FINBRIDGE

Maschinelle Informationsextraktion aus gescannten Dokumenten

David Israel

Jördis Helmers

Photo by Paulius Dragunas on Unsplash, Abgerufen am 11.08.2020.

25. August 2020

Maschinelle Informationsextraktion aus gescannten Dokumenten

Im Geschäftsleben generell und auch in der Bankenwelt spielen Verträge eine zentrale Rolle und die Arbeit mit diesen nimmt viel Zeit im Tagesgeschäft ein. Die entsprechenden Dokumente werden heutzutage üblicherweise gescannt und in einem digitalen Format vorgehalten, die Weiterverarbeitung erfolgt aber häufig immer noch durch einen menschlichen Sachbearbeiter. Dieser liest und durchsucht den Vertrag als PDF-Dokument nach bestimmten Vertragsdetails und gleicht diese beispielsweise gegen vorgegebene Kriterien ab oder überträgt sie per Hand in ein IT-System. Solche manuellen Vertragsanalysen und -auswertungen sind oft aufwendig und fehleranfällig und nutzen die Möglichkeiten der Digitalisierung im Vorteil gegenüber der herkömmlichen Arbeit mit einem gedruckten Vertrag nicht vollumfänglich. Bei einer großen Menge an auszuwertenden Dokumenten sind vollständige Analysen mit herkömmlichen Methoden mitunter auch nicht in einem akzeptablen Zeitrahmen möglich. Es ist daher sinnvoll, auch die Weiterverarbeitung von gescannten Vertragsdokumenten möglichst weitgehend zu automatisieren und damit die Effizienz der Bearbeitung deutlich zu steigern.

Im Folgenden beschreiben wir typische Elemente, beispielhafte Schritte und auftretende Herausforderungen bei der automatisierten Informationsextraktion aus gescannten Dokumenten näher und zeigen verschiedene Anwendungsmöglichkeiten auf.

Maschinelle Texterkennung (OCR)

Bei einem gescannten Dokument muss zunächst grundsätzlich eine Texterkennung, auch Optical Character Recognition (OCR) genannt, durchgeführt werden, da Bilddateien nur aus Farbpixeln bestehen und per se keinen Text enthalten. Nur dann kann der Vertragstext im Anschluss auch maschinell weiterverarbeitet werden. Viele Scanner können zwar prinzipiell über mitgelieferte

Software direkt eine Texterkennung durchführen, häufig werden gescannte Dokumente aber nur als einfache Bilddatei oder als bildbasiertes PDF abgespeichert. Eine Texterkennung kann mithilfe von kommerzieller Software oder auch über Cloud Services durchgeführt werden. Alternativ kann sie als Bestandteil von Automatisierungsroutinen selbst programmiert werden, gewöhnlich unter Zuhilfenahme von Open-Source-Tools wie Tesseract.

In der praktischen Anwendung wird die Texterkennung häufig durch eine schlechte Scanqualität erschwert, weshalb eine Vorverarbeitung der Scans nötig ist, z.B. Drehen oder Begradigen von Seiten, Entfernen von Artefakten oder Erhöhung des Kontrastes. Außerdem kann eine Layoutanalyse durchgeführt und entschieden werden, ob die gesamte Seite als fortlaufender Fließtext interpretiert werden soll oder ob die Texterkennung alternativ in einzelnen Bereichen separat durchgeführt wird. Letzteres ist bei strukturierten Dokumenten wie Verträgen häufig sinnvoll, da Inhalte oft als zusammenhängende Textblöcke, Tabellen oder Ähnliches formatiert sind.

Ebenso kann optional eine Nachbearbeitung des erkannten Textes durchgeführt werden, um sicherzustellen, dass nur zulässige Wörter erkannt werden oder die erkannten Wörter im jeweiligen Kontext Sinn ergeben. So kann die Erkennungs-qualität zusätzlich gesteigert werden. Weiterhin sollten typische Buchstaben- und Zeichenfehler aus der Texterkennung bei der Weiterverarbeitung in sogenannten Regulären Ausdrücken berücksichtigt werden. So bereitet das in Vertragstexten häufig auftretende Paragrafenzeichen „§“ vielen Texterkennungssystemen Probleme und es werden gelegentlich andere Zeichen wie „$“, „&“, „8“ oder „B“ erkannt. Auch Verwechslungen der sich stark ähnelnden Zeichen „5“ und „S“ oder „1“ und „I“ sind typisch. Mittels Regulärer Ausdrücke können z.B. mehrere mögliche Zeichen als gleich angesehen und somit solche Erkennungsfehler ignoriert werden.

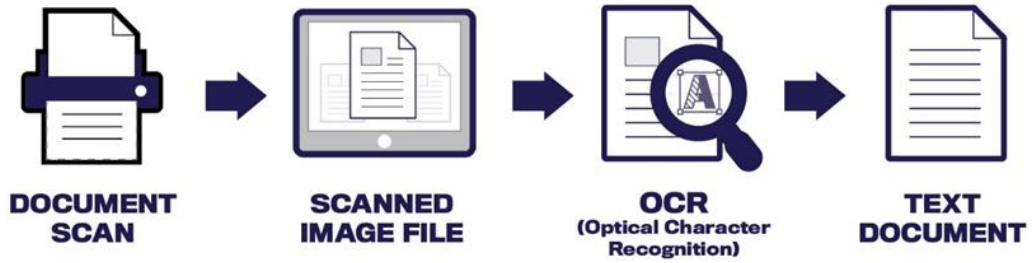


Abbildung 1: Mittels Optical Character Recognition (OCR) wird der in einem gescannten Dokument enthaltene Text erkannt und zur maschinellen Weiterverarbeitung verfügbar gemacht.

Maschinelle Dokumentauswertung

Zur Auswertung von Dokumenten reicht die oben beschriebene einfache Texterkennung, z.B. zeilenweise über das gesamte Dokument, für gewöhnlich nicht aus, da Informationen in Verträgen und strukturierten Dokumenten in unterschiedlicher Form vorliegen können. Das Layout und die Anordnung der Textpassagen spielt also eine Rolle und muss zur korrekten Interpretation berücksichtigt werden. Dem tragen sogenannte Parser Rechnung.

Ein Parser ist im Allgemeinen ein Computerprogramm, das für die Zerlegung und Umwandlung einer Eingabe in ein für die Weiterverarbeitung geeigneteres Format zuständig ist. Ein Dokumenten-Parser extrahiert die ausformulierten oder formatiert dargestellten Vertragsdetails aus einem Textdokument und überführt sie in einfache Datensätze. Dabei sind jeweils individuelle Parsing-Regeln pro Dokumenttyp (z.B. Formular mit einem speziellen Layout) und Ergebniselement (z.B. Kundenadresse) nötig, mittels welcher die Informationsextraktion erfolgt. Eine Parsing-Regel besteht dabei aus mehreren Einzelschritten, um die gewünschten Informationen je nach Layout und Vorkommen auszulesen und gegebenenfalls anders zu formatieren. Parsing-Regeln werden anfangs vom Nutzer spezifiziert oder von der Anwendung gelernt und können dann zur automatisierten Verarbeitung einer großen Menge an Dokumenten verwendet werden.

Grundsätzlich lässt sich also eine maschinelle Vertragsauswertung grob in vier Schritte unterteilen:

1. Klassifikation des jeweiligen Dokuments
2. Anwendung der passenden Parsing-Regel
3. Plausibilitätsprüfung
4. Ergebnisgenerierung

Nachfolgend erläutern wir diese Schritte und dabei bestehende Alternativen näher.

Klassifikation des jeweiligen Dokuments

Im ersten Schritt muss das auszuwertende Dokument klassifiziert werden, um entscheiden zu können, welche zum Dokumenttyp passende Parsing-Regel angewendet werden soll. Falls die Dokumente nicht vorklassifiziert oder z.B. entsprechend gruppiert abgespeichert wurden, muss die Analysesoftware diese Klassifikation möglicherweise selbst leisten. Prinzipiell sind drei verschiedene Herangehensweisen für die Klassifikation möglich: textbasiert, layoutbasiert oder durch Nutzung von Metadaten.

Textbasierte Klassifikation

Das natürliche Vorgehen zur Klassifikation von Dokumenten ist textbasiert. So können beispielsweise häufig auftretende Schlüsselwörter oder abgeleitete, im Dokument auftauchende inhaltliche Themen zur Zuordnung zu einer bestimmten Dokumentenklasse verwendet werden. Bei strukturierten Dokumenten wie Verträgen bietet es sich an, gezielt bestimmte Textpassagen zu analysieren. So liefert z.B. die Überschrift des Dokuments häufig relevante Hinweise auf den Inhalt (z.B. „Auszahlungsbestätigung zu ihrer Baufinanzierung“). Da Vertragsdokumente normalerweise aus vorgegebenen Textbausteinen generiert werden, führt die Suche nach solchen charakteristischen Textbausteinen auch oft zum Erfolg.

Layoutbasierte Klassifikation

Alternativ ist es möglich, auf den gescannten Dokumentenseiten direkt einen Bildklassifikationsalgorithmus anzuwenden. Dabei kann bei einer charakteristischen Seitenstruktur zum einen die gesamte Seite des Dokuments

betrachtet werden, alternativ kann sich auch wieder nur auf bestimmte Regionen mit Firmenlogos oder Ähnlichem beschränkt werden. Bei der Bildklassifikation werden die Bildpixel direkt und nicht der gedruckte Text betrachtet und zur Entscheidungsfindung herangezogen. Mit diesem Vorgehen spart man sich also die meist rechenaufwändige Texterkennung. Hierbei ist es allerdings erforderlich, dass sich die relevanten Seiten der unterschiedlichen Dokumentenklassen optisch deutlich voneinander unterscheiden. Zusätzlich kann man so auch die Seiten eines Dokuments nach Art des Inhaltes einteilen, z.B. in die Klassen „Deckblatt“, „Fließtext“, „Leerseite“ und „Term Sheet“.

Klassifikation mittels Metadaten

Eine dritte Möglichkeit stellt die Verwendung von Metadaten des Scans dar. Beispiele sind hier die Seitenzahl des Dokuments, Dateiname, Dateityp oder Dateigröße. In manchen Anwendungsfällen lässt sich auch so schon auf die Art des Dokuments schließen oder zumindest eine Grobklassifikation vornehmen.

Anwendung der Parsing-Regel

Wurde das Dokument einer bekannten Dokumentenklasse zugeordnet, findet die passende Parsing-Regel Anwendung, um die gewünschten Informationen auszulesen. Im Folgenden betrachten wir drei typische Dokumentenformate und stellen denkbare zugehörige Parsing-Regeln dar:

Formatierte Textblöcke

Bei diesem häufig auftretenden Fall stehen zusammengehörige Informationen in einem Block und sollten auch blockweise ausgelesen und interpretiert werden. Ein einfaches zeilenweises Vorgehen vermischt nicht-zusammengehörige Text-passagen.


Bird-Bank

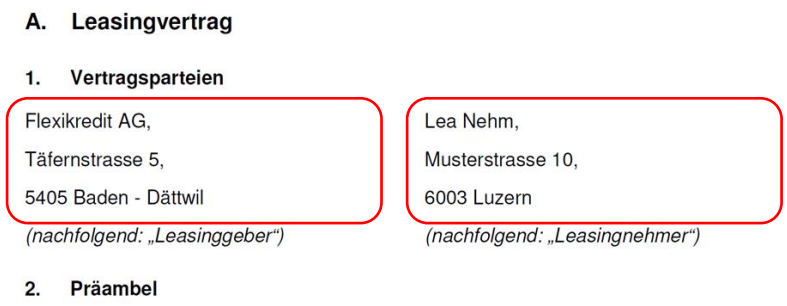
-vertraulich-
Team Kapitalmarkt Compliance

Bird-Bank GmbH

Depotnummer: 1239994545
Depotinhaber: Jürgen-Roland Schulte
Vermerk der Bank: Duplikat

Wertpapierabrechnung	**Kauf**
Ordernummer	**13281196.001**
ISIN (WKN)	LU0274211480 (DBX1DA)
Wertpaperbezeichnung	Xtrackers DAX Inhaber-Anteile 1C o.N.

Abbildung 2: Der relevante, rot umrahmte Textblock wird separat ausgelesen. Informationen in „Key-Value-Form" lassen sich interpretieren und direkt in ein strukturiertes Format übertragen, z.B. Key = „Depotnummer", Value = 1239994545.


A. Leasingvertrag

1. Vertragsparteien

Flexikredit AG,
Täfernstrasse 5,
5405 Baden - Dättwil
(nachfolgend: „Leasinggeber")

Lea Nehm,
Musterstrasse 10,
6003 Luzern
(nachfolgend: „Leasingnehmer")

2. Präambel

Abbildung 3: Die rot umrahmten Adressblöcke werden separat ausgelesen und mittels ihres standardisierten Aufbaus interpretiert.

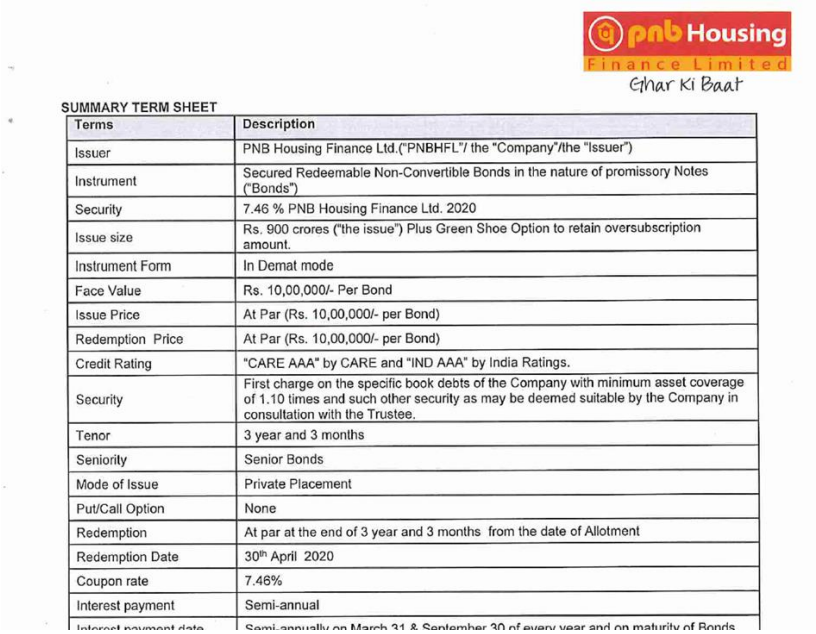
Schritte der Parsing-Regel:

1. Lokale Texterkennung im Bereich des jeweiligen Textblocks (Zonal OCR): Die zusammenhängenden Blöcke werden aus festgelegten Zonen separat ausgelesen. Bei gegebenem Layout kann der jeweilige Textblock immer im gleichen, in der Parsing-Regel festgelegten Bildausschnitt, gefunden werden.

2. Interpretation je nach Format (z.B. „Key-Value-Form", Adressblock, o. Ä.): Je nach Format des Textblocks findet die Informationsextraktion statt. Bei Informationen in „Key-Value-Form" ist eine unmittelbare Zuordnung von Werten zu den jeweiligen Ergebnisfeldern möglich (vgl. Abbildung 2). Adressblöcke haben einen immer gleichen Aufbau, mittels welchem sie sich standardisiert interpretieren lassen, z.B. „Name – Straße – Hausnummer – Postleitzahl – Ort" (vgl. Abbildung 3).

Tabelle oder Term Sheet

Eine Tabelle besteht aus durch Rahmenlinien getrennten Zellen, wobei diese Rahmenlinien jeweils zusammengehörige Textblöcke voneinander abgrenzen. Ein OCR-Tabellen-Parser muss den erkannten Text den richtigen Zellen zuordnen.



pnb Housing Finance Limited
Ghar Ki Baat

SUMMARY TERM SHEET

Terms	Description
Issuer	PNB Housing Finance Ltd.("PNBHFL"/ the "Company"/the "Issuer")
Instrument	Secured Redeemable Non-Convertible Bonds in the nature of promissory Notes ("Bonds")
Security	7.46 % PNB Housing Finance Ltd. 2020
Issue size	Rs. 900 crores ("the issue") Plus Green Shoe Option to retain oversubscription amount.
Instrument Form	In Demat mode
Face Value	Rs. 10,00,000/- Per Bond
Issue Price	At Par (Rs. 10,00,000/- per Bond)
Redemption Price	At Par (Rs. 10,00,000/- per Bond)
Credit Rating	"CARE AAA" by CARE and "IND AAA" by India Ratings.
Security	First charge on the specific book debts of the Company with minimum asset coverage of 1.10 times and such other security as may be deemed suitable by the Company in consultation with the Trustee.
Tenor	3 year and 3 months
Seniority	Senior Bonds
Mode of Issue	Private Placement
Put/Call Option	None
Redemption	At par at the end of 3 year and 3 months from the date of Allotment
Redemption Date	30th April 2020
Coupon rate	7.46%
Interest payment	Semi-annual
Interest payment date	Semi-annually on March 31 & September 30 of every year and on maturity of Bonds

Abbildung 4: Beim Auslesen der Tabelle muss der erkannte Text jeweils der richtigen Zelle zugeordnet werden.

Schritte der Parsing-Regel:

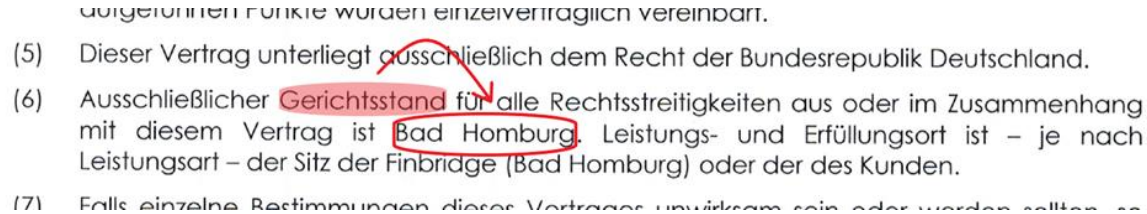
1. Optische Analyse zur Erkennung von Rahmenlinien und Positionen der Zellen:
   Zuerst müssen die horizontalen und vertikalen Rahmenlinien erkannt und daraus die Struktur und Lage der Tabelle errechnet werden. Die Pixel-Positionen müssen in die entsprechenden Zeilen und Spalten der Tabelle überführt werden, also z.B. Box an Position (x, y) entspricht Zelle B3.

2. Zellenweise Texterkennung:
   Folgend wird eine zellenweise Texterkennung durchgeführt. Durch die Beschränkung auf jeweils eine Zelle repräsentierende Bildausschnitte kann der erkannte Text eindeutig einer Tabellenzelle zugeordnet werden.

3. Weitergehende Interpretation der Tabelle:
   Zusätzlich kann eine weitergehende Interpretation oder Umformatierung der Tabelle erfolgen. Im Beispiel könnte man festlegen, dass Spalte „Terms“ den Feldbezeichnern (Key) und Spalte „Description“ den zugehörigen Feldwerten (Value) eines Datenbankeintrags entspricht (vgl. Abbildung 4).

Informationen im Fließtext

Ausformulierte oder im Vertragstext „versteckte“ Informationen sind generell schwieriger zu extrahieren. Hierbei kommt es nicht so sehr auf die korrekte Interpretation des Layouts an, sondern es ist Textverständnis oder evtl. sogar Fachwissen zur Informationsextraktion notwendig. Auch eine möglicherweise von Vertrag zu Vertrag unterschiedliche Formulierung muss korrekt verarbeitet werden und jeweils das entsprechende Ergebnis liefern. Daher ist für eine automatisierte Informationsextraktion auch ein gewisses maschinelles Textverständnis nötig. Solche Parsing-Regeln können Elemente des Natural Language Processing (NLP) enthalten und werden z.B. mit Hilfe von Künstlichen Neuronalen Netzen realisiert.


aufgeführten Punkte wurden einzelvertraglich vereinbart.

(5) Dieser Vertrag unterliegt ausschließlich dem Recht der Bundesrepublik Deutschland.

(6) Ausschließlicher Gerichtsstand für alle Rechtsstreitigkeiten aus oder im Zusammenhang mit diesem Vertrag ist Bad Homburg. Leistungs- und Erfüllungsort ist – je nach Leistungsart – der Sitz der Finbridge (Bad Homburg) oder der des Kunden.

(7) Falls einzelne Bestimmungen dieses Vertrages unwirksam sein oder werden sollten, so

Abbildung 5: Beispiel „Gerichtsstandsextraktion“: Die gesuchte Bezeichnung muss im Vertragstext gefunden werden.

Schritte der Parsing-Regel:

1. Texterkennung über den gesamten Vertrag:
   Es wird eine zeilenweise Texterkennung über den gesamten Vertrag durchgeführt, um den Vertragstext im Anschluss maschinell weiterverarbeiten zu können. Somit ist kein einheitlicher Aufbau oder eine einheitliche Formulierung der Verträge notwendig.

2. Suche nach Schlüsselwort (z.B. „Gerichtsstand“):
   Zum Auffinden des relevanten Abschnittes im Vertrag wird eine Schlüsselwortsuche durchgeführt. Dabei wird davon ausgegangen, dass das gesuchte Ergebnis in unmittelbarer Nähe des Schlüsselwortes im Vertragstext zu finden ist.

3. Entscheidung über Umfang des zugehörigen Absatzes (Kontext):
   Zur weiteren Eingrenzung der relevanten Textpassage wird der zum Schlüsselwort gehörige Kontext, also ein inhaltlich zusammenhängender Textabschnitt, extrahiert. Dazu können Textmarker verwendet werden, welche üblicherweise zusammengehörige Absätze voneinander trennen, z.B. in Verträgen §1, (a), (5). Möglich wäre auch die Betrachtung einzelner Sätze. Im Folgenden wird nur dieser zugehörige Kontext weiter analysiert, welcher auch selbst schon als Ergebniselement denkbar wäre.

4. Extraktion der gesuchten Information:
   Im finalen Schritt wird der extrahierte Kontext mit Hilfe von Sprachmodellen weitergehend analysiert. Zur Bestimmung des gesuchten Gerichtsstandes eines Vertrages beispielsweise muss die zugehörige Ortsbezeichnung extrahiert werden (vgl. Abbildung 5). Dafür wird eine sogenannte „Named-Entity Recognition“ durchgeführt, eine automatische Identifikation und Klassifikation von Eigennamen im Text.

Plausibilitätsprüfung

Zur Qualitätssicherung und Sicherstellung der Korrektheit der ausgelesenen Ergebnisse sollte nach Anwendung des Parsers eine Überprüfung und evtl. Anpassungen erfolgen. Neben generellen Fehlern in der Auswerteroutine können auch hier eine schlechte Scanqualität oder eine unerwartete Änderung im Dokumentenaufbau zu Problemen führen. Denkbar ist zum einen eine maschinelle Plausibilisierung der vorläufigen Ergebnisse, z.B. auf plausible Datentypen, korrektes Format oder mittels Prüfziffern, welche beispielsweise bei vielen Identcodes, IBAN, ISIN usw. vorhanden sind. Fehlerhafte Werte können im

Zuge der Plausibilitätsprüfung gegebenenfalls auch direkt korrigiert werden. Alternativ kann der Auslesealgorithmus mit einer angepassten Parametrisierung auch erneut ausgeführt werden. Zumindest aber sollte der Anwender über unplausible oder auch fehlende Ergebnisse informiert werden. Darüber hinaus kann eine manuelle Nachbearbeitung und Plausibilisierung erfolgen, zumindest stichprobenartig, um Fehler zu korrigieren oder fehlende Werte zu ergänzen.

Ergebnisgenerierung

Das Ergebnis sind prinzipiell gewünschte Informationen in einem strukturierten Format. Die im Dokument als Text ausformulierten oder in einem speziellen Layout vorliegenden Informationen werden ausgelesen und in Datensätze bestehend aus Feldbezeichner und Feldinhalt überführt. Danach kann eine maschinelle Weiterverarbeitung dieser Daten erfolgen, z.B. ein maschineller Abgleich, ein Eintrag in eine Datenbank, die Erzeugung einer formatierten Excel-Datei, eine Abspeicherung im CSV-Format, o. Ä.

Unser Angebot

Die hier vorgestellten Methoden und Lösungsansätze zur Durchführung einer automatischen Informationsextraktion mittels OCR aus gescannten Verträgen und anderen Dokumenten bieten umfangreiche Anwendungsmöglichkeiten für die maschinelle Vertragsanalyse und Weiterverarbeitung. Die folgenden Themen stehen beispielhaft dafür:

- Aufbau einer Dokumentendatenbank, inklusive Extraktion von Metadaten und Vertragsinhalten zur Vorhaltung in einem strukturierten Format,
- Automatische Vertragserfassung in einem Banksystem,
- Maschineller Abgleich von Vertragsdetails zwischen gescannten Dokumenten und Einträgen in einer Datenbank zur Qualitätssicherung oder als Prüfungshandlung,
- Automatisierte Kreditantragsprüfung, Risikobewertung, o. Ä.,
- Dokumentenkontrolle im Rahmen der Compliance- oder Personaltätigkeit,
- Maschinelle Auswertung vorhandener Verträge oder externer Dokumente nach besonderen Eigenschaften durch Entwicklung maßgeschneiderter Parsing Regeln.

Wie aufgezeigt sind zur Digitalisierung und Prozessautomatisierung meist individuelle, von den konkreten Anforderungen und gegebenen Rahmenbedingungen abhängige Lösungen erforderlich. Zur Umsetzung solcher individuellen Lösungen bringen unter anderem unsere folgenden Kernkompetenzen einen maßgeblichen Mehrwert zur effizienten Zielerreichung:

- Schnelle, gründliche und bereichsübergreifende Prozessanalyse
- Detailliertes Prozessverständnis, insbesondere beim Design zur Digitalisierung und/oder Automatisierung
- Fachliches und technisches Problemlösungs-Know-how, Ergebnisorientierung und Integrationsfähigkeit bei der Entwicklung methodischer Ansätze
- Erfahrung in der Entwicklung von individuellen Prototypen mittels Best Practice Lösungen sowie in der Begleitung der Integration dieser entwickelten Lösungen in eine vorhandene IT-Landschaft

Gern beraten wir Sie im Rahmen einer Vorstudie bei der Analyse möglicher Lösungsvarianten und helfen Ihnen außerdem bei der Entwicklung einer individuellen Lösung, die im Rahmen der fachlichen und technischen Möglichkeiten adäquat zu den Rahmenbedingungen Ihres Unternehmens passt. Darauf aufbauend unterstützen wir Sie ebenfalls gern bei der technischen Umsetzung sowie bei der Einbettung in vorhandene Prozesse. Die Schulung interner Mitarbeiter und eine bedarfsgerechte Anwendungsbetreuung sind für uns ebenso selbstverständlich wie ein kundenorientierter Lösungsansatz und die Umsetzung eventuell benötigter individueller Gestaltungen. Gern unterstützen wir darüber hinaus im Projekteinsatz oder/und dem regelmäßigen Betrieb.

Wir hoffen Ihr Interesse an unserer Beratung geweckt zu haben und freuen uns auf Ihre Kontaktaufnahme!

Team

David Israel
Senior Financial Engineer
Financial Engineering
eMail | LinkedIn | Xing

Jördis Helmers
Senior Manager
Financial Engineering
eMail | Xing

Über Uns

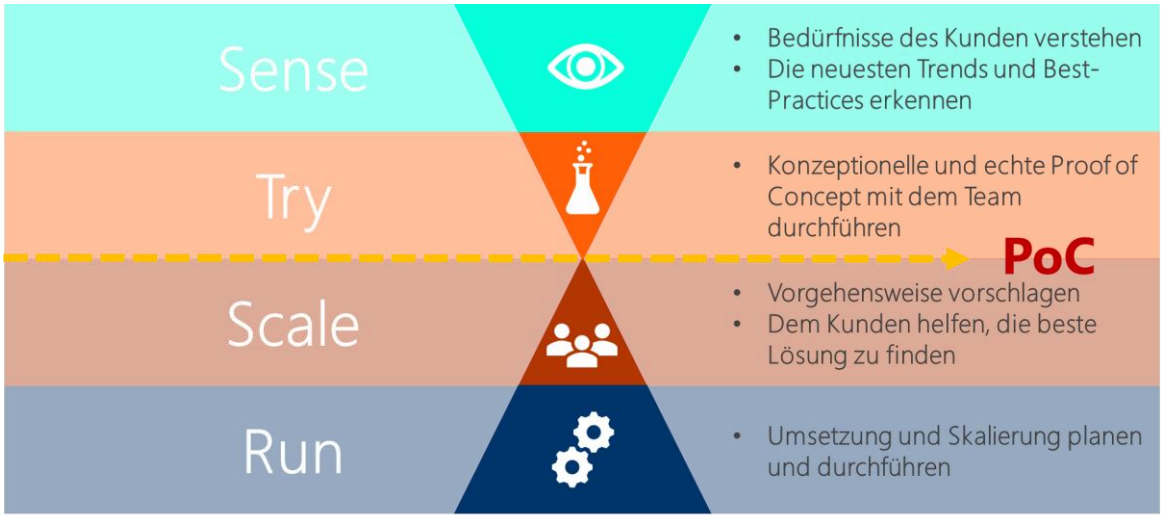
Finbridge GmbH & Co. KG ist ein unabhängiges, spezialisiertes Beratungsunternehmen im Bereich Financial Services und unterstützt die gesamte Prozesskette von Finanzprodukten in Kredit, Kapitalmarkt, Treasury, Risikocontrolling, Compliance, Accounting und Meldewesen.

Digital Transformation @ Finbridge

Digital Transformation ist die neueste Initiative von Finbridge, die die Einführung innovativer Methoden und Technologien bei unseren Kunden fokussiert.

Finbridge arbeitet integriert und strukturiert an verschiedenen Fronten der Digital Transformation. Wir unterstützen unsere Kunden bei der Bewältigung individueller Herausforderungen, insbesondere im Kontext der Digitalisierung, wenn vorhandene, klassische Technologien und Prozesse an ihre Grenzen stoßen.

Unsere Experten profitieren von langjähriger Erfahrung aus verschiedensten Projekteinsätzen und sind bestens vertraut mit den Herausforderungen, die sich im täglichen Betrieb unserer Kunden ergeben.



Innovationspfad: wie können wir unsere Kunden unterstützen?
Quelle: Peter Hinssen / Finbridge

FINBRIDGE

Mehr Insights

Finbridge GmbH & Co. KG
Louisenstraße 100
61348 Bad Homburg v. d. H.
www.finbridge.de