



# DataLab WestSax

Zukunfts- und Wertschöpfungslabor Westsachsen: Ein regionaler Katalysator Für Datenbasierte Wertschöpfung

# EXTRAKTION VON INFORMATIONEN AUS (SEMI-)STRUKTURIERTEN DOKUMENTEN

... bezieht sich auf den Prozess der automatisierten Identifizierung, Extraktion und Verarbeitung von relevanten Informationen aus Dokumenten, die ein bestimmtes Format oder eine bestimmte Struktur aufweisen. Semi-strukturierte Dokumente können z.B. Tabellen, Formulare oder XML-Dateien sein, während unstrukturierte Dokumente Texte ohne erkennbare Struktur sind.

---

## Inhalt

|  |    |
|--|----|
| Extraktion von Informationen aus (semi-)strukturierten Dokumenten..... | 4  |
| Indikatoren für die Nutzung .....                                      | 5  |
| Vorteile .....   | 6  |
| Nachteile.....   | 8  |
| Use Cases .....  | 9  |
| Hinweise .....   | 10 |
| IDP Technologien im Vergleich.....                                     | 11 |
| IDP Technologien im Vergleich.....                                     | 12 |
| Kontakt .....  | 13 |

---

# EXTRAKTION VON INFORMATIONEN AUS (SEMI-)STRUKTURIERTEN DOKUMENTEN

Die Extraktion von Informationen aus solchen Dokumenten ist relevant, weil Unternehmen und Organisationen oft große Mengen von Dokumenten besitzen, die wichtige Informationen enthalten. Die extrahierten Informationen können für verschiedene Zwecke verwendet werden, z. B. für die Erstellung von Berichten, die Analyse von Kundenfeedback, die Überwachung von Vertragsbedingungen, die Erfassung von Finanzdaten und vieles mehr.

Um Informationen aus (semi-)strukturierten Dokumenten zu extrahieren, werden Technologien wie OCR (optische Zeichenerkennung), NLP (natürliche Sprachverarbeitung) und Machine Learning-Algorithmen eingesetzt. Diese Technologien helfen dabei, relevante Informationen aus den Dokumenten zu extrahieren und in strukturierter Form bereitzustellen, so dass sie für weitere Verarbeitungsschritte in anderen Systemen verwendet werden können.



## Indikatoren für die Nutzung

Automatisierte Verfahren zur Extraktion und Verarbeitung von Informationen aus Dokumenten können **Zeit und Ressourcen einsparen**, da manuelle Verarbeitungsschritte entfallen. Technologien wie OCR und NLP **verbessern die Genauigkeit** bei der Verarbeitung von unstrukturierten oder schlecht formatierten Dokumenten. Eine höhere Zuverlässigkeit und Konsistenz bei bestimmten Informationen aus Dokumenten ist für **Compliance-Anforderungen** besonders wichtig, während die **Prozessautomatisierung** zu einer höheren Effizienz und Produktivität führen kann.

---

# VORTEILE

Die Automatisierung von Prozessen ist ein entscheidender Vorteil der automatisierten Verarbeitung von Dokumenten. Durch die Minimierung von menschlichen Fehlern und die Beschleunigung von Arbeitsprozessen können Unternehmen und Organisationen erhebliche Einsparungen erzielen. Die Verarbeitung großer Datenmengen wird schnell und effizient durchgeführt, was die Leistungsfähigkeit der Organisation verbessert.

Die Genauigkeit der automatisierten Verarbeitung von Dokumenten wird durch die Verwendung von OCR, NLP und Machine Learning erhöht. Diese Technologien ermöglichen es, selbst unstrukturierte oder schlecht formatierte Dokumente genau zu verarbeiten, ohne dass menschliches Eingreifen erforderlich ist. Darüber hinaus kann die Skalierung der automatisierten Verarbeitung von Dokumenten ohne Beeinträchtigung der Genauigkeit durchgeführt werden.

Die Analysemöglichkeiten bei der automatisierten Verarbeitung von Dokumenten bieten einen weiteren Vorteil. Durch die Identifikation von Mustern und Zusammenhängen in den Daten können Unternehmen und Organisationen wertvolle Erkenntnisse gewinnen, die als Grundlage für Entscheidungen dienen können. Die automatisierte Verarbeitung von Dokumenten kann somit auch einen entscheidenden Beitrag zur Optimierung von Geschäftsprozessen und zur Steigerung der Effizienz leisten.



**Zeitersparnis**

**Entscheidungsgrundlage**

**Hohe Genauigkeit**

---

## NACHTEILE

Einer der größten Nachteile ist die Fehleranfälligkeit bei der OCR-Texterkennung. Wenn ein Dokument schlecht gescannt oder unleserlich ist, kann dies zu Fehlern bei der Texterkennung führen. Dadurch können falsche Informationen extrahiert und weiterverarbeitet werden, was zu Problemen führen kann.

Unternehmen müssen sicherstellen, dass die Qualität der Dokumente ausreichend ist, um eine genaue Verarbeitung zu gewährleisten.

Auch die Anpassung an neue Dokumenttypen erfordert oft neue Konfigurationen, was Zeit und Ressourcen benötigt. Darüber hinaus hat die automatisierte Verarbeitung von Dokumenten ein begrenztes Verständnis von Kontext, was zu ungenauen Ergebnissen führen kann. Schließlich spielen Kosten eine große Rolle, denn die Umsetzung von Machine-Learning- und Deep-Learning-Technologien kann sehr kostspielig sein.

**Genauigkeits- und Kostenprobleme können bei der Extraktion von Informationen auftreten.**



---

# USE CASES

Automatisierte Verfahren zur Extraktion und Verarbeitung von Informationen aus Dokumenten finden in einer Vielzahl von Anwendungsbereichen Einsatz. Im Rechnungswesen können Rechnungen und Zahlungsdokumente automatisiert verarbeitet werden, indem relevante Daten extrahiert und mit Bestellungen abgeglichen werden. Auch in der Versicherungsbranche können Schadensfälle automatisiert bearbeitet werden, um eine schnellere Schadensregulierung zu ermöglichen.

In Personalabteilungen können Lebensläufe und Bewerbungsunterlagen automatisiert verarbeitet werden, um den Bewerbungsprozess zu beschleunigen. Im Gesundheitswesen können Patientenakten, Versicherungsansprüche und Abrechnungen automatisiert verarbeitet werden, um den administrativen Aufwand zu reduzieren. In der Logistik können Frachtbriefe und Lieferscheine automatisiert verarbeitet werden, um den Versandprozess zu beschleunigen.

In Banken und Finanzdienstleistungen können Kontoauszüge, Finanzdaten und Antragsformulare automatisiert verarbeitet werden, um den Prozess der Kontoeröffnung oder der Kreditvergabe zu beschleunigen. Im E-Commerce können Bestellungen, Rechnungen und Versandketten automatisiert verarbeitet werden, um eine schnellere Abwicklung von Bestellungen und eine Reduktion des manuellen Aufwands zu ermöglichen.

Insgesamt zeigen die zahlreichen Use Cases für automatisierte Verfahren zur Extraktion und Verarbeitung von Informationen aus Dokumenten, dass diese Technologien in nahezu allen Branchen und Anwendungsbereichen eingesetzt werden können, um Prozesse zu beschleunigen, Kosten zu reduzieren und die Effizienz zu steigern.

---

# HINWEISE

Um eine erfolgreiche automatisierte Extraktion von Informationen aus Dokumenten zu gewährleisten, gibt es einige wichtige Hinweise zu beachten. Eine gründliche Analyse der Dokumente ist unerlässlich, um die Struktur und den Inhalt der Dokumente zu verstehen und mögliche Schwierigkeiten bei der Extraktion zu erkennen. Die Auswahl geeigneter Tools hängt von den Anforderungen des jeweiligen Use Cases ab und es muss entschieden werden, welche Technologien oder Kombinationen von Technologien am besten geeignet sind.

Eine Integration mit externen Systemen kann dazu beitragen, eine nahtlose Verarbeitung der extrahierten Informationen zu gewährleisten. Es ist wichtig, die Extraktionsergebnisse regelmäßig zu überwachen und die Extraktionsmodelle kontinuierlich zu verbessern, um eine höhere Genauigkeit und Zuverlässigkeit der Ergebnisse zu erzielen. Bei der Auswahl einer Software sollten die unterstützten Dokumenttypen, die zu extrahierenden Daten, die unterstützten Datenformate, die benötigten Integrationen und Schnittstellen sowie die Erfahrung und Expertise des Benutzers berücksichtigt werden. Die Benutzerfreundlichkeit, Skalierbarkeit, Genauigkeit und Sicherheit sind ebenfalls wichtige Faktoren, die bei der Entscheidung berücksichtigt werden sollten.

# IDP TECHNOLOGIEN IM VERGLEICH

|                               | TESTVERSION | ON-PREM DEPLOYMENT | CLOUD DEPLOYMENT | SAP INTEGRATION | MACHINE LEARNING |
|-------------------------------|-------------|--------------------|------------------|-----------------|------------------|
| OpenText Intelligent Capture  | ✗           | ✓                  | ✓                | ✓               | ✓                |
| Kofax TotalAgility            | ✗           | ✗                  | ✓                | ✓               | ✓                |
| UiPath Document Understanding | ✓           | ✓                  | ✓                | ✗               | ✓                |
| ABBYY FlexiCapture            | ✗           | ✗                  | ✓                | ✓               | ✓                |

|                            | TESTVERSION | ON-PREM DEPLOYMENT | CLOUD DEPLOYMENT | SAP INTEGRATION | MACHINE LEARNING |
|----------------------------|-------------|--------------------|------------------|-----------------|------------------|
| Automation Anywhere IQ Bot | ✗           | ✗                  | ✓                | ✓               | ✓                |
| Datamatics TruCap+         | ✗           | ✗                  | ✓                | ✓               | ✓                |

# IDP TECHNOLOGIEN IM VERGLEICH

OpenText Intelligent Capture

hohe Verarbeitungsmenge, einfache Integration in Enterprise-Systeme

Kofax TotalAgility

Echtzeit-Lernfähigkeit, Automatisierung von Abläufen, intelligentes Dokumentenmanagement

UiPath Document Understanding

präzise Dokumentenverarbeitung, Automatisierung von Dokumentationsprozessen

ABBYY FlexiCapture

hohe Genauigkeit bei der Extraktion von Daten, Automatisierung von Dokumentenmanagement

## NACHTEILE

steile Lernkurve, kontraintuitive Benutzeroberfläche

manche Tasks nicht konsistent, schwierige Anpassung von Workflows

hohe Systemanforderungen, Support begrenzt in Sprache

Support begrenzt in Sprache, eingeschränkte Flexibilität bei speziellen Anwendungsfällen

Automation Anywhere IQ Bot

Automatisierung von Dokumentenmanagement, intelligente Automatisierung, leistungsstark

Datamatics TruCap+

intuitiver Konfigurator, skalierbare Automatisierung von Dokumentenmanagement

## NACHTEILE

Support begrenzt in Sprache, schwierige Anpassung von Workflows

Support begrenzt in Sprache, schwierige Anpassung von Workflows, eingeschränkte Flexibilität bei speziellen Anwendungsfällen

---

**Melden Sie sich gern!  
Daten. Wissen(s)Wert.**

## KONTAKT

E-Mail: [kontakt@datalab-vestsax.de](mailto:kontakt@datalab-vestsax.de)

Webseite: <https://datalab-vestsax.de/>

**Fraunhofer IMW Leipzig**  
Neumarkt 9-19  
04109 Leipzig

**Westsächsische Hochschule**  
Zwickau  
Kornmarkt 1  
08056 Zwickau

### Besucheradresse

Westsächsische Hochschule Zwickau  
Fakultät Wirtschaftswissenschaften  
Professur Wirtschaftsinformatik  
Scheffelstraße 39  
Haus 3, Zimmer 3111  
08066 Zwickau