Analysis". In: (2022). URL: <http://wiley.com/en-ie/Introduction+to+Linear+Regression+Analysis%2C+Solutions+Manual-p-9781119578765>.
- [Ass24] Association P.E.P. *PEP Ecopassport*. Offizielle Website der Initiative für Umweltdeklarationen elektronischer Produkte. 2024. URL: <https://www.pep-ecopassport.org/> (besucht am 18.10.2025).
- [Aue+24] Christoph Auer u. a. "Docling Technical Report". In: (2024). DOI: [10.48550/arXiv.2408.09869](https://doi.org/10.48550/arXiv.2408.09869). URL: <https://arxiv.org/abs/2408.09869>.
- [Nad+24] Rohaan Nadeem u. a. "Extraction of User-Defined Information from PDF". In: (2024). DOI: [10.1109/DASA63652.2024.10836169](https://doi.org/10.1109/DASA63652.2024.10836169). URL: <https://ieeexplore.ieee.org/document/10836169>.
- [AA25] Narayan S. Adhikari und Shradha Agarwal. "A Comparative Study of PDF Parsing Tools Across Diverse Document Categories". In: (2025). DOI: [10.48550/arXiv.2410.09871](https://doi.org/10.48550/arXiv.2410.09871). URL: <https://arxiv.org/abs/2410.09871>.
- [Aue+25] Christoph Auer u. a. "Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion". In: (2025). DOI: [10.48550/arXiv.2501.17887](https://doi.org/10.48550/arXiv.2501.17887). URL: <https://arxiv.org/abs/2501.17887>.
- [Mor+25] José Teófilo Moreira-Filho u. a. "Automating Data Extraction From Scientific Literature and General PDF Files Using Large Language Models and KNIME: An Application in Toxicology". In: (2025). DOI: [10.1002/wcms.70047](https://doi.org/10.1002/wcms.70047). URL: <https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wcms.70047>.
- [YCZ25] Wen Yang, Feifei Cao und Xueli Zhao. "Extraction of PDF Table Data Based on the Pdfplumber Method". In: (2025). DOI: [10.1145/3696474.3696731](https://doi.org/10.1145/3696474.3696731). URL: <https://dl.acm.org/doi/full/10.1145/3696474.3696731>.