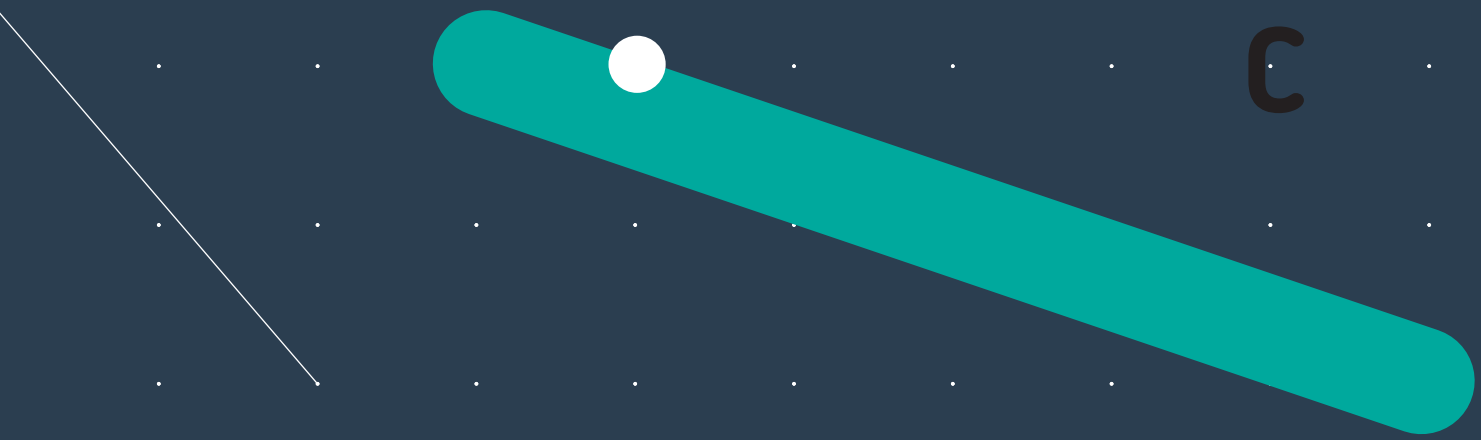
Hochschule Konstanz
Department of Computer Science

Submitted by
Jonas Mayer
Student Number 305630

jonas.mayer@htwg-konstanz.de



Bachelor Thesis

Datenanalyse PEP

Konstanz, TODO

Hochschule Konstanz
Department of Computer Science

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

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 Ecopassports, 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 Umweltproduktdeklarationen 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*) und die Referenzlebensdauer, hier meist 10 bis 20 Jahre. Weiterhin werden 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 (Elektronik, Sonstiges). Im Beispiel des ABB EQ Meter entfallen 28 % auf Kunststoffe, 53 % auf Metalle und 19 % auf weitere Komponenten. 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.

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

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. [Sel95] 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 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. Herausforerungen 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]
- *Diakritische Zeichen:* Buchstaben mit Diakritika (z. B. à, ã) 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.[Sel95]

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 strukturierts Format zu bringen, um sie anschließend weiter zu analysieren.

Klassische Verfahren (regelbasierte Parser) Ein Beispiel für ein klassisches, regelbasiertes Extraktionstool 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 potenzielle 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 Ex-

traktionstools 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 Layouts7 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 Extraktionstools, 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 Extraktionstools 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-Analysemodell *DocLayNet* identifiziert auf Basis eines trainierten Objektdetektors verschiedene Seitenelemente und deren Begrenzungsrahmen – etwa Absätze, Überschriften, Listen, Abbildungen oder Tabellen. [Aue+24] Die-

se 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.

2.3. PDF-Datenextraktion (?)

Theoretische Grundlagen PDFs und Daten (Aufbau Textlayer usw), gpt Modell für extraktion (Vor-/Nachteile)

2.4. Statistische Methoden

Welche statistischen Methoden werden eingesetzt und worauf basieren sie?

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

Extraktionsaspekt	**pdfplumber (regelbasiert)**	**Docling (KI-gestützt)**
Zeichen- und Worterkennung	Gut: präzise Koordinaten- und Schriftanalyse, robust für editierbare PDFs	Gut: kombiniert geometrische und visuelle Merkmale, tolerant gegenüber Layoutabweichungen
Lesereihenfolge (Reading Order)	Schlecht: rein geometrisch, keine semantische Korrektur bei mehrspaltigen Layouts	Gut: Layout-Analysemodell erkennt Spalten, Absätze und Lesefluss kontextsensitiv
Absatz- und Textstruktur	Teilweise: rekonstruiert Absätze heuristisch anhand von Zeilenabständen	Gut: erkennt logische Abschnitte, Überschriften, Listen und Beschriftungen
Tabellenerkennung	Teilweise: funktioniert bei klaren Linienstrukturen; scheitert bei verschachtelten oder visuellen Tabellen	Gut: rekonstruiert Tabellenstruktur semantisch über das *TableFormer*-Modell
Grafiken und eingebettete Objekte	Nicht gelöst: keine Analyse von Abbildungen oder Vektorobjekten	Teilweise: erkennt Abbildungen und Bildunterschriften, jedoch keine inhaltliche Bildanalyse
Metadatenextraktion	Nicht gelöst: keine Erkennung oder Harmonisierung von Titel, Autor, Referenzen	Gut: Post-Processing extrahiert zentrale Metadaten und Sprachinformationen
Mehrsprachige Dokumente	Eingeschränkt: kein Sprachmodell integriert	Teilweise: automatische Spracherkennung und Layout-Korrektur im Post-Processing
Nicht-textuelle Inhalte (z. B. OCR)	Nicht gelöst: unterstützt nur editierbare PDFs	Gut: optionale OCR-Erkennung für gescannte Dokumente über *EasyOCR*
Komplexe Layouts (mehrspaltig, technische Doks)	Schwach: häufige Fehlsegmentierung und falsche Lesereihenfolge	Gut: robuste Layout-Analyse durch *DocLayNet*-Modell
Semantische Rollen (z. B. Caption, Footnote)	Nicht gelöst	Gut: semantische Klassifizierung unter-

3

Pipeline und Datenbasis

Im folgenden Kaiptel wird die Pipeline zur Datenextraktion aus PEP Ecopassport PDFs sowie die daraus gewonnene Datenbasis beschrieben.

3.1. Pipeline

Beschreibung vorheriger Pipeline, Probleme und Endresultat

3.1.1. PEP Recherche

Probleme mit der PEP Recherche und wie ich sie deshalb durchgefürt habe

3.1.2. Vorhandene Pipeline

Lennys Pipeline und Probleme

3.1.3. Pipeline-Versuch1

Lennys Pipeline versucht upzugrade. Hier werden vor allem die Probleme gezeigt

3.1.4. Docling und LLM basierte Pipeline

Docling beschreiben, neue gpt-api usw.

3.2. Datenbasis

Wie viele PDFs, wie ichs noch weiter flachgezogen habe, Probleme usw

3.2.1. Normalisierung

Mapping Material, Strommodell usw

3.2.2. Datenbereinigung

Wie Daten bereinigt (zb zu leere PDFs aussortiert), Ausreißer manuell analysiert, Werte nachtragen

Literatur

[Sel95] Lenny Selg. In: (1995). URL: https://nottingham-repository.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.

[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.

[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: 10.1145/3085228.30852. URL: https://dl.acm.org/doi/abs/10.1145/3085228.3085278.

[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. URL: https://arxiv.org/abs/2408.09869.

[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. 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. URL: https://arxiv.org/abs/2501.17887.

[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. URL: https://dl.acm.org/doi/full/10.1145/3696474.3696731.