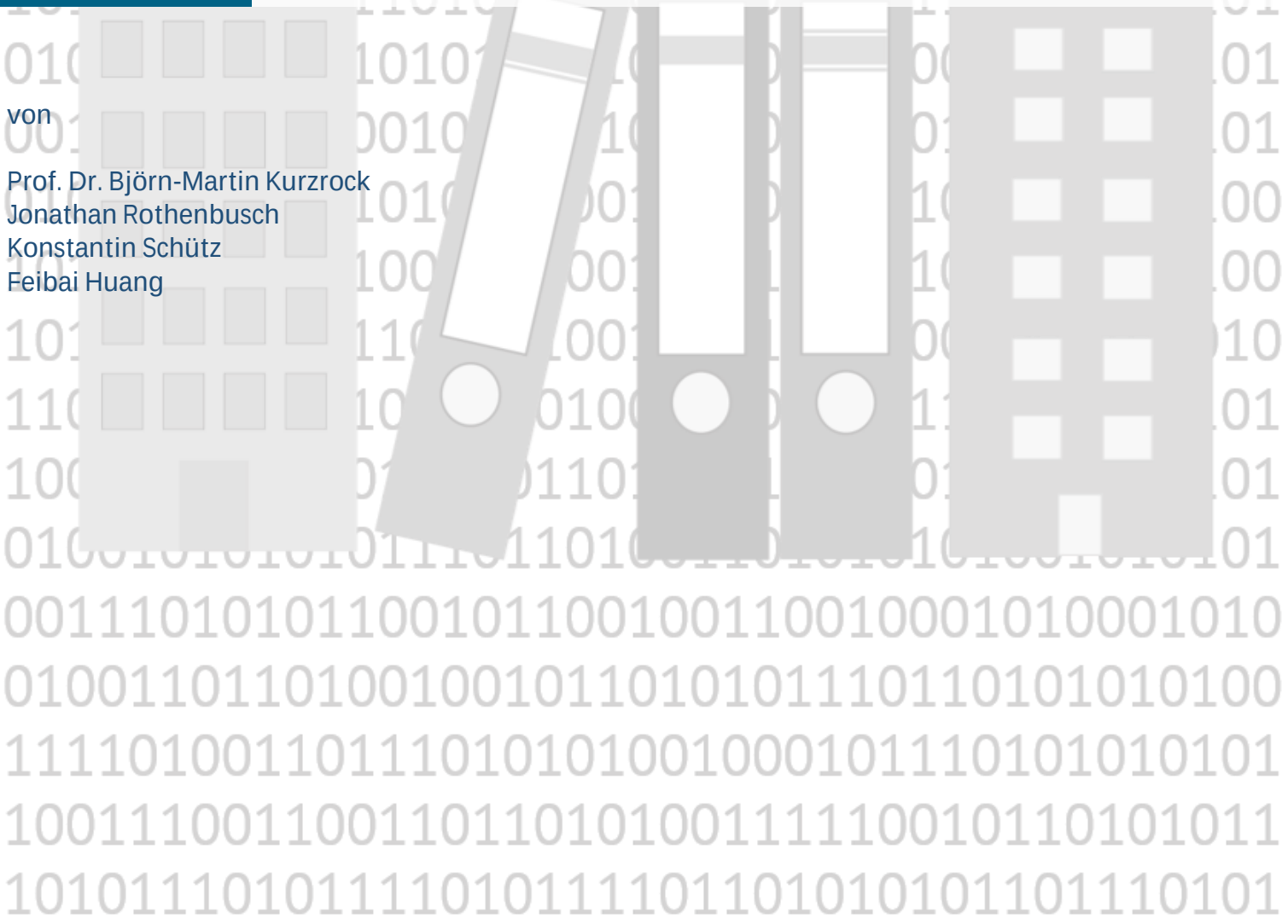
ZUKUNFT BAU
FORSCHUNGSFÖRDERUNG

Bundesinstitut
für Bau-, Stadt- und
Raumforschung
im Bundesamt für Bauwesen
und Raumordnung

BBSR-
Online-Publikation
48/2023

Maschinelles Lernen im Bereich Gebäudedokumentation



von

Prof. Dr. Björn-Martin Kurzrock
Jonathan Rothenbusch
Konstantin Schütz
Feibai Huang

Maschinelles Lernen im Bereich Gebäudedokumentation

Grundlagen der Informationsextraktion für Energieeffizienz- und Lebenszyklusanalysen (ML-BAU-DOK)

Gefördert durch:

Bundesministerium
für Wohnen, Stadtentwicklung
und Bauwesen

aufgrund eines Beschlusses
des Deutschen Bundestages

ZUKUNFT BAU
FORSCHUNGSFÖRDERUNG

Dieses Projekt wurde gefördert vom Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) im Auftrag des Bundesministeriums für Wohnen, Stadtentwicklung und Bauwesen (BMWSB) aus Mitteln des Innovationsprogramms Zukunft Bau.
Aktenzeichen: 10.08.18.7-20.26
Projektlaufzeit: 04.2021 bis 01.2023

IMPRESSUM

Herausgeber
Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR)
im Bundesamt für Bauwesen und Raumordnung (BBR)
Deichmanns Aue 31–37
53179 Bonn

Fachbetreuerin
Bundesinstitut für Bau-, Stadt- und Raumforschung
Referat WB 3 „Forschung und Innovation im Bauwesen“
Anne Bauer
anne.bauer@bbr.bund.de

Autoren
RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Fachgebiet Immobilienökonomie
Prof. Dr. Björn-Martin Kurzrock (Projektleitung)
bjoern.kurzrock@rptu.de

Jonathan Rothenbusch
jonathan.rothenbusch@rptu.de

Konstantin Schütz
konstantin.schütz@bauing.uni-kl.de

Feibai Huang
huangf@rhrk.uni-kl.de

Redaktion
RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Fachgebiet Immobilienökonomie
Prof. Dr. Björn-Martin Kurzrock
Jonathan Rothenbusch

Stand
Januar 2023

Satz und Layout
RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Fachgebiet Immobilienökonomie
Prof. Dr. Björn-Martin Kurzrock
Jonathan Rothenbusch

Bildnachweis
Titelbild: RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Fachgebiet Immobilienökonomie

Zitierweise
Kurzrock, Björn-Martin; Rothenbusch, Jonathan; Schütz, Konstantin; Huang, Feibai, 2023: Maschinelles Lernen im Bereich Gebäudedokumentation: Grundlagen der Informationsextraktion für Energieeffizienz- und Lebenszyklusanalysen (ML-BAU-DOK). BBSR-Online-Publikation 48/2023, Bonn.

ISSN 1868-0097 Bonn 2023

Inhaltsverzeichnis

Kurzfassung

ML-BAU-DOK legt Regeln und Methoden für die Informationsextraktion von Informationen aus Gebäudedokumentation dar. Die Informationsextraktion basiert auf einer zuvor durchgeführten Segmentierung der Dokumente eines Massenscanverfahrens und der anschließenden Klassifizierung der Dokumentation nach anwendungsspezifischen Dokumentenklassen.

Der Forschungsbericht beginnt mit den grundlegenden Regeln der Digitalisierung gebäudebezogener Dokumentation. Zunächst müssen Standards festgelegt werden, um die gesamte Dokumentation auf ein einheitliches Niveau zu bringen. Insbesondere der Umgang mit papierbasierter Dokumentation und die damit einhergehenden Regeln der Digitalisierung von Bestandsdokumentation werden zu Beginn dargelegt.

Anschließend wird der Anwendungsfall Energieeffizienz- und Lebenszyklusanalyse abgegrenzt und analysiert. Dabei werden relevante Daten auf der Basis konventioneller deutscher Standards definiert, die im weiteren Projektverlauf zur zielgerichteten Definition von Schlüsselinformationen und Schlüsseldokumenten(-klassen) dienen.

Hierauf aufbauend werden Dokumentenklassen betrachtet und die verschiedenen Klassifizierungssysteme miteinander verglichen. Grundlage hierfür ist der Dokumentenklassenstandard nach Müller (2023). Dieser wurde unter Berücksichtigung von gängiger Gebäudedokumentation und der zukünftigen ML-basierten Auswertungen erweitert bzw. reduziert und entsprechend angepasst.

Die Anforderungen an Dokumente werden im Hinblick auf die maschinelle Verarbeitung beschrieben und die zugrunde gelegte Dokumentbasis auf Maschinenlesbarkeit und Zeichengenauigkeit untersucht. Um zukünftig Dokumente nach ihrer Datenqualität zu priorisieren, werden zwei anwendungsspezifische Modelle zur Bestimmung der Dokumentenklassen und Aussagekraft der Dokumente präsentiert.

Schließlich werden Codes für die automatisierte Segmentierung von Dokumenten aus einem Massenscanverfahren entwickelt und die Ergebnisse ausgewertet. Mit Hilfe von Klassifikationscodes werden Dokumente automatisiert klassifiziert. Dazu werden verschiedene Algorithmen beschrieben, die je nach vorhandener Datenbasis bevorzugt eingesetzt werden sollten. Der Bericht schließt mit einem allgemeinen Fazit inklusive Einschränkungen und Ausblick.

Abstract

ML-BAU-DOK provides rules and methods for the information extraction of data from building documentation. The research report opens with the basic rules for digitizing building-related documentation. First of all, standards have to be established in order to bring all documentation to a uniform level. In particular, the handling of paper-based documentation and the associated rules for the digitization of as-built documentation are defined at the beginning of the report.

Subsequently, the use case of energy efficiency and life cycle analysis is analyzed. Relevant data are defined on the basis of conventional German standards, which are used in the further course of the project for the target-oriented definition of key information and key documents (key document classes).

Document classes are then considered and the different class systems are compared. According to the application, the document class standard of Müller (2023) was used. This was extended or reduced to take account of common building documentation and future ML-based evaluations, and adapted accordingly.

In the following, requirements for documents with regard to automated processing are described and the document basis used in *ML-BAU-DOK* is checked for its machine readability and character accuracy. In order to prioritize documents according to their data quality in the future, two user-specific models for determining the document classes and expressiveness of the documents are presented.

Finally, codes for the automated segmentation of documents from a mass scanning process are defined and the results evaluated. Classification codes are used to classify documents automatically. For this purpose, different codes are described that should be used depending on the available database.

1 Einleitung

Die vorliegende Arbeit ist der Endbericht des Projektes *ML-BAU-DOK*, das die Grundlagen der Informationsextraktion für Energieeffizienz- und Lebenszyklusanalysen thematisiert. *ML-BAU-DOK* wurde gefördert vom Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) im Auftrag des Bundesministeriums für Wohnen, Stadtentwicklung und Bauwesen (BMWSB) aus Mitteln des Innovationsprogramms Zukunft Bau.

1.1 Problemstellung

Der Umgang mit Dokumenten aus dem Lebenszyklus der Immobilie ist kaum bis gar nicht geregelt. Dennoch sind sie die Basis für verantwortungsvolle Entscheidungen und ein nachhaltiges Management über den gesamten Lebenszyklus von Immobilien.

Weder die Art und Weise der optimalen Digitalisierung und damit verbundene Reproduktion des Dokuments als digitales Abbild noch die ordnungsgemäße Ablage, Namensgebung und Ordnung der gebäudebezogenen Dokumentation sind vereinheitlicht und als Regelwerk zur Verfügung gestellt.

Nicht nur in Archiven, sondern auch im Tagesgeschäft, ist Papier als Träger von relevanten Informationen omnipräsent. Dabei ist nicht nur der arbeitsplatzunabhängige Zugang ein wichtiger Faktor für Scans, sondern auch die Nutzbarmachung der Dokumente für zukünftige digitale Prozesse. Insbesondere die zukünftige Anwendung Maschinellen Lernens (ML) und allgemein Künstlicher Intelligenz (KI) erfordert eine Aufbereitung der Dokumente, um im Nachgang digitale Tools darauf anwenden zu können.

Derzeit werden Daten aus Dokumenten häufig noch manuell in Folgesoftware übertragen. Dies ist fehleranfällig und zeitaufwändig, ermöglicht aber eine schnelle Auswertung der Informationen. Allerdings ist nicht jede wichtige Information auch in solchen technischen Systemen abgebildet. Dies hat zur Folge, dass sich die Suchzeiten bei der Auswertung von spezifischen Informationen, die nicht in Softwareprodukten erfasst, sondern nur auf den Dokumenten selbst vorgehalten werden, enorm erhöhen. Suchzeiten von bis zu zwei Arbeitsstunden für ein Objekt sind alltäglich.[1] Die Lösung für dieses Problem ist die maschinenbasierte Extraktion der entscheidenden Informationen aus den Dokumenten. Für eine zielgerichtete Extraktion müssen jedoch Voraussetzungen, insbesondere in der Texterkennung, Dokumentenaufteilung und Klassifizierung jedes einzelnen Dokuments erfüllt werden.

Aus diesem Grund widmet sich *ML-BAU-DOK* dem Lebenszyklus von Dokumenten, die im Rahmen der Bewirtschaftung von Gebäuden genutzt werden. Dabei werden die Prozesse der effizienten Digitalisierung, die Priorisierung der Dokumente nach Datenqualität und Maschinenlesbarkeit und die ordnungsgemäße Klassifizierung als Voraussetzung für die Informationsextraktion dargelegt. Zusätzlich werden Open-Source Algorithmen für eine automatisierte Segmentierung von großen Dokumentenmengen entwickelt und auf ihre Anwendung geprüft.

1.2 Zielsetzung und forschungsleitende Fragestellung

ML-BAU-DOK soll die Voraussetzungen zukünftiger maschinenbasierter Informationsextraktion aus gebäude- und anlagenbezogenen Dokumenten schaffen. Dazu sollen Regeln für die Digitalisierung von papierbasierter Gebäudedokumentation aufgestellt und speziell für die beiden Anwendungsbereiche Energieeffizienz- und Lebenszyklusanalysen definiert werden, Schlüsselinformationen dieser Anwendungsbereiche identifiziert und auf Basis dieser Informationen die Schlüsseldokumente und Dokumentenklassen erkannt werden, um diese abschließend zu priorisieren und die Dokumente den Klassen automatisiert zuordnen zu können. Die Ergebnisse sollen auf andere Anwendungsbereiche übertragbar sein. Für das Projekt wurden folgende Teilziele definiert:

1. Formulierung von Regeln für die effektive Digitalisierung von gebäude- und anlagenbezogenen Dokumenten als Grundvoraussetzung für Maschinelles Lernen.

[1] Vgl. Kyocera Document Solutions Deutschland GmbH 2018, S. 17.

2. Spezifikation der wesentlichen Dokumentenklassen für die Anwendungsbereiche Energieeffizienz und Lebenszyklusanalysen.
3. Priorisierung der Dokumentenklassen für die definierten Anwendungsbereiche, wobei die Methodik und Kriterien auf andere Anwendungsbereiche übertragbar sein sollen.
4. Darlegung der Möglichkeiten für (automatisches) Klassifizieren von gebäude- und anlagenbezogenen Dokumenten

Für die Erreichung der Teilziele wurden folgende begleitende Forschungsfragen als Kernbereiche von *ML-BAU-DOK* formuliert:

1. Welche Regeln müssen bzw. sollten für die effektive Digitalisierung von gebäude- und anlagenbezogenen Dokumenten eingehalten werden?
2. Welche wesentlichen Dokumentenklassen sind relevant für die spezifizierten Anwendungsbereiche Energieeffizienz und Lebenszyklusanalysen?
3. Wie können die Dokumentenklassen für die spezifizierten Anwendungsbereiche priorisiert werden, um Effizienz und Effektivität einer maschinenbasierten Informationsextraktion sicherzustellen?
4. Welche Möglichkeiten bestehen für das (automatische) Klassifizieren von gebäude- und anlagenbezogenen Dokumenten? Wie können diese künftig verbessert werden?

Auf diese Weise können folgende Kernthesen bewertet werden:

1. Für die optimierte Planung und Durchführung von (baulichen) Maßnahmen ist eine möglichst vollständige (und priorisierte) digitale Gebäudedokumentation grundlegende Voraussetzung.
2. Die Nutzung von Optical Character Recognition (OCR) (Optische Zeichenerkennung) und ML erlaubt eine teilweise automatisierte Informationsextraktion aus digitalen Dokumenten. Spezialisierten Technologieunternehmen im Bereich Dokumentenmanagement/Informationsextraktion und Unternehmen in der Bau- und Immobilienwirtschaft fehlt hierzu jeweils relevantes Wissen.
3. Die Projektergebnisse sind ein wichtiger Schritt zur Digitalisierung in der Bau- und Immobilienwirtschaft mit dem Einsatz von Maschinellem Lernen im Bereich Gebäudedokumentation.

Durch die Beantwortung der Forschungsfragen soll der Grundstein für weitere Forschungsprojekte im Bereich der digitalen Gebäudedokumentation und deren Auswertung gelegt werden.

ML-BAU-DOK gibt zusätzlich Aufschluss über die Anwendung des Massenscanverfahrens und zeigt Möglichkeiten der automatisierten Dokumentensegmentierung auf. Es beschreibt die Messbarkeit von Datenqualität mit einem System zur Integration in Unternehmensprozesse und abschließend Möglichkeiten der automatisierten Dokumentenklassifikation.

1.3 Vorgehensweise und Methodik

Mit *ML-BAU-DOK* werden Regeln für das grundlegende Vorgehen im Rahmen der Klassifikation von Gebäudedokumentation dargelegt. Dabei wird der gesamte Lebenszyklus der Gebäudedokumentation vom Posteingang bis zur Extraktion betrachtet.

Zu Beginn wird ein noch nicht digitalisierter Teil der Gebäudedokumentation aus dem technischen Facility Management des Praxispartners Stiftung Kloster Eberbach (SKE) digitalisiert und auf Basis dessen Regeln für die effiziente Digitalisierung von Gebäudedokumentation aufgestellt. Der Scan erfolgt entgegen der üblichen Praxis als Massenscan, um mit einer größeren Geschwindigkeit große Dokumentenmengen zu digitalisieren. Die Regeln der Digitalisierung werden anschließend mit dem Praxispartner Architrave/Property Care validiert.

Aufgrund der Nutzung des Massenscans wurde das Schreiben eines Segmentierungsalgorithmus nötig für die ordnungsgemäße Aufbereitung der Dokumente. Dieser Algorithmus ermöglicht die automatisierte Trennung von massenhaft gescannter Gebäudedokumentation.

Anschließend werden die Anwendungsgebiete Energieeffizienz- und Lebenszyklusanalyse definiert und auf wesentliche Schlüsselinformationen reduziert, um auf dieser Basis die Schlüsseldokumente ausfindig zu machen. Die Schlüsseldokumente wiederum ermöglichen die Abgrenzung der relevanten Dokumentenklassen.

Um die qualitativ hochwertigsten Daten identifizieren zu können, werden Methoden zur Überprüfung der Datenqualität von Dokumenten definiert. Darüber hinaus wird die Maschinenlesbarkeit verschiedener Dokumente aus dem Massenscanverfahren geprüft.

Abschließend werden unterschiedliche Möglichkeiten der Klassifikation von Gebäudedokumentation aufgezeigt und ausgewertet. Je nachdem, wie die Unterlagen zur Verfügung stehen, kann somit der geeignete Algorithmus zur Aufbereitung der Gebäudedokumentation verwendet werden.

1.4 Gang der Untersuchung

Die Untersuchung wird in sechs Arbeitspaketen durchgeführt, die sukzessive an der Vorgehensweise und Methodik orientiert sind. Die genaue inhaltliche Abfolge der Arbeitspakete ist wie folgt:

1. Digitalisierung der Gebäudedokumentation und Formulierung von Regeln für die Digitalisierung
2. Definition Schlüsselinformationen der Anwendungsbereiche Energieeffizienz- und Lebenszyklusanalysen
3. Identifikation der Dokumente und Dokumentenklassen auf Basis der Schlüsselinformation
4. Priorisierung der Dokumente und Dokumentenklassen auf Basis der Datenqualität und Prüfung der Maschinenlesbarkeit
5. Automatisierung der Klassifikation durch Anwendung der Algorithmen
6. Dokumentation und Dissemination der Ergebnisse

Entlang der dargestellten Arbeitspakete wird der Bericht strukturiert und sukzessive dargelegt.

2 Digitalisierung von Immobilienbestandsdokumenten

Im Laufe des Lebenszyklus einer Immobilie fallen zahlreiche Dokumente an. Daher ist es wichtig, sie in einer bestimmten Ordnung und Systematik aufzubewahren und zugänglich zu machen. Dieses Kapitel behandelt den Umgang mit Papierdokumenten und deren dauerhafte Digitalisierung für die wissenschaftliche und praktische Weiterverwendung.

2.1 Dokumente als Basis der Immobilienwirtschaft

Die Menge der produzierten Informationen war noch nie so groß wie heute, dennoch besteht die Gefahr des Verlustes historischer Erinnerung.[2] Dies gilt auch für Datenerfassung und Dokumentation in der Bau- und Immobilienwirtschaft, wo Dokumente und Daten im Lebenszyklus von Gebäuden oft erst nach Jahren oder Jahrzehnten wieder benötigt werden. Die Menge an Informationen führt nicht zwangsläufig zu einer qualitativen Steigerung des Wissens: Dafür muss der Umgang mit der Menge an Informationen konsequent geregelt werden und Personen müssen in der Lage sein, die Informationen richtig zu nutzen.[3] Die Pflege und Weitergabe relevanter Informationen auch über Generationen hinweg ist Aufgabe von Archiven, die mehr und mehr digitalisiert werden sollten.[4]

Bei einer Lebensdauer von 100 Jahren besteht eine Immobilie über vier Generationen. Der Informationswert älterer, aber wichtiger Dokumente muss durch umfassende systemische, technische und organisatorische Maßnahmen optimiert und erhalten werden.[5] Die Archivierung bestehender Dokumente und die Ablage von Neudokumentation müssen einem geregelten System unterliegen, um die Nutzbarmachung der Daten zu optimieren und den individuellen Einfluss von Institutionen, Organisationen und Personen auf den Dokumentenbestand zu minimieren.[6] Dazu werden in *ML-BAU-DOK* Probleme untersucht und Regeln für die Digitalisierung aufgezeigt.

Die Archivierung der Dokumentation wird sich grundlegend ändern.[7] Zukünftig müssen elektronisch erstellte Dokumente bereits bei Fertigstellung in der dafür vorgesehenen Ablage platziert werden. Hierdurch können Archive in Echtzeit entstehen ohne eine aufwändige Nachbearbeitung durch Digitalisierungsprozesse. Durch die Umwandlung der Dokumente in ein Langzeitarchivierungsformat und Auswertung und Extraktion relevanter Informationen lässt sich eine maximale Auskunftsfähigkeit aus Dokumenten erzielen. *ML-BAU-DOK* betrachtet papierbasierte archivierte Dokumente sowie bereits digitalisierte Dokumente, die als Originalquellen in den Betreiberprozess integriert werden. Je nach Liegenschaft kann es sich um kleine bis sehr große Ordnerbestände handeln, deren Auskunftsfähigkeit durch unstrukturierte, papierbasierte Ablage sehr gering ist. Demnach braucht es ein Regelwerk, das aus bereits archivierten heterogenen Dokumentenmassen durch Digitalisierung eine homogene auskunftsfähige Dokumentation macht.

2.2 DMS als digitales Archiv

Die Dokumente des Praxispartners SKE befinden sich in einem Dokumentenmanagementsystem (DMS), das den simultanen ort- und zeitunabhängigen Zugriff auf die Datenbasis erlaubt. Weiterführende Disziplinen eines gängigen DMS, wie Prozessorganisation durch ein workflowfähiges Datenmanagement, werden in dem vorliegenden Nachschlagesystem nicht genutzt.[8] Es ist damit ein elektronisches Archiv im Sinne der digitalen Langzeitarchivierung und dient als Nachschlagewerk für die Liegenschaftsverwaltung. Deshalb wird

[2] Vgl. Lin, Wang, Li et al. 2019, S. 197.

[3] Vgl. Bawden et al. 2009, S. 184.

[4] Vgl. Schroeder 2006, S. 34.

[5] Vgl. Krcmar 2005, S. 54.

[6] Vgl. RICS 2017, S. 6.

[7] Vgl. Reimann 2004, S. 209.

[8] Vgl. Götzer, Maier, Schmale et al. 2014, S. 9–11.

nachfolgend der Fokus auf den Aufbau, die Organisation, den Nutzen und die Pflege von digitalen Archiven gelegt. Im Rahmen von *ML-BAU-DOK* bezieht sich der Begriff ‚Archiv' auf die digitale Archivierung großer Mengen von Dokumenten aus dem Lebenszyklus einer Immobilie und deren optimierte Nutzung.

Archive werden weitläufig als Orte assoziiert, an denen unterschiedlichste Sachgegenstände aufbewahrt oder gesammelt werden. Im Kontext von *ML-BAU-DOK* wird der Begriff ‚Archiv' als *„Zusammenfassung bestimmter Archivalien oder Archivbestände"*[9] aufgefasst. Archive existieren in diversen Formen in jeder Organisationseinheit, unabhängig von deren Größe oder Historie. Je nach Organisation kann der Inhalt eines Archivs grundlegend unterschiedlich sein. Der ‚Ort' eines Archivs wurde bisher als räumlicher Bereich betrachtet, der sich häufig innerhalb oder in der Nähe der Organisation befindet, um einen jederzeitigen Zugang zur Einsichtnahme zu ermöglichen. Seit geraumer Zeit und insbesondere gegenwärtig werden Archive durch elektronische Datenverarbeitung (EDV) unterstützt, wodurch eine Verlagerung des räumlichen Archivs in ein abstrakteres digitales Archiv, beispielsweise auf Datenträgern oder in einer Cloud, möglich ist.

Der Umgang mit Archivgut umfasst das *„erfassen, ordnen, verwahren, betreuen und erschließen"*[10] von Dokumenten und Informationen, unabhängig davon, ob das Archiv räumlich oder digital abgebildet wird. Fraglich ist jedoch, was tatsächlich als Archivgut geführt werden muss. Die Entstehung eines Archivgutes ist auf ein dokumentverursachendes Ereignis zurückzuführen. Hierdurch entstehen Archivale, die in ihrer Konstitution und Aussagekraft einmalig sind.[11] Diese können in der Bau- und Immobilienwirtschaft verschiedene Dokumentenformate wie Textdokumente (Schriftverkehr), Tabellendokumente (Kalkulationen) oder Abbildungen (Pläne) sein.[12] Die Sinnhaftigkeit der Aufbewahrung entsteht jedoch nicht allein durch das Ereignis selbst. Eine Archivierung ist aus zweierlei Gründen erforderlich. Zum einen, um einen verpflichtenden/zwangsläufigen nachträglichen Zugriff auf das Dokument zu ermöglichen, um beispielweise Ansprüche aus dem Dokument geltend machen zu können (z.B. im Zuge von Bauleistungen innerhalb des Gewährleistungszeitraums). Zum anderen dienen die Dokumente als Quelle für Vorhaben, also als Dokumentation eines Zustandes und dessen zeitliche Veränderungen (z.B. baulicher Substanzänderungen, im Sinne der Auskunftsgebung).[13]

Durch die digitale Ablage der Dokumente ergeben sich zahlreiche Vorteile, die den Einsatz eines DMS als Archiv unverzichtbar machen. Hierzu gehören in erster Linie die Revisionssicherheit der Daten durch einfache Ablage im digitalen Archiv und der damit einhergehenden Langzeitarchivierung. Zudem wird der unternehmensübergreifende Austausch von Dokumenten beschleunigt. Allerdings muss auch der unternehmensinterne Prozess beschleunigt werden, da die Erstellung neuer Dokumente die Ablagekapazität übersteigt. Dies führt zu unkonventionellen, nutzerspezifischen Ablagekonzepten und langfristigen Hemmnissen bei der Nutzung von Dokumenten.[14] Durch eine digitale Archivierung wird ein ort- und zeitunabhängiger Zugriff auf die Dokumentation ermöglicht. So wird nicht nur ein simultaner Zugriff auf die Dokumente, sondern auch eine schnellere Verfügbarkeit geschaffen.[15] Neben der optimierten Archivierung finden sich entscheidende Kostenersparnisse durch die Reduzierung von Lagerflächen papierbasierter Ablageordner und durch Steigerungen der Arbeitsgeschwindigkeit.[16] Gleichzeitig steigt die Nutzerzufriedenheit bei der Anwendung unterstützender digitaler Systeme.[17] Der größte Nutzen liegt jedoch in der Reduzierung der Suchzeiten des Personals um 20 % und der Maximierung der

[9] Vgl. Reimann 2004, S. 19.

[10] Vgl. Reimann 2004, S. 20; vgl. Ugale et al. 2017, S. 217–218.

[11] Vgl. Reimann 2004, S. 23.

[12] Vgl. Müller et al. 2021, S. 105.

[13] Vgl. Reimann 2004, S. 21.

[14] Vgl. Ugale et al. 2017, S. 217; vgl. Götzer, Maier, Schmale et al. 2014, S. 340.

[15] Vgl. Götzer, Maier, Schmale et al. 2014, S. 269.

[16] Vgl. Ugale et al. 2017, S. 217.

[17] Vgl. Sprague 1995, S. 36.

Auskunftsfähigkeit aus der vorhandenen Dokumentation für neu entstehende Immobilienmanagementprozesse.[18]

2.3 Dokumentenklassen nach dem Pertinenzprinzip

Neben der Digitalisierung der Dokumente ist das Ordnungssystem des Archivs von entscheidender Bedeutung. Bei der Archivierung von Gütern werden zwei gegensätzliche Formen unterschieden, das Provenienz- und das Pertinenzprinzip. Das Provenienzprinzip beschreibt die Zusammenführung der Archivgüter auf den gemeinsamen Ursprung/Herkunft. Pertinenzprinzip bedeutet die Zusammenführung verschiedener Archivgüter unabhängig von deren Herkunft nach einem betreffenden sachlichen Umstand. Die Bildung einer Struktur kann durch organisatorische Maßnahmen wie Kennzeichnungen, Ordnungen und einen Aktenplan hergestellt werden.[19] Das Pertinenzprinzip wurde in historischen Archiven überworfen, da die entworfenen Systematiken nicht zeitlos waren.[20]

Das Pertinenzprinzip ist die bevorzugte Archivierungsmethodik in Unternehmensdokumentationen und auskunftsgebenden Systemen, insbesondere um Zusammenhänge eines Betreffs zu bilden.[21] Diese Ordnung nach Betreffen ist auch in gängigen Verwaltungssystemen des Immobilien- und Facility Managements zu finden. Beispielsweise ordnet Computer-Aided Facility Management (CAFM) Software häufig nach DIN 276 oder GEFMA-Richtlinie 198. Somit sollte die Archivierung von Dokumenten durch ein Ordnungssystem gestützt werden, das die Massendokumentation mit einer einheitlichen, konsequenten und idealerweise zeitlosen Systematik überzieht.

Bei der Digitalisierung von Massendokumentation eines Archivs stellt sich die Frage, welches Ordnungssystem angewandt werden sollte. Vorab sollte hinterfragt werden, ob die räumliche archivische Ordnung übernommen werden kann, oder ob eine neue systematische Ordnung im Rahmen der Digitalisierung erstellt werden sollte. Es gibt unterschiedliche Möglichkeiten der Ordnung, die je nach Geschäftsbereich differieren können. Für die systematische Ablage von Geschäftsdokumenten sind insbesondere numerische, alphanumerische und stichwortbezogene Konventionen geläufig, wobei die Ablage stets einer zeitlichen/chronologischen Systematik unterliegen sollte.[22] Diese systematische Ablagemethodik erhöht den Pertinenzwert über ein System und bietet idealisiert prozessunterstützende Daten, die wiederum einen Teil neuer prozessgenerierender Daten bilden können und Einfluss auf dessen Evidenzwert haben.[23] Das gewählte Ordnungssystem/Aktenplan sollte stets in einer Unternehmensrichtlinie ausformuliert werden und sich an Normen des Dokumentenmanagement, wie DIN EN ISO 11442, DIN EN 61355-1, DIN EN 62023, DIN EN 82045 und VDI 4500, orientieren.

Die Dokumentenbasis der SKE im Rahmen von *ML-BAU-DOK* lag in einem räumlichen Archiv ohne erkennbare Systematik vor. Im Zuge der Digitalisierung wurden die einzelnen Aktenordner und deren zuordenbarer Inhalt durch eine händische Sichtprüfung einer systematischen Ordnung unterzogen, die in einer Dokumentationsrichtlinie festgelegt wurden.

Als entscheidende Informationen für das Auffinden der Dokumente wurden das Gebäude und die Dokumentenkategorie gewählt. Die Kurzbezeichnung des Objektes wurde durch eine abkürzende Buchstabenkombination des jeweiligen Gebäudes bestehend aus drei Ziffern erstellt (z.B. BIB = Bibliothek). Die Dokumentenkategorien wurden an die Dokumentenklassen der GEFMA-Richtlinie 198 angelehnt. Die Namenskonventionen des einzelnen Dokuments sind in alphanumerischer Form und setzen sich aus einem Buchstaben von A-I und mindestens einer Zahl von 1-9 zusammen. Die Zahlen können zudem noch Unterkategorien von 1-9 bilden (z.B. D2 = Dienstleistungsverträge; D2.1 =

[18] Vgl. Schroeder 2006, S. 37–39.

[19] Vgl. Reimann 2004, S. 26–27.

[20] Vgl. Reimann 2004, S. 23–24.

[21] Vgl. Schmude et al. 06.2020, S. 11–12.; vgl. Diaz-Bone, Weischer 2015, S. 417.

[22] Vgl. May 2011, S. 15–16.

[23] Vgl. Keller 2014, S. 253–254.

Anfragen/Leistungsbeschreibungen). Die Dokumentationsrichtlinie der SKE in Kombination mit der Dokumentationsklasse der GEFMA-Richtlinie 198 bilden somit das Regelwerk für die digitale Indexierung und das Tagging der Bestandsdokumentation anhand einer festgelegten Systematik. So blieb die Ordnung unverändert und wurde lediglich um eine weiterführende technologische Ordnung durch die Möglichkeiten des DMS erweitert. Bei dieser weiterführenden Ordnung handelt es sich um die Indexierung des Dokumentenbestandes nach dem vorgenannten Ordnungssystem. Eine Anpassung der Ordnung des Archivs an eine möglicherweise veränderte zukünftige Ordnung ist demnach nicht notwendig, wenn die unterschiedlichen Ordnungssysteme unternehmensintern festgelegt werden.

Im Immobilienmanagement existieren standardisierte Dokumentenklassen, entworfen durch die Gesellschaft für immobilienwirtschaftliche Forschung (gif). Die Richtlinie DMS beinhaltet die normierten Dokumentenklassen, nach denen das operative Immobilienmanagement, die digitale Dokumentenablage in Datenräumen oder DMS organisieren kann. Der Index besteht aus vielzähligen Klassen, in die Dokumente mit gleichen Inhalten und Eigenschaften zusammengeführt werden können. Die Zusammenführung von Dokumenten gleicher Art beschleunigt die Auswertung und Anwendung der Dokumente durch zielgenaues Filtern und Suchen der Dokumente.

In dem nachfolgenden Punkt wird detaillierter auf die Indexierung eingegangen.

2.3.1 Indexierung

Ordnungssysteme in DMS werden in der Regel durch Indexierung hergestellt. Durch Indexierung wird in einem DMS die Ordnerstruktur eines konventionellen Aktenordners ersetzt und somit eine detailliertere Suche durch manuelles Vorarbeiten innerhalb der digitalen Struktur ermöglicht. Hierdurch wird der Suchaufwand des DMS durch Filterfunktionen oder Ordner erweitert und die Erfolgsquote der Suche deutlich erhöht.[24] Der Index umfasst dabei ein vorgefertigtes und kontrolliertes Vokabular (Thesaurus), das an den Wirtschaftszweig, das Unternehmen und nutzende Personen angepasst wird und in der Summe einen Aktenplan ergibt. Im Immobilienmanagement werden international verschiedene Indizes, angewendet. Entgegen des Taggings sollte der Index ein festes und unveränderliches Statut in einer Unternehmensrichtlinie darstellen.[25] Die Indexierung der Dokumente kann unternehmensintern abgewickelt, bei eigenem Standard, oder an einen Dienstleister vergeben werden. Unternehmensinternes indexieren kann den Suchaufwand reduzieren und die Akzeptanz für das Ordnungssystem steigern.[26] Unabhängig davon, ob einzelne Dokumente mit einem eindeutigen Index verschlagwortet oder in ein entsprechendes Ordnersystem abgelegt werden, wird der Zweck der schnelleren Dokumentenauskunft erzielt.

2.3.2 Tagging

Das ‚Tagging' ist ein Hervorheben wichtiger Inhalte des Dokuments mittels Markierung und Verschlagwortung. Tagging ist eine Unterform der Indexierung, die ebenfalls eine Klassifizierung mit sich bringt, welche jedoch durch persönliche Kenntlichmachung entsteht.[27] Es unterliegt somit keinem einheitlichen, normierten und unveränderlichen Regelwerk. Aufgrund der individuellen Form sollte das Tagging nur als ergänzendes System zu einem klar definierten Index dienen und nicht durch gegenläufige Tags die Aussagekraft mindern.[28] Durch die Markierung des Dokuments mit einem ‚Tag' kann die Suchfunktion des DMS nochmals gestärkt und die Auskunftsfähigkeit gesteigert werden.[29] Bei benutzerdefinierten Kennzeichnungen muss jedoch immer die Veränderung des Indexes berücksichtigt werden. Bei einem ausgefeilten Index sollte daher auf eine individuelle Kennzeichnung verzichtet werden. Indexierung und

[24] Vgl. Ugale et al. 2017, S. 217.

[25] Vgl. Kipp, Campbell 2006, S. 17.

[26] Vgl. Schroeder 2006, S. 37.

[27] Vgl. Kipp, Campbell 2006, S. 3.

[28] Vgl. Kipp, Campbell 2006, S. 17.

[29] Vgl. Ugale et al. 2017, S. 218.

Tagging der Dokumente ermöglicht eine detaillierte Filterfunktion bei der Nutzung des DMS. Der Filter des DMS kann wesentlich gezielter angewandt werden, als wenn der gesamte Dokumentenbestand nach einzelnen Worten innerhalb der Volltextgenerierung suchen müsste. Hierdurch kann die Auskunftsfähigkeit des DMS maßgeblich gesteigert werden.[30] Indexierung und Tagging sind demnach Katalysatoren für ein erfolgreiches DMS und dienen der maximalen Auskunftskraft.

Die Indexierung schafft eine Zuordnung der Dokumentation zu Betreffen im Sinne des Pertinenzprinzips und bildet demnach eine detaillierte Kategorisierung des Datenbestandes. Diese Kategorisierung muss jedoch einer festgelegten Systematik folgen. Die individuelle Ablage einzelner Personen kann somit vermieden werden, wobei der Index durch individuelle Verschlagwortung um eine persönliche Ordnung ergänzt werden kann. Durch Volltextgenerierung, Indexierung und Tagging kann so aus einem heterogenen papierbasierten Dokumentenbestand eine homogene, auswertbare Dokumentensammlung entstehen.[31]

2.4 Scan und Volltextgenerierung

Grundvoraussetzung für die indexierte Ablage der Massendokumentation in einem DMS ist zunächst die Digitalisierung des Dokumentenbestandes. Diese erfordert die Einhaltung entscheidender Parameter (z. B. Farb-, Scanqualität, geringe Papierverschmutzung, etc.) für eine zukünftig hohe Auskunftsfähigkeit der digitalisierten Dokumente, um das Ziel der maximal möglichen Wiedergabe der relevanten Inhalte zu ermöglichen.[32] Die maximale Wiedergabefähigkeit lässt sich zum einen durch die Bereitstellung und Nutzbarmachung und zum anderen durch die Vernetzung der Informationen zu einer digitalen Infrastruktur messen.[33] Die reine Verbildlichung des Dokuments in Form eines Scans stellt dabei den geringsten Grad der Nutzbarmachung dar. Es sollte bei der Dokumentenbereitstellung bereits eine Selektion erforderlicher Dokumente geben. Dabei gilt der Grundsatz, eher mehr als weniger zu digitalisieren. Zudem sollte ein detaillierter qualitativer Ausschluss von Dokumenten erst in der digitalen Fassung geschehen.[34] Hierdurch wird eine Nutzung der Dokumente für tiefgreifende Analyseprozesse wie automatisierte Klassifikation oder Extraktion ermöglicht.

Bei einer Auflösung in Höhe von 300 dots per inch (dpi), einer Farbtiefe von 8 bis 24 Bit, dem idealen Abgleich (digitaler Zwilling) von Vorlage- und Ausgabeformat und der Speicherung der Dokumente in dem standardisierten PDF/A-Langzeitarchivierungsformat ist die digitale Sichtung gewährleistet.[35] Durch die Berücksichtigung vorgenannter Regeln können die entscheidenden Parameter an die Digitalisierung von Dokumenten, Qualität, Langlebigkeit, Interoperabilität, erbracht werden.[36] Eine Steigerung der Nutzbarmachung kann und sollte zudem durch die Volltextgenerierung geschaffen werden. Diese ist entweder durch die Anwendung einer OCR-fähigen Software, ein OCR-fähiges DMS oder durch das aufwändige Abschreiben/Transkribieren des Schriftgutes möglich.[37] Eine zusätzliche manuelle Korrektur der OCR kann ebenfalls vollzogen werden. Dies wurde im Rahmen der Dokumentendigitalisierung der SKE, aufgrund der geringen Scanqualität und der darauf basierenden OCR, vorgenommen. Die geringe Scanqualität war in der vorliegenden Dokumentenbasis auf altersbedingte Verschmutzungen der Dokumente zurückzuführen. Eine manuelle Nachkorrektur ist bei großen Massendigitalisierungen nur mit hohen zeitlichen und finanziellen Aufwendungen realisierbar. Eine Nachkorrektur besonders bei wichtigen Dokumenten (z.B. Urkunden) ist zu empfehlen, da Verschmutzungen, Schatten, handschriftliche Notizen oder unterstrichene

[30] Vgl. Ugale et al. 2017, S. 217.

[31] Vgl. Gödert, Lepsky, Nagelschmidt 2012, S. 36.

[32] Vgl. DFG 2016, S. 6.

[33] Vgl. DFG 2016, S. 11.

[34] Vgl. Schroeder 2006, S. 36.

[35] Vgl. DFG 2016, S. 15–22.

[36] Vgl. DFG 2016, S. 14.

[37] Vgl. DFG 2016, S. 34.

Wörter häufig Fehler in der Zeichenerkennung bewirken.[38] Bei der Ablage von Dokumenten in ein DMS oder in einen Datenraum ist zu beachten, ob das jeweilige System über eine eigene Volltextgenerierung verfügt. Dies kann Auswirkungen auf die Qualität der OCR und die Lesbarkeit der Daten haben.

Die Volltextgenerierung ist ein entscheidendes Element in der maximalen Güte von digitalen Dokumenten und sollte genauestens geplant werden. Sie entscheidet nicht nur über das Finden und Kopieren textlicher Inhalte eines Dokuments, sondern auch über die Analysefähigkeit ganzer Dokumentenbestände für tiefgreifende Ordnung, Klassifikationen und Extraktionen von bzw. aus Dokumenten.

2.5 Prozess der digitalen Archivierung

Digitale Dokumentation braucht ein digitales System für den Abruf der Dokumente, eine pertinente Ordnung in Form von Dokumentenklassen (Index) und das digitalisierte Dokument als Scan inkl. der Voraussetzungen einer Volltextgenerierung. Zudem bedarf es für die Umsetzung einer Dokumentendigitalisierungsstrategie auch der Beschreibung des Digitalisierungsprozesses. Nachfolgend wird der Digitalisierungsprozess chronologisch bearbeitet. Die nachfolgend beschriebene Prozessfolge wurde mit dem Unternehmen PropertyCare GmbH validiert.

Ausschreibung

Bevor der tatsächliche Prozess der Dokumentendigitalisierung beschrieben wird, muss zunächst auf die Ausschreibung einer Scandienstleistung eingegangen werden. Eine solche Ausschreibung erfordert Information über die ungefähre Seitenzahl der gesamten Dokumentation, die überwiegenden Formate der Dokumentation (DIN A4, DIN A3 oder groß- oder kleinformatigere Dokumente). Zudem ist eine Spezifizierung der Dokumentenstruktur in Tabellen-, Textdokumente und Plandokumenten (Abbildungen) notwendig. Für die Benamung der digitalen Dokumente im Anschluss an die Digitalisierung muss ein einheitliches Kennzeichnungssystem zugrunde gelegt werden. Bei einem ordnerweisen Massenscan kann eine Benamung entsprechend des Aktenordnertitels gewählt werden. Bei einem Einzeldokumentenscan erfordert die Benennung der einzelnen Dokumente mehr Fachwissen und eine Namenswahl bezogen auf das jeweilige Dokument.

Neben den zuvor genannten Bestimmungen bzgl. der Ausschreibung einer Scandienstleistung müssen grundlegende und spezifische Anforderungen an den Umgang mit Dokumenten festgelegt werden. Grundlegende Anforderungen erfordern Auskunft der bietenden Unternehmen über den gesicherten Umgang mit den Dokumenten, den Ort der Vollbringung der Dienstleistung, den Zeitrahmen für die Digitalisierung sowie Mechanismen zur Qualitätssicherung. Spezifische Anforderungen beziehen sich auf den tatsächlichen Digitalisierungsprozess und welche Leistungen durch den Dienstleister erbracht werden sollen. Insbesondere Aufgaben der Vor- und Nachbereitung der Dokumente werden hier beschrieben. Die Vorbereitung (Pre-Processing) erfordert das Glätten der Papiere, Entfernen der Heftungen, Klammern und Notizzettel, sowie die individuelle Anpassung des Scanners auf Vorlagenqualität, Papierstärke und Formate. Die Nachbereitung bezieht sich insbesondere auf die Rückführung der Dokumentation nach dem Scan in den Ausgangszustand. Ein weiterer wichtiger Aspekt der spezifischen Anforderungen ist die Beschreibung des Kennzeichnungssystems der digitalen Dokumentation. Abschließend werden in den Anforderungen die Parameter an den Scan wie in 2.4 beschrieben. Auch die Erstellung einer Volltextgenerierung mit oder ohne manuelle Korrektur erhält Einzug in die spezifischen Anforderungen, wenn diese beauftragt werden soll.

Vergabe

Nachdem die Anforderungen an die Digitalisierungsdienstleistungen festgelegt wurden, kann die Ausschreibung veröffentlicht werden. Die Ausschreibung der Leistung erfordert durch den Bieter die genaue Bepreisung der Druckformate (DIN A4, DIN A3, Großformat). Vergabekriterien können in Preis, zeitlicher Aufwand, Referenzen und Fachkunde, in absteigender Gewichtung definiert werden.

Für die Kalkulation sind die exakte Anzahl der Dokumente in den jeweiligen Formaten, die Beschaffenheit der Dokumente sowie die Information, ob die Dokumente einseitig oder beidseitig bedruckt sind, von besonderer Bedeutung. Zudem konnte festgestellt werden, dass die Paginierung der Dokumente für die Qualitäts- und

[38] Vgl. DFG 2016, S. 36.

Vollständigkeitsprüfung wichtig ist. Bei unpaginierten Seiten muss die Prüfung manuell vorgenommen werden, da die Vollständigkeit dann nicht anhand der Seitennummern festgestellt werden kann. Somit muss jede Seite einzeln mit dem Original abgeglichen werden. Ebenso ist entscheidend, mit welchem Scansystem (Dokumenten-, Flachbett-, Durchzugs- oder Aufsichtsscanner) die Dokumente digitalisiert werden sollen. Die Scansysteme bergen jeweils Vor- und Nachteile in der Bearbeitung und haben Auswirkungen auf die Qualität des digitalen Dokuments. Im Rahmen einer OCR-fähigen Dokumentendigitalisierung sollte die höchste Scanqualität erreicht werden, um ein maschinelles Auslesen der Dokumente zu ermöglichen.Im Anschluss an die Vergabe der Leistung an einen Scandienstleister erfolgt die tatsächliche Digitalisierung der Dokumente.Hierzu muss der Digitalisierungsprozess papierbasierter Dokumente betrachtet werden. Abbildung 1 beschreibt in sechs Schritten den Prozess der Digitalisierung eines papierbasierten Dokumentenarchivs durch Beauftragung eines Dienstleisters.[39] Moderne Technologien wie die automatisierte Dokumentensegmentierung und die Dokumentenklassifizierung finden hier noch keine Berücksichtigung. Die Darstellung des konventionellen Ablaufs soll im späteren Verlauf des Projekts vergleichend die Prozessverbesserungen durch die Algorithmen aus *ML-BAU-DOK* verdeutlichen.

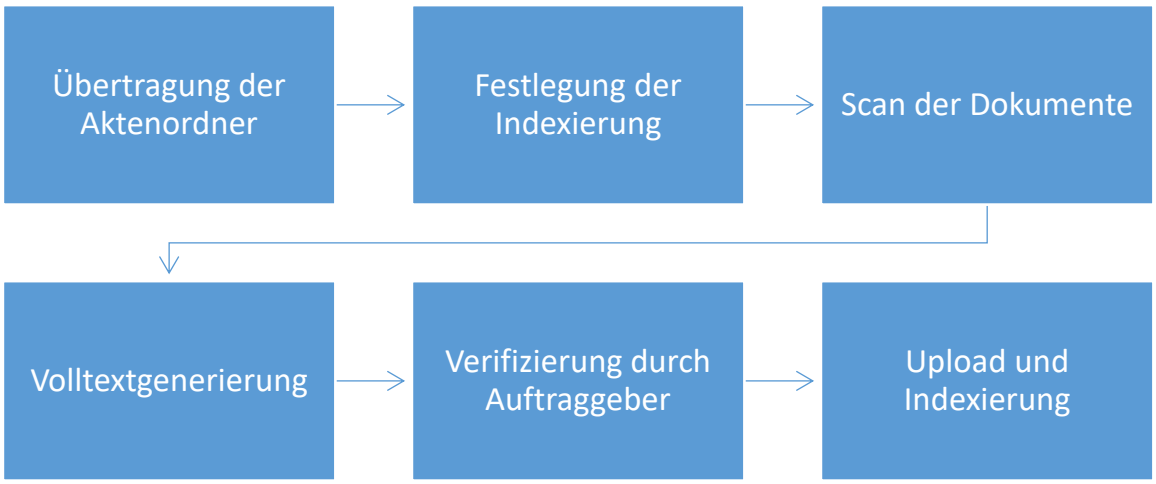


Abbildung 1: Prozess der Digitalisierung von Papierdokumenten[40]

1) Übertragung der zu digitalisierenden Ordner und Dokumente

Bei der herkömmlichen Vergabe der Scanleistung an einen Digitalisierungsdienstleister, werden zunächst die vorsortierten und ausgewählten Massendokumente von dem Auftraggeber an den Auftragnehmer übertragen. Die Digitalisierung kann in den Räumlichkeiten des Auftragnehmers oder des Auftraggebers vorgenommen werden. Dies ist insbesondere abhängig vom Auftraggeber und der Sicherheitsstufe der Dokumentation.

Die Erfassung der Dokumente ist abhängig von dem Umgang und der Nutzung der Dokumente. Bei einer Digitalisierung in wenige Dateien mit vielen PDF-Seiten, also der reinen Übertragung der Aktenordner in ein digitales Archiv, reduzieren sich Aufwand und Vorbereitung des Scans. Hierdurch können zeitliche und finanzielle Ressourcen eingespart werden. Allerdings mindert der Massenscan auch die Weiterverarbeitungsfähigkeit der Dokumente für eine automatisierte Klassifikation oder die Informationsextraktion. Die nachträgliche Separierung der Dokumente in Einzeldokumente ist auch unter Zuhilfenahme von Algorithmen nur schwer zu realisieren, da die Merkmale – Seitenzahlen, unvollständige

[39] Vgl. Ugale et al. 2017, S. 218; vgl. Reimann 2004, S. 209–210.; vgl. Schroeder 2006, S. 34–35.

[40] Eigene Darstellung.

Sätze – für eine Separierung nicht konsequent auf allen Dokumentenseiten vorhanden sind. Dennoch wurden in *ML-BAU-DOK* Segmentierungsalgorithmen programmiert, die je nach Dokumentenbasis Anwendung finden können (siehe 5.2). Alternativ kann ein aufwändigeres Pre-Processing durch manuelle Selektion der zu digitalisierenden Dokumente und deren Einzelscan erfolgen oder durch die Markierung von Aktenordnern mit einem einheitlichen Schema (Farbmarkierung, Barcode, Makro-Trennung) für eine algorithmusbasierte Auftrennung des Massenscans.[41] Das aufwändigere Pre-Processing ist zu empfehlen, da hierdurch auch der behutsame Umgang mit Stickern und wertvollen Urkunden sichergestellt werden kann. Zum anderen können Textstellen, die durch handschriftliche Markierungen oder Verschmutzungen von der OCR nur schlecht erfasst würden, durch Markierung für eine manuelle Nachbearbeitung der OCR ausgewählt werden. Ebenso sollten Dokumente, die bekanntermaßen schlecht lesbar, alt oder schlecht formatiert sind, für eine manuelle Nachbearbeitung markiert werden. Die markierten Dokumente werden dann mit größerer Sorgfalt bearbeitet. In *ML-BAU-DOK* kommt ein Massenscanverfahren zum Einsatz, das die Vorverarbeitung lediglich durch das Lösen von Heftklammern und das Heften der Dokumentenstapel aus dem Ordner berücksichtigt. Ziel ist es, ohne diesen aufwändigen Arbeitsschritt die Dokumentenmengen in Einzeldokumente zu trennen und anschließend klassifizieren zu können.

2) Analyse und Sortierung des Ordnersystems und der Dokumente (Festlegung der Indexierung)

Nach der Übertragung der Dokumente und Ordner werden die Dokumente analysiert, vorbereitet und vorsortiert. Durch Analyse der Ordnerstruktur werden Anhaltspunkte für eine spätere Indexierung geprüft. Sollte die Indexierung vorgegeben oder gar das Ordnersystem überworfen werden, müssen die Dokumente zusätzlich in die neue Ordnung gebracht und die Indexierung entsprechend aufgenommen werden.

Die Gebäudedokumentation sollte stets einer pertinenten Ordnung unterzogen werden (2.3). Allerdings ist die mehrschichtige Ordnung in einem gewöhnlichen Baumsystem nicht mehr notwendig und zielführend. Durch die Erstellung oder Nutzung eines eindeutig definierten/standardisierten Indexes (2.3.1) können die Dokumente, ohne eine aufwändige, ständig erweiterbare Baumstruktur, einer Ordnung (Dokumentenklassen) unterzogen werden. Eine Indexierung sollte einmalig festgelegt werden und nicht veränderbar sein. Es sollte in der Ergebnisfindung kein falsch-negativen Ergebnisse, aber auch möglichst wenige falsch-positiven Ergebnisse vorkommen. Falsch-negative Zuordnungen fehlen in der weiteren Bearbeitung und werden so ggf. übersehen. Falsch-positive Ergebnisse führen zu einer Erhöhung der Dokumentenmenge und damit ebenfalls zu vermeidbarem Aufwand. Eine zu weiche Verteilung des Indexes führt zu ungenauen Suchen und einer reduzierten Auskunftsfähigkeit des Systems. In 5.3 werden Methoden der Klassifizierung nach einem standardisierten Index dargestellt. Das Tagging der Dokumente (2.3.2) nach subjektiven Systemen sollte unterbunden werden, da es die Suchfunktion des DMS beeinträchtigt. Im Falle eines zusätzlichen Taggings sollte stets auf einen einheitlichen Thesaurus der Immobilienwirtschaft zurückgegriffen werden.

3) Scan der Dokumente

Im Anschluss an die Aufbereitung der Dokumente werden diese gescannt. Der Scanprozess sollte unter Berücksichtigung der vorab festgelegten Rahmenbedingungen vollzogen werden. Durch Stichproben kann die Qualität des Scans geprüft werden.

Die Digitalisierung von Dokumenten durch Scannen unterliegt bereits anerkannten Regeln der Technik, die eine digitale Nutzung ermöglichen. Es konnte durch dpi- und Bit-Tests im Rahmen von *ML-BAU-DOK* festgestellt werden, dass die Rahmenbedingungen der Deutschen Forschungsgemeinschaft (DFG) an den Scanprozess keine Anpassung benötigen. Somit ist ein Scan mit den genannten Parametern (200-300 dpi, 8-24 Bit, PDF) zu bevorzugen. Im Falle eines Massenscans müssen vor dem Scan die Trennungsmechanismen (Barcode, Farbmarkierung, Makro-Trennung) integriert werden. Erwartungsgemäß ist die automatisierte Texterkennung von handschriftlichen Erzeugnissen oder durch Verunreinigungen beeinflusste Buchstaben nicht möglich. Ein großer Nachteil bei Massenscans heterogener Dokumentenmengen ist die mangelnde Kontrolle beeinflusster oder verunreinigter Textstellen.

[41] Vgl. Schroeder 2006, S. 37.

4) OCR, Abschreiben oder Transkribieren der Dokumente = Fertigstellung der PDF inkl. Bild und Volltext

Sobald die Dokumente vollständig digitalisiert wurden, kann die Volltextgenerierung vorgenommen werden. Je nachdem, welche Vereinbarungen getroffen wurden, müssen die Scans entweder durch eine professionelle OCR-Software ausgelesen, manuell abgeschrieben oder transkribiert werden. Gängige OCR-Software ermöglicht die nachträgliche Korrektur des Volltextes, wodurch sich wesentlich bessere Resultate generieren lassen. Einige DMS verfügen über eine integrierte OCR-Funktion, die die Dokumente beim Hochladen in das System automatisch ausliest. Somit gilt es bei einer Digitalisierung zu beachten, ob ein solches System genutzt wird.

Um die Tauglichkeit eines Massenscans für gängige Dokumente der Bau- und Immobilienwirtschaft bewerten zu können, wurde eine digitalisierte heterogene Dokumentenmasse nach bekannten Fehlern der Texterkennung geprüft werden, um deren Auswirkungen zu messen. Ein häufiger Fehler der Texterkennung ist der Zeichenfehler, also die Verwechslung von Buchstaben und oder Zahlen und die damit einhergehende falsche Schreibweise eines Wortes. Gängige Zeichenfehler sind I – l; l – 1; B – 8; B – ß; c – e; rn – m; O – 0; b – 6; t – f.[42] Eine Prüfung der Maschinenlesbarkeit erfolgt in 5.1.2. Es konnte im Zuge der Analyse festgestellt werden, dass die OCR den Buchstaben O und die Zahl 0, sowie auch I – l als Synonym akzeptiert. Somit belasten Satzzeichenfehler die Auskunftsfähigkeit der OCR nur geringfügig und stellen keine schwerwiegende Beeinflussung des erkannten Textes dar. Neben der Prüfung der Satzzeichen wurde auch geprüft, ob es Unterschiede in der Erkennungsquote der entscheidenden Dokumente (Text-, Tabellendokumente, Pläne) im Immobilienlebenszyklus gibt. Die richtige Erkennung von Tabellen- und Textdokumenten ist gleichermaßen hoch, die Erkennung von Plandokumenten führte häufiger zu einer fehlerhaften Satzzeichenwahl.

Systeme der Volltexterkennung, die die Qualität des bildlichen Abbildes der PDF reduzieren oder verändern, können Auswirkungen auf die Weiterverarbeitungsfähigkeit eines Dokuments haben. Dies ist insbesondere dann der Fall, wenn die ursprüngliche Datei im Zuge der Volltextgenerierung in seiner Größe oder sogar in seiner Schrift verändert wird. Diese Veränderungsprozesse werden bei manchen Softwareprodukten angeboten, um den Speicherbedarf zu minimieren oder die Editierbarkeit des Dokuments zu ermöglichen. Die Senkung der Bildqualität sorgt für eine Reduzierung der Erkennungsquote des Textes und sollte daher vermieden werden.

5) Verifizierung des Scans durch Auftraggeber und Rückführung/Konservierung/Vernichtung der Originale

Wenn die Dokumente einschließlich der Inhalte vollständig erstellt sind, werden diese und die Scans zur Kontrolle an den Auftraggeber rückgeführt. Die Originale können im Anschluss an die Digitalisierung verwahrt oder vernichtet werden. In der Praxis wird dem Auftraggeber häufig eine ‚Schamfrist' zur Prüfung der Dokumentation gewährt. Nach Ablauf dieser Frist wird die Dokumentation in Papierform vernichtet, wenn innerhalb dieser Frist keine Einsichtnahme erfolgt ist. Dokumente von hohem Wert, insbesondere solche mit Schriftformerfordernis, wie Verträge und Urkunden, sollten in Papierform aufbewahrt werden.

Neben der Aufbewahrungsentscheidung der Dokumentation sollte der Auftraggeber die Qualität der OCR und der Digitalisate prüfen. Falls im Vorhinein durch den Auftraggeber entschieden worden ist, dass ausgewählte Dokumente nicht gescannt werden, so können diese durch den Dienstleister farblich getrennt in den jeweiligen Aktenordnern abgeheftet werden. Unabhängig davon muss der Auftraggeber die ordnungsgemäße Rückführung der Dokumentation prüfen.

6) Upload und Indexierung der Dokumente in DMS

Sobald der Auftraggeber die Digitalisate verifiziert hat, stehen diese für den Upload bereit. Durch den Upload in ein DMS können die Scans der vorgegebenen Indexierung unterzogen werden. Als Alternative bieten sich auch virtuelle Datenräume an, die sich in ihrer Funktion immer weiter an DMS annähern.[43] Eine digitale

[42] Vgl. Baierer et al., S. 74.

[43] Vgl. gif e.V. 2021.

Erschließung der Dokumentation sollte verfolgt werden. Bei der Wahl eines cloudbasierten Archivs ermöglicht eine zweckspezifische Einrichtung des jeweiligen DMS die Anpassung an das jeweilige Unternehmen. Sobald die Dokumente in die Cloudumgebung integriert wurden, können diese der geforderten Indexierung unterzogen werden. Abschließend steht das DMS als Massenarchiv für die Auskunftsgebung zur Verfügung.

Hindernisse des Digitalisierungsprozesses

Durch die Auswertung von Bieterrückfragen und interner Digitalisierungsverfahren im Rahmen von *ML-BAU-DOK* können nachfolgend Hindernisse entlang des Prozesses der Archivierung von Papierdokumenten beschrieben werden.

Das größte Hindernis stellt die OCR dar. Die Feinfühligkeit der Erkennung handelsüblicher OCR-Software, insbesondere bei Falten oder Verunreinigungen, reduziert die Auswertbarkeit der Dokumente. In der Praxis sind nachträglich Korrekturen üblich, erfordern jedoch große Personalkapazitäten. Weiterhin sind Heftungen aufwändig zu verarbeiten, da die Seiten einzeln über einen Flachbrettscanner digitalisiert werden müssen. Leerseiten in Ordnern und handbeschriebene Sticker sowie Notizen führen zu regelmäßigen Rückfragen während der Beauftragung, der Umgang mit solchen Einflüssen muss bereits im Rahmen der Ausschreibung festgelegt werden. Scans sollten stets in Leserichtung der Seite gemacht werden, da ansonsten die Anwendung der OCR eingeschränkt werden könnte.

Die auftretenden Hindernisse lassen sich entweder vertraglich oder durch ein aufwändiges Pre-Processing des Auftraggebers reduzieren bzw. vermeiden. Allerdings erfordert das Pre-Processing großen Personaleinsatz für die Durchsicht gesamter Dokumentenbestände. Demnach ist die Berücksichtigung in den Ausschreibungs- und Vergabedokumenten zur bevorzugen.

2.6 Exkurs: Umgang mit Papiergut

Bei einer Digitalisierung von Archiven oder auch einzelner Archivgüter wird die Trennung von Information und Informationsträger vollzogen. Hierdurch entsteht ein Widerspruch zum konventionellen Archivwesen, wonach Informationsträger und Informationsgut eine untrennbare Einheit bilden.[44] Demnach kann ein digitaler Zwilling im Langzeitarchivierungsformat gebildet werden. Fraglich ist, wie mit dem Papieroriginal umgegangen werden soll, denn eine hybride Haltung von Papier und E-Dokumenten stellt keine dauerhafte Lösung dar.[45] Gemäß Handelsgesetzbuch (HGB) müssen Geschäftsdokumente über einen Zeitraum von sechs bis zehn Jahren aufbewahrt werden. Durch die Novellierung der Grundsätze zur ordnungsmäßigen Führung und Aufbewahrung von Büchern, Aufzeichnungen und Unterlagen in elektronischer Form sowie zum Datenzugriff (GoBD) im Jahre 2020 können aufbewahrungspflichtige Papierdokumente künftig durch das ersetzende Scannen und die damit einhergehende Erstellung eines redundanten digitalen Zwillings vernichtet werden. Allerdings ist die vollumfängliche rechtliche Beweiskraft noch nicht gegeben. Zudem fehlen aktuell noch eindeutige Anforderungen an ein rechtssicheres ersetzendes Scannen.[46] Grundvoraussetzung ist jedoch, dass die Qualität und Vollständigkeit des Scans geprüft werden muss.[47] Aus den genannten Gründen ist die Akzeptanz für die tatsächliche Vernichtung der Originale gering.[48] Der digitale Zwilling sollte in seiner Detailtiefe mindestens die genannten Anforderungen erfüllen, um eine spätere ideale Nutzbarmachung der Daten innerhalb der Dokumente zu ermöglichen. Bei bleibender Skepsis müssen die Originale in dem bereits vorher genutzten Archiv oder in einem externen Archiv konserviert werden. Sollte man sich dennoch für ein

[44] Vgl. Reimann 2004, S. 22.

[45] Vgl. Krüger et al. 2016, S. 245.

[46] Vgl. Krüger et al. 2016, S. 242.

[47] Vgl. Krüger et al. 2016, S. 241.

[48] Vgl. Zink et al. 2015, S. 11.

ersetzendes Scannen entscheiden, sollten wichtige Dokumente wie Urkunden auf jeden Fall aufbewahrt werden.[49]

2.7 Folgerung: Regeln der Digitalisierung von Immobilienbestandsdokumenten

Die Regeln für die Digitalisierung von gebäude- und anlagenbezogenen Dokumenten sind abhängig von der weiterführenden Nutzung und den geplanten Anwendungsbereichen während des Gebäudebetriebs. In *ML-BAU-DOK* werden die Regeln auf Basis einer Massendokumentation erstellt. Nachfolgend werden die Regeln entlang des Digitalisierungsprozesses dargestellt, der bereits in 2.5 beschrieben wurde. Die Aufstellung wurde mit einem Marktführer für digitale Ablage- und Scanleistungen validiert und als richtig bewertet. Abbildung 2 wurde in Zusammenarbeit erstellt.

Das Wichtigste bei der Digitalisierung der Unternehmensdokumentation ist jedoch, dass die zukünftige Echtzeit-Archivierung bereits geplant und strukturell – durch eine Archivierungsrichtlinie – in die Organisation eingebunden ist. Der Archivierungsprozess wird im Rahmen von Neudokumentation auf den Kopf gestellt, denn die Indexierung muss bereits bei Generierung oder Eingang eines Dokuments feststehen. Sollte dies nicht der Fall sein, wird eine weitere Digitalisierung notwendig sein und die Akzeptanz beim Personal sinken. Ziel ist es somit, eine vollständige digitale Transformation für alle Folgeprozesse im Unternehmen zu definieren.

[49] Vgl. Krüger et al. 2016, S. 243.

Abbildung 2: Regeln der Digitalisierung von Immobilienbestandsdokumenten[50]

[50] Eigene Darstellung, validiert mit Property Care GmbH.

3 Analyse der Anwendungsbereiche und Schlüsselinformationen

In Arbeitspaket 1 von *ML-BAU-DOK* wurden Regeln für die Digitalisierung von Bestandsdokumenten aus dem Immobilienlebenszyklus dargelegt. Insbesondere die Indexierung des Dokumentenbestandes wurde als Parameter für Suchgeschwindigkeiten und maximale Auskunftsfähigkeit digitalisierter Dokumente erkannt. Um eine maximale Interoperabilität zwischen verschiedenen Anwendungsbereichen zu ermöglichen, sollte der Fokus auf standardisierten Dokumentenklassen liegen. Für die Spezifizierung der Indexierung der Anwendungsbereiche Energieeffizienz- und Lebenszyklusanalysen werden zunächst die jeweiligen Schlüsselinformationen und darauf aufbauend die Schlüsseldokumente analysiert. Durch die Definition der Schlüsseldokumente lassen sich die entscheidenden Dokumentenklassen definieren und für eine spätere Klassifikation spezifizieren.

3.1 Energieeffizienzanalysen

Die Energieeffizienz und die damit einhergehende Reduktion des Primärenergieverbrauchs sind der wesentliche Treiber bei der Umsetzung weltweiter Klimaziele.[51] Energieeffizienz ist einer der wichtigsten Betrachtungspunkte innerhalb der Bestandsdokumentation.[52] Anhand der Bestandsdokumentation lassen sich Schwachstellen und Modernisierungsbereiche identifizieren. Im Folgenden werden die Energieeffizienzanalyse im Gebäudesektor definiert und die wesentlichen Schlüsselinformationen für die weitere Klassifizierung analysiert.

3.1.1 Definition

Für die Analyse der Energieeffizienz im Rahmen von *ML-BAU-DOK* werden Informationen über den Energiebedarf bzw. den Energieverbrauch eines Gebäudes, dessen baulichen Zustand und den zur Energiebereitstellung eingesetzten Energieträger benötigt. Grundlage für die Vergleichbarkeit der Energieeffizienz von Gebäuden ist der Energieausweis. Dieser stellt die Energieeffizienz in den Maßstab einer umfassenden allgemeingültigen Betrachtung. Mit wenigen signifikanten Werten ist eine Auskunft über den energetischen Zustand eines Gebäudes möglich.

Hierfür müssen nach DIN 18599 alle technischen Anlagen und deren Verbräuche in Betracht gezogen werden, insbesondere Kühlung, Heizung, Lüftung, Warmwasser und Beleuchtung. Der Verbrauch wird entsprechend der Kennwerte des jeweiligen Energieträgers gemessen. Die Energieträger unterscheiden sich nach DIN 2010 in nicht leitungsgebundene (Holz, Öl, Kohle und Flüssiggas) und leitungsgebundene (Fernwärme, Strom und Gas).

Der reine Verbrauch eines Energieträgers gibt allerdings noch keine Auskunft über die Effizienz einer technischen Anlage. Eine Analyse wird erst über Bezugsgrößen relevant und vergleichbar. Demnach müssen neben den Verbrauchskennzahlen auch Daten über die bedienten Flächen und/oder den Zeitraum der Bereitstellung gegeben sein. Der Verbrauchskennwert nach VDI 3807 ermittelt flächenbezogene Kennwerte, indem der Energieverbrauch mit der Bezugsfläche des Gebäudes über einen Zeitraum von einem Jahr in Verhältnis gesetzt wird.

Der Energieausweis wird unterschieden in den Bedarfs- und den Verbrauchsausweis für Wohngebäude und Nichtwohngebäude. Gemäß Gebäudeenergiegesetz (GEG) müssen je nach Ausweisform unterschiedliche Informationen abgerufen werden. Die Anforderungen für einen Energieverbrauchsausweis können als Mindestmaß für Energieeffizienzanalysen dienen. Diese Daten können mit weiterführenden Daten angereichert werden.

[51] Vgl. BMWi 2019, S. 8.

[52] Vgl. Ramseier et al. 2020, S. 27.

Das Ziel der Energieeffizienzanalysen durch Ermittlung des Verbrauchswertes geht über die ökologische Bewertung eines Gebäudes hinaus. Es dient auch der Beurteilung der Verbräuche, dem Benchmarking und der Analyse von Nutzungsverhalten. Diese Erkenntnisse können gemäß VDI 3807 wiederum in ein Controlling einfließen, um ökonomische Schlussfolgerungen treffen zu können.

3.1.2 Parameter der Energieeffizienz

Die Schlüsselinformationen für Energieeffizienzanalysen ergeben sich aus den Angaben, die bei der Beantragung eines Energieausweises abgefragt werden. Grundlage ist das im Jahr 2020 novellierte GEG, das EnEV, EnEG und EEWärmeG in sich vereint. Nachfolgend werden die Schlüsselinformationen aus Sicht des Antragsstellers kurz dargestellt.

Tabelle 1: Schlüsselinformationen aus Energieausweisen für Wohn- und Nichtwohngebäude unterteilt in allgemeine Informationen, Energiebedarfsausweisinformationen und Angaben für den Energieverbrauchsausweis[53]

	Schlüsselinformation allgemeine Informationen (Seite 1) Wohngebäude		Schlüsselinformation allgemeine Informationen (Seite 1) Nichtwohngebäude
1	Gebäudetyp	1	Hauptnutzung / Gebäudekategorie
2	Adresse	2	Adresse
3	Gebäudeteil (bei Teilbewertungen)	3	Gebäudeteil (bei Teilbewertungen)
4	Baujahr Gebäude	4	Baujahr Gebäude
5	Baujahr des Wärmeerzeugers / Baujahr Übergabestation (Fernwärme)	5	Baujahr des Wärmeerzeugers / Baujahr Übergabestation (Fernwärme)
6	Anzahl Wohnungen	6	Nettogrundfläche=Netto-Raumfläche nach DIN 277 (nur beheizter/gekühlter Bereich)
7	Gebäudenutzfläche	7	wesentliche Energieträger für Heizung
8	wesentliche Energieträger für Heizung	8	wesentliche Energieträger für Warmwasser
9	wesentliche Energieträger für Warmwasser	9	Erneuerbare Energien Art & Verwendung
10	Erneuerbare Energien Art & Verwendung	10	Art der Lüftung
11	Art der Lüftung	11	Art der Kühlung
12	Art der Kühlung	12	Inspektionspflichtige Klimaanlage / kombinierte Lüftungs- Klimaanlage Fälligkeit Inspektion
13	Inspektionspflichtige Klimaanlage / kombinierte Lüftungs- Klimaanlage Fälligkeit Inspektion	13	Anlass der Ausstellung
14	Anlass der Ausstellung	14	Verbrauchs- / Bedarfsausweis
15	Verbrauchs- / Bedarfsausweis	15	Datenerhebung durch Eigentümer / Aussteller
16	Datenerhebung durch Eigentümer / Aussteller		

	Schlüsselinformation Energiebedarf (Seite 2) Wohngebäude		Schlüsselinformation Energiebedarf (Seite 2) Nichtwohngebäude
17	Treibhausgasemission berechnet aus Jahresprimärenergiebedarf	16	Treibhausgasemission berechnet aus Jahresprimärenergiebedarf
18	Endenergiebedarf dieses Gebäudes für Wärme, Heizung, Warmwasserbereitung, Kühlung, Lüftung, Beleuchtung	17	Primärenergiebedarf dieses Gebäudes für Wärme, Heizung, Warmwasserbereitung, Kühlung, Lüftung, Beleuchtung
19	Primärenergiebedarf dieses Gebäudes für Wärme, Heizung, Warmwasserbereitung, Kühlung, Lüftung, Beleuchtung	18	Primärenergiebedarf nach GEG § 15 Ist-Wert | Anforderungswert
20	Primärenergiebedarf nach GEG § 15 Ist-Wert | Anforderungswert	19	Mittlerer Wärmedurchgangskoeffizient
21	Energetische Qualität der Gebäudehülle Hi' Ist-Wert | Anforderungswert	20	Einhaltung sommerlicher Wärmeschutz (DIN 4108-2)
22	Einhaltung sommerlicher Wärmeschutz (DIN 4108-2)	21	Verfahren der Energiebedarfsberechnung
23	Verfahren der Energiebedarfsberechnung	22	Endenergiebedarf (unterteilt nach Energieträger und die einzelnen Verbrauchern)
24	Endenergiebedarf dieses Gebäudes	23	Endenergiebedarf Wärme
25	Anteil an Deckung v. Wärme- Kälteenergiebedarf / Anteil zur Pflichterfüllung	24	Endenergiebedarf Strom
26	Angaben zur Nutzung erneuerbarer Energien für Wärmeenergiebedarf	25	Angaben zur Nutzung erneuerbarer Energien für Wärmeenergiebedarf
27	Angaben zur Nutzung erneuerbarer Energien für Kälteenergiebedarf	26	Angaben zur Nutzung erneuerbarer Energien für Kälteenergiebedarf
28	Maßnahmen zur Einsparung (nur Neubau)	27	Gebäudezonen
		28	Maßnahmen zur Einsparung

	Schlüsselinformation Energieverbrauch (Seite 3) Wohngebäude		Schlüsselinformation Energieverbrauch (Seite 3) Nichtwohngebäude
29	Treibhausgasemission als äquivalente Kohlendioxidemission	29	Endenergieverbrauch dieses Gebäudes für Wärme, Heizung, Warmwasserbereitung
30	Endenergieverbrauch dieses Gebäudes für Wärme, Heizung, Warmwasserbereitung, Kühlung, Lüftung, Beleuchtung	30	Endenergieverbrauch dieses Gebäudes für Strom
31	Primärenergieverbrauch dieses Gebäudes für Wärme, Heizung, Warmwasserbereitung, Kühlung, Lüftung, Beleuchtung	31	Verbrauchserfassung
32	Endenergieverbrauch dieses Gebäudes für Wärme, Heizung, Warmwasserbereitung, Kühlung, Lüftung, Beleuchtung	32	Primärenergieverbrauch dieses Gebäudes
33	Verbrauchserfassung - Heizung und Warmwasser	33	Treibhausgasemissionen dieses Gebäudes (in CO2-Äquivalenten)
		34	Gebäudenutzung

[53] Eigene Darstellung.

Energieausweise bestehen, unabhängig von der Einstufung als Wohn- oder Nichtwohngebäude, aus fünf Seiten. Die erste Seite beinhaltet allgemeine Informationen über das Gebäude, die zweite Seite gibt für Neubauten oder umfangreiche Modernisierungen Auskunft über den Energiebedarf und die dritte Seite stellt den Energieverbrauch dar. Die Seiten vier und fünf bestehen aus Modernisierungsempfehlungen und allgemeinen Erläuterungen. Die Inhalte der verschiedenen Energieausweise auf den entscheidenden ersten drei Seiten werden in Tabelle 1 gegenübergestellt, besonders die Schlüsselinformationen, die bei Antragstellung zur Verfügung gestellt werden müssen. Die Inhalte sind in der Reihenfolge der Abfrage des formularisierten Energieausweises aufgelistet. Es handelt sich bei den einzelnen Zeilen um die Benamung der einzelnen Felder, die dazugehörigen Inhalte werden im nachfolgenden Punkt erläutert. Energieausweise für Wohn- und Nichtwohngebäude unterscheiden sich hauptsächlich durch die Abgrenzung zwischen Bedarf und Verbrauch. Beim Verbrauchsausweis entfallen die Umrechnung des Primärenergieverbrauchs in Treibhausgasemissionen, die Betrachtung von erneuerbaren Energien und die zonierte Erfassung des Energiebedarfs. Die Gebäudesubstanz wird nur beim Bedarfsausweis berücksichtigt.

ML-BAU-DOK soll die Grundlagen für die Erstellung einer digitalen Gebäudedokumentation aus Bestandsdokumenten schaffen, die bei Energieeffizienz- und Lebenszyklusanalysen unterstützt. Aus dem Energieverbrauchsausweis können einige Schlüsselinformationen entnommen werden. Auf Grundlage der Verbrauchsanalyse sind in gewissem Umfang Rückschlüsse auf nutzerindividuelle Einflüsse und auf die bauliche Struktur möglich.

Aufgrund der verfügbaren Gebäudedokumentation werden die Schlüsselinformationen aus Energieverbrauchsausweisen für Nichtwohngebäude betrachtet.

3.1.3 Priorisierung der Schlüsselinformationen

Nachfolgend werden die Schlüsselinformationen für Energieeffizienzanalysen aus den Pflichtfeldern des Energieverbrauchsausweises analysiert. Die Betrachtung des Energieverbrauchsausweises für Nichtwohngebäude umfasst die Seiten eins und drei des standardisierten Energieausweises (Tabelle 1). Für eine weiterführende Analyse wurde untersucht, welche Daten für das jeweilige Pflichtfeld erforderlich sind und auf welchen Normen und Richtlinien die Informationen basieren. Zudem wurden die Pflichtfelder reduziert auf die wichtigsten Schlüsselinformationen. Als Schlüsselinformationen für die Energieeffizienzanalyse wurden die Positionen des Energieverbrauchsausweises für Nichtwohngebäude in Tabelle 2 bestimmt.

Tabelle 2: Schlüsselinformationen des Energieverbrauchsausweises für Nichtwohngebäude[54]

Nr.	Schlüsselinformation allgemeine Informationen (Seite 1) Nichtwohngebäude
1	Hauptnutzung / Gebäudekategorie
2	Adresse
3	Gebäudeteil (bei Teilbewertungen)
4	Baujahr Gebäude
5	Baujahr des Wärmeerzeugers / Baujahr Übergabestation (Fernwärme)
6	Nettogrundfläche=Netto-Raumfläche nach DIN 277 (nur beheizter/gekühlter Bereich)
7	wesentliche Energieträger für Heizung
8	wesentliche Energieträger für Warmwasser
9	Erneuerbare Energien Art & Verwendung
10	Art der Lüftung
11	Art der Kühlung
	Schlüsselinformation Energieverbrauch (Seite 3) Nichtwohngebäude
29	Endenergieverbrauch dieses Gebäudes für Wärme, Heizung, Warmwasserbereitung
30	Endenergieverbrauch dieses Gebäudes für Strom
31	Verbrauchserfassung
32	Primärenergieverbrauch dieses Gebäudes
33	Treibhausgasemissionen dieses Gebäudes (in CO2-Äquivalenten)
34	Gebäudenutzung

In Tabelle 2 sind die Überschriften der erforderlichen Schlüsselinformationen aufgelistet. Zur Herleitung wurden zunächst die Schlüsselinformationen des Energieverbrauchsausweises für Nichtwohngebäude, dargestellt auf Seite 1 des Ausweises, betrachtet. Diese Schlüsselinformationen setzen sich aus allgemein beschreibenden Informationen, wie der Gebäudekategorie, dem Standort, dem Baujahr sowie Flächenangaben, zusammen. Des Weiteren werden die Energieträger für Heizung und Warmwasser und die Art und Verwendung von erneuerbaren Energien festgestellt. Zusätzlich müssen Angaben über ggf. eingebaute Anlagen zur Lüftung und Kühlung gemacht werden.

Die Schlüsselinformationen über den Energieverbrauch sind in Tabelle 2 ab Zeile 29 aufgelistet. Der Endenergieverbrauch für Wärme, Heizung, Warmwasserbereitung und Strom wird summiert und als Einzelaufstellung abgefragt. Zudem ist die Angabe von Primärenergieverbrauch, Treibhausgasemissionen und Art der Gebäudenutzung erforderlich. Die Schlüsselinformationen aus Tabelle 2 bilden die Grundlage zur Auswertung der Dokumentenklassen in 5.1.3.1.

Um Informationen in Dokumenten wiederfinden zu können, ist es notwendig, den Detaillierungsgrad der benötigten Informationen zu kennen. Hierzu ist es nützlich, die Betrachtung um die erforderliche Detailtiefe zu erweitern. Diese Erweiterung ist in Tabelle 3 dargestellt.

[54] Eigene Darstellung.

Tabelle 3: Erforderliche Detailtiefe und Nachschlagewerke für die Ermittlung der Schlüsselinformationen[55]

Nr.	Erforderliche Detailtiefe, inkl. erforderlicher Nachschlagewerke für Ermittlungen
1	Kategorie nach Bauwerkszuordnungskatalog
2	Straße, Hausnummer, PLZ, Ort
3	Gebäudeteil nach Teilungserklärung
4	Baujahr in Jahreszahlen
5	Baujahr in Jahreszahlen
6	Berechnung von A nach DIN 277
7	Holz, Öl, Kohle, Flüssiggas, Erdgas, Fernwärme, Strom (Mehrfachnennungen möglich)
8	Holz, Öl, Kohle, Flüssiggas, Erdgas, Fernwärme, Strom (Mehrfachnennungen möglich)
9	Solar-, Wind-, Bioenergie, Geothermie, Wasserkraft; erforderliche Angaben: Anteil in % Deckung | Pflichterfüllung
10	Fensterlüftung, Schachtlüftung, Lüftungsanlage mit Wärmerückgewinnung, Lüftungsanlage ohne Wärmerückgewinnung
11	Passive Kühlung, Gelieferte Kälte, Kühlung aus Strom, Kühlung aus Wärme
	Erforderliche Detailtiefe, inkl. erforderlicher Nachschlagewerke für Ermittlungen
29	kWh/m²*a erweitere Angabe ob WW und/oder Kühlung enthalten sind, Abgrenzung über Messwerte (falls nicht möglich über Berechnungsverfahren der Heizkostenverordnung); Berücksichtigung der Leerstands- und Witterungsbereinigung
30	kWh/m²*a erweiterte Angabe ob Zusatzheizung, Warmwasser, Lüftung, eingebaute Beleuchtung, Kühlung, Sonstiges inbegriffen Aufteilung der Verbräuche auf Verbraucher
31	Zeitraum (mind. 36 Monate zusammenhängender Zeitraum, jüngste nicht älter als 18 Monate) |Energieträger | Primärenergiefaktor | Energieverbrauch Wärme | Anteil Warmwasser | Anteil Kälte | Anteil Heizung | Klimafaktor | Energieverbrauch Strom
32	kWh/m²*a Multiplikation des Endenergieverbrauchs mit dem für den jeweiligen Energieträger zugehörigen Primärenergiefaktor aus Anlage 4 GEG
33	kg/m²*a Multiplikation des Primärenergieverbrauchs mit dem für den jeweiligen Energieträger zugehörigen Emissionsfaktor aus Anlage 9 Nummer 3 GEG
34	Gebäudekategorie | Flächenanteil | Vergleichswert Wärme | Vergleichswert Strom Vergleichswerte aus Bekanntmachung der Regeln für Energieverbrauchswerte und der Vergleichswerte im Nichtwohngebäudebestand vom 15. April 2021 (BAnz AT 03.05.2021 B1)

In Tabelle 3 ist die erforderliche Detailtiefe der Schlüsselinformationen inklusive der notwendigen Nachschlagewerke zur Ermittlung der jeweiligen Werte dargestellt. Die Reihenfolge gleicht der Abfolge in Tabelle 2. Demnach ist in intuitiven Zellen wie der Adresse oder dem Baujahr hinterlegt, welche Informationstiefe für die Erstellung des Energieverbrauchsausweises erforderlich ist. Bei Zellen, in denen tiefgreifendere Informationen notwendig sind, wurden die erforderlichen Werte integriert und deren Zuordnung zu bestimmten Nachschlagewerken dargestellt. Daraus ergeben sich Zusammenhänge zwischen der Gebäudekategorie und der Zuordnung zu einer Kategorie aus dem Bauwerkszuordnungskatalog. Die

[55] Eigene Darstellung.

Berechnungsmethoden unterscheiden sich nach Wohn- und Nichtwohngebäuden, weshalb der Hinweis auf die Berechnung nach DIN 277 ergänzt wurde. Weitere Nachschlagewerke finden sich in der Messung des Energieverbrauchs und den hierfür zu berücksichtigenden Regeln und Werken. Abschließend wird für die Berechnung des Primärenergieverbrauchs und der Treibhausgasemissionen auf die Anforderungen des GEG verwiesen.

Die tabellarische Darstellung zeigt, welche Informationen für die Erstellung eines Energieverbrauchsausweises benötigt werden und welche Beiwerke für die ordnungsgemäße Vervollständigung notwendig sind. *ML-BAU-DOK* basiert auf Bestandsdaten des Facility Managements. Der Fokus liegt auf der Ermittlung der Schlüsseldokumente und deren automatisierter Zuordnung. Auf die Leerstands- und Witterungsbereinigung der Endenergieverbräuche wird im Rahmen der Energieeffizienzanalyse verzichtet. Beide beziehen sich auf Flächen- und Verbrauchswerte, die bereits in der Messung der flächenspezifischen Verbrauchswerte aufgenommen werden. Aus Gründen der Einfachheit wird auf eine doppelte Erfassung verzichtet.

Tabelle 3 gibt detailliert vor, in welchem Umfang Daten bei einer automatisierten Informationsextraktion durch einen Algorithmus extrahiert werden müssen. Durch die Definition der Schlüsselinformationen können die Schlüsseldokumente und damit die Dokumentenklassen definiert werden. Bei der Anwendung des Algorithmus kann dann in der jeweiligen Dokumentenklasse, auf Basis der Schlüsselinformationen und -dokumente, nach den spezifischen Informationen gesucht und anschließend extrahiert werden.

Aufbauend auf die Energieeffizienzanalyse werden im folgenden Abschnitt die Schlüsselinformationen für den Anwendungsbereich der Lebenszyklusanalyse dargestellt.

3.2 Lebenszyklusanalysen

Nachfolgend werden die Bestandteile von Lebenszyklusanalysen definiert und deren zentrale Parameter dargestellt. Hieraus ergeben sich die prüfungsrelevanten Schlüsselinformationen, die dann als Indikator für die Klassenzuordnung dienen können.

3.2.1 Definition

Die Lebenszyklusanalyse kann nach unterschiedlichen Ansätzen vorgenommen werden, da keine einheitliche Definition und Herangehensweise existiert.

Beim klassischen Ansatz wird auch von einer Ökobilanzierung (engl. Life Cycle Assessment (LCA)) gesprochen. Dabei kann die Ökobilanz über den gesamten Lebenszyklus erstellt werden – „cradle to crave" bzw. „von der Wiege bis zur Bahre" – und so die aus der Materialität resultierenden Umweltwirkungen in die Bilanzierung miteinbeziehen. Auch Teilformen sind weit verbreitet. So können beispielsweise lediglich die Herstellungsphase und die daraus resultierenden Umweltwirkungen berücksichtigen werden – „cradle to gate". Sonderformen der Ökobilanzierung betrachten einzelne Wirkungskategorien, beispielsweise beschreibt der Carbon-Footprint lediglich die CO_2-Bilanzierung einer Betrachtungseinheit.[56] [57]

Ein weiterer Ansatz der Lebenszyklusanalyse ist die Betrachtung von LCA und der Lebenszykluskosten (engl. Life Cycle Costs (LCC)). Bei diesem Ansatz wird also zusätzlich zur Ökobilanzierung die ökonomische Komponente erfasst. Auch hier bestehen weitere Sonderformen, die lediglich Teilbereiche betrachten.[58]

Der dritte Ansatz ist die ganzheitliche Betrachtung der Lebenszyklusanalyse, auch ganzheitliche Bilanzierung oder Life Cycle Engineering (LCE) genannt. Diese umfasst LCA, LCC und soziale Gesichtspunkte nach dem Ansatz der Triple Bottom Line – den drei Säulen des Nachhaltigen Bauens.[59]

Für die weitere Bearbeitung in *ML-BAU-DOK* wurde eine Sonderform des klassischen Ansatzes gewählt. Die genaue Beschreibung der Sonderform sowie die Begründung ergibt sich aus der Erklärung der Systemgrenzen

[56] Vgl. Klöpffer, Grahl 2012, S. 1.

[57] Vgl. Rössig, S. 48.

[58] Vgl. König 2017, S. 97.

[59] Vgl. Fischer, S. 4–6.

und der Anwendung der Systemgrenzen auf *ML-BAU-DOK*. Die Systemgrenzen werden im weiteren Verlauf des Abschnitts näher betrachtet.

Nach dem klassischen Ansatz entspricht die Definition der Lebenszyklusanalyse in diesem Fall der Definition der Ökobilanz. Der Begriff „Ökobilanz" ist durch DIN EN ISO 14040 und 14044 definiert als:

„Zusammenstellung und Beurteilung der Input- und Outputflüsse und der potentiellen Umweltwirkungen eines Produktsystems im Verlauf seines Lebensweges" (*DIN EN ISO* 14040, S. 9)

Eine Ökobilanz ermöglicht die systematische Analyse einer Vielzahl von Umweltwirkungen, die direkt oder indirekt von einer Immobilie und den in ihr verbauten Materialien über ihren gesamten Lebenszyklus hinweg ausgehen.[60]

Aufgrund der internationalen Normung kann die Ökobilanz – im Gegensatz zu anderen Umweltbewertungsinstrumenten – durch ein standardisiertes Vorgehen durchgeführt werden und gewährleistet somit eine gute Vergleichbarkeit. Dabei sollte jedoch beachtet werden, dass die Tiefe und Breite von Ökobilanzen gemäß DIN EN ISO 14040 je nach gewähltem Untersuchungsrahmen, Zielsetzung und Datenqualität beträchtliche Schwankungen aufzeigen können.

In Abbildung 3 sind die Bestandteile und Zusammenhänge einer Ökobilanz nach DIN EN ISO 14040 und die Lebenswegmodule gemäß DIN EN 15804 über den gesamten Lebenszyklus einer Immobilie dargestellt.

Für die Bewertung bzw. die Erstellung einer Ökobilanz ist die Festlegung von Zieldefinitionen mit Systemgrenzen grundlegend. Die Systemgrenzen bestimmen die Prozesse, die für die weitere Bewertung berücksichtigt werden und bilden somit den Untersuchungsrahmen.[61]

Bei Neubauten umfasst die Systemgrenze den gesamten Lebenszyklus (Phase A1 bis C4), bei einem bestehenden Gebäude sämtliche Phasen der verbleibenden Nutzungsdauer nach DIN EN 15978.

Anschließend können alle relevanten Lebenszyklusphasen über die Lebenszyklusschritte analysiert werden und die daraus resultierenden Input- und Outputflüsse in der Sachbilanz (engl. life cycle inventory analysis – LCI) aufgeführt werden. Die Datenlage für diesen Prozessschritt sieht vor, dass hauptsächlich Primärdaten verwendet werden. Sollten keine Daten zur Verfügung stehen, kann auf Sekundärdaten (Durchschnittswerte) zurückgegriffen werden. Sollten diese Daten ebenfalls nicht vorliegen, müssen Schätzungen vorgenommen werden.

Im nächsten Schritt kann eine Auswertung auf Basis der Wirkungsabschätzung (engl. life cycle impact assessment (LCIA)) und der Zieldefinition erfolgen. Die Ergebnisse sollen anschließend als Basis für Schlussfolgerungen, Empfehlungen und Entscheidungshilfen gemäß DIN EN ISO 14040 zusammengefasst werden.

[60] Vgl. Brockmann et al. 2019, S. 5.

[61] Vgl. Fischer, S. 4.

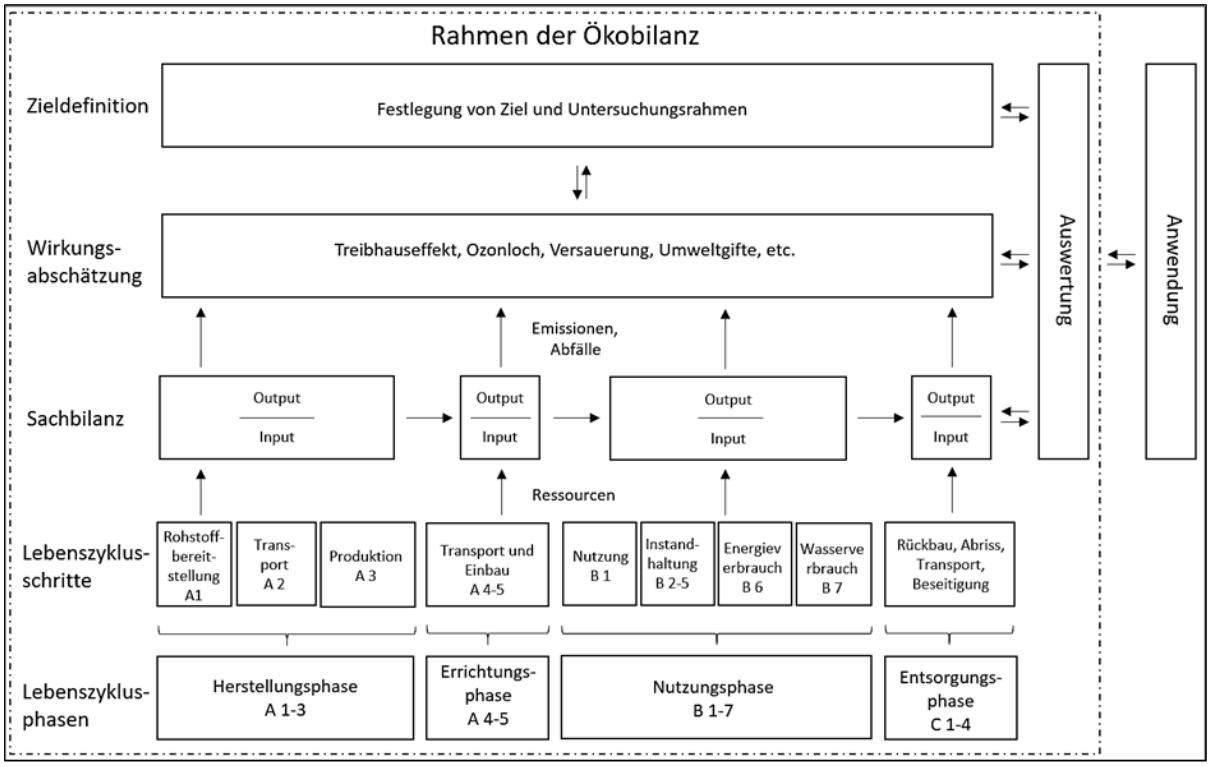


Abbildung 3: Bestandteile und Zusammenhänge einer Ökobilanz nach DIN EN ISO 14040 und DIN EN 15804[62]

Nachdem unter dem Begriff Lebenszyklusanalyse eine ganze Reihe an Verfahren und Vorgehensweisen aufgefasst werden, wurde mit Bezug auf die Ökobilanzierung nach DIN EN ISO 14040 und 14044 ein normiertes Vorgehen ermöglicht. Allerdings gibt es eine Vielzahl von Variationsmöglichkeiten und bei langlebigen Produkten müssen Annahmen und Szenarien getroffen werden, sodass die Ergebnisse nur bedingt vergleichbar sind. Erschwerend kommt hinzu, dass Umweltauswirkungen, auch Impact Kategorien genannt, nicht immer gegeneinander aufgerechnet werden können, z.B. der Wasserverbrauch in m^3/a nicht direkt gegen CO_2 Emissionen.

3.2.2 Parameter des Lebenszyklus

Der erste Prozessschritt einer Lebenszyklusanalyse besteht gemäß DIN EN ISO 14040 in der Festlegung von Zielen, Untersuchungsrahmen, Systemgrenzen, Untersuchungsgegenstand, Datenqualität und Detailierungsgrad. Die Systemgrenzen sollten so weit wie nötig und so eng wie möglich abgesteckt werden.[63]

Aus der digitalen Gebäudedokumentation sollen durch *ML-BAU-DOK* Daten und Maßnahmen für die Lebenszyklusanalyse abgeleitet werden können. Da die Lebenszyklusanalyse nicht für eine bestimmte Gebäudeart, sondern für jegliche Bestandsgebäude bestimmt werden soll, muss der Untersuchungsrahmen breiter und weniger detailliert gefasst werden.

Die relevanten Parameter einer Lebenszyklusanalyse lassen sich in die vier Lebenswegmodule nach DIN EN 15804 einteilen, beginnend mit der Herstellungsphase, der Errichtungsphase und der Nutzungsphase und endend mit der Entsorgungsphase (Abbildung 3). Da sich die Lebenswegmodule über Jahrzehnte erstrecken, ist eine Berechnung äußerst komplex und umfangreich. Weitere erschwerende Faktoren sind die mangelhafte Datenqualität bei Bestandsgebäuden und die eingeschränkte Datenverfügbarkeit, durch die viele Daten nur mit sehr viel Aufwand oder gar nicht erfasst werden können.[64] [65] So hat beispielsweise ein Gebäudefundament

[62] Eigene Darstellung.

[63] Vgl. Frischknecht 2020, S. 15.

[64] Vgl. May et al. 2022, S. 205.

[65] Vgl. Rodeck et al., S. 15.

einen erheblichen Einfluss auf die Ökobilanz.[66] Die genauen Mengen und Materialien lassen sich aber bei mangelhafter Gebäudedokumentation im Bestand nicht mehr nachvollziehen und oft nur schätzen. Ggf. müssen Szenarien mit Annahmen über die Mengen und die Auswirkungen gebildet werden.[67] Selbst bei bekannten Massen muss aufgrund des Mangels an verfügbaren herstellerspezifischen Datensätzen für Ökobilanzen oft auf generische Datensätze zurückgegriffen werden.[68]

Für *ML-BAU-DOK* wird die Nutzungsphase betrachtet, da diese Lebenszyklusphase bei Bestandsgebäuden diejenige Phase mit dem stärksten Einfluss auf den Gesamtenergieverbrauch und daher von größter Bedeutung ist.[69]

Die Nutzungsphase umfasst gemäß DIN 15804 die Lebenszyklusschritte Nutzung B1, Instandhaltung B2 bis B5, Energieverbrauch B6 und Wasserverbrauch B7. Die ersten fünf Nutzungsphasen beziehen sich auf die Bausubstanz, die Nutzungsphasen B6 und B7 auf den Betrieb des Gebäudes. Aufgrund der gewählten Systemgrenzen werden die Nutzungsphasen B6 und B7 betrachtet.

In dem Sinne ist die Lebenszyklusanalyse die periodische Auswertung der Energieeffizienzanalyse, und kann auch für Monitoring bzw. Benchmarking genutzt werden. Insbesondere können Entwicklungen der Verbräuche nachvollzogen und Ursachen dafür analysiert werden. So können Verbrauchstreiber identifiziert und durch Instandhaltung oder Modernisierung beseitigt werden.

3.2.3 Priorisierung der Schlüsselinformationen

Durch die gewählten Systemgrenzen sind die Schlüsselinformationen für die Lebenszyklusanalyse nahezu identisch mit denen für die Energieeffizienzanalyse (Abschnitt 3.1). In Tabelle 4 sind die Schlüsselinformationen für Lebenszyklusanalysen zusammengefasst.

Die Schlüsselinformationen der Lebenszyklusanalyse wurden gegenüber denen der Energieeffizienzanalyse um die Punkte 35 Menge Kaltwasser und 36 Menge Warmwasser erweitert. Die allgemeinen Informationen dienen zur eindeutigen Kennung des Gebäudes und der eineindeutigen Zuordnung der Daten. Angaben über Alter und Art der Wärmeerzeuger, Lüfter oder Kühler können im Benchmarking der Lebenszyklusanalysen bei der Auswertung berücksichtigt werden. Die Nettogrundfläche (NGF) dient der Bildung von Bezugsgrößen und ermöglicht das Benchmarking mit weiteren Gebäuden derselben Nutzungsart.

[66] Vgl. John et al. 2010, S. 341.

[67] Vgl. König, Kohler, Kreißig et al. 2012, S. 41.

[68] Vgl. DGNB 2018, S. 11.

[69] Vgl. DGNB 2018, S. 8.

Tabelle 4: Schlüsselinformationen Lebenszyklusanalyse[70]

	Schlüsselinformationen Lebenszyklusanalyse, allgemeine Informationen
2	Adresse
3	Gebäudeteil (bei Teilbewertungen)
4	Baujahr Gebäude
5	Baujahr des Wärmeerzeugers / Baujahr Übergabestation (Fernwärme)
6	Nettogrundfläche=Netto-Raumfläche nach DIN 277 (nur beheizter/gekühlter Bereich)
7	wesentliche Energieträger für Heizung
8	wesentliche Energieträger für Warmwasser
9	Erneuerbare Energien Art & Verwendung
10	Art der Lüftung
11	Art der Kühlung
	Schlüsselinformationen Lebenszyklusanalyse, Energieverbrauch
29	Endenergieverbrauch dieses Gebäudes für Wärme, Heizung, Warmwasserbereitung
30	Endenergieverbrauch dieses Gebäudes für Strom
31	Verbrauchserfassung
32	Primärenergieverbrauch dieses Gebäudes
33	Treibhausgasemissionen dieses Gebäudes (in CO2-Äquivalenten)
34	Gebäudenutzung
	Schlüsselinformationen Lebenszyklusanalyse, Wasserverbrauch
35	Menge Kaltwasser
36	Menge Warmwasser

Da hinsichtlich der Informationen über den Wasserverbrauch (Zeile 35 und 36) Übereinstimmung zwischen Energieeffizienzanalyse und Lebenszyklusanalyse besteht, kann Tabelle 3: *Erforderliche Detailtiefe und Nachschlagewerke für die Ermittlung der Schlüsselinformationen*Tabelle 3 um die Zeilen 35 und 36 aus Tabelle 5 ergänzt werden. Dadurch ist für alle relevanten Datenfelder die erforderliche Detailtiefe beschrieben.

Tabelle 5: Erforderliche Detailtiefe und Nachschlagewerk für die Ermittlung der Schlüsselinformation einer Lebenszyklusanalyse[71]

	Erforderliche Detailtiefe
35	m³/a
36	m³/a

[70] Eigene Darstellung.

[71] Eigene Darstellung.

3.3 Folgerung: Informationssymmetrien und Schlüsselinformationen zwischen Energieeffizienz- und Lebenszyklusanalysen

Die Anwendungsbereiche Energieeffizienz- und Lebenszyklusanalysen können weit gefasst und je nach Anwendung mit unterschiedlichen Detailtiefen beschrieben werden. In *ML-BAU-DOK* wurden beide zusammengefasst. Dabei basiert die Lebenszyklusanalyse auf der Energieeffizienzanalyse und erweitert diese um eine mehrperiodische Betrachtung im Sinne eines Monitorings oder Benchmarkings.

Die Energieeffizienzanalyse umfasst grundlegende Schlüsselinformationen des Gebäudes. Diese werden mit Informationen über die installierte Anlagentechnik und Verbräuche ergänzt. Durch die Erweiterung um relevante Bezugsgrößen, wie z.B. Flächen, kann die Energieeffizienz zeitpunkt- und zeitraumbezogen messbar gemacht werden. Dies ermöglicht den relationalen Vergleich von Gebäuden und die Analyse nutzungsbezogener Verbräuche. Auf Basis dieser Auswertungen können Kosten- und Verbrauchstreiber identifiziert und Maßnahmen zur Gebäudeoptimierung abgeleitet werden.

4 Dokumente der Energieeffizienz- und Lebenszyklusanalysen

Das folgende Kapitel beinhaltet die in Arbeitspaket 3 durchgeführten Analysen von Dokumentenklassen. Dabei werden Dokumente identifiziert, aus denen Schlüsselinformationen für die beiden Anwendungsbereiche extrahiert werden können. Neben Sinn und Nutzen von Dokumentenklassen werden die marktüblichen Klassifizierungssysteme aufgezeigt. Die Arbeit basiert auf dem Klassifizierungssystem nach Müller 2023, das insbesondere Dokumentenklassen der Technischen Due Diligence berücksichtigt. Das gewählte Klassifizierungssystem wird auf Vollständigkeit für die Anwendungsbereiche Energieeffizienz- und Lebenszyklusanalysen geprüft. Nach der Vervollständigung der Klassen werden diese durch Analyse der notwendigen Datenfelder priorisiert. Auf das Klassifizierungssystem wird dann der Algorithmus für die automatisierte Separierung und Klassifizierung von Dokumenten angewandt.

4.1 Dokumentenklassen als digitale Ordnungssystematik

Ordnungssystematiken in Wirtschaft und Verwaltung unterlagen bisher einer vorgegebenen hierarchischen Struktur. Diese Strukturen wurden in der Regel an individuellen Aufgabenprofilen orientiert und durch sich ändernde oder erweiternde Anforderungen angereichert. Diese konventionelle Art der Dokumentenablage wurde zeitweise auch in digitale Ablagesysteme übertragen. Durch die zunehmende Nutzung von cloudbasierten Ablagesystemen wandelt sich die hierarchische Ablagestruktur hin zu einer Ablagestruktur nach Dokumentenklassen.

Unter Dokumentenklassen versteht man die Zusammenfassung von Dokumenten mit gleichen inhaltlichen Attributen, d.h. eine Gruppierung von Dokumenten mit ähnlichen oder gleichen Inhalten. Dokumentenklassen können wiederum Teil eines Dokumentenindex sein, der eine hierarchische Strukturierung ausgewählter Klassen schafft und somit eine Rückführung in konventionelle hierarchische Strukturen darstellt. Die eineindeutige Zuordnung eines Dokuments zu seiner Klasse wird gemäß gif DMS 2.2 (gif-Richtlinie „Standard zum Aufbau eines Immobiliendatenraums und Dokumentenmanagementsystems (DMS)“, Version 2.2) sowohl durch Attribute als auch durch Metadaten ermöglicht. Metadaten beschreiben die organisatorischen Eigenschaften eines Dokuments.

Der Austausch von Dokumenten zwischen Dokumentenklassen unterschiedlicher Klassifizierungssysteme ist durch Mapping der jeweiligen Standards möglich. Durch Nutzung einer Zuordnungsmatrix sollte eine verlustfreie Übertragung gewährleistet werden können. Hierbei sollte jedoch beachtet werden, dass nicht jeder aktuelle Klassifizierungsstandard zwangsläufig mit anderen kompatibel ist. Ist dies der Fall kann ohne aufwändige manuelle Nacharbeitung kein automatisiertes Mapping stattfinden, da die Dokumentenklassen einen unterschiedlichen Detailierungsgrad aufweisen.

Die Anreicherung der Dokumente mit Attributen sowie die Zuordnung von Dokumenten zu Klassen ermöglicht die effiziente Such- und Filterfunktion aller Dokumente eines Datenraums oder DMS. Hierdurch lassen sich Suchzeiten reduzieren und Arbeitsabläufe effizienter gestalten.

4.2 Zusammenhang zwischen Klassen und Dokumenten

Zu Beginn der Etablierung von Dokumentenklassen wurde jedem Dokument im Immobilienlebenszyklus eine eigene Klasse zugeordnet. Hierdurch wurde eine eineindeutige Zuweisung von Dokument und Klasse ermöglicht. Die daraus entstehenden Dokumentenklassen wurden einer hierarchischen Ablagestruktur zugewiesen, die sich aus Aufgabenfeldern und Prozessen im Immobilienmanagement ergab.

Die neuesten Versionen von standardisierten Dokumentenklassen, hier das Beispiel gif-Richtlinie DMS 2.2, orientiert sich nicht mehr an hierarchischen Ebenen. Die Ordnung entsteht einzig durch die in einem Dokument befindlichen Daten. Diese Attribute und Metadaten ermöglichen die Analyse- und Suchfunktion der Dokumente über den Datenraum oder das DMS, ohne diese nach einer strengen vorgegebenen Hierarchie zu ordnen. Die Attribute werden jeder einzelnen Klasse über vier Dimensionen zugeordnet. Durch diese

Herangehensweise konnte die Anzahl an Dokumentenklassen von anfänglich mehr als 1.000 auf 192 Klassen reduziert werden. Nur hierdurch lässt sich die Zahl der Dokumentenklassen minimieren.

In Kapitel 3 wurden die Anwendungsbereiche Energieeffizienz- und Lebenszyklusanalyse beschrieben und die notwendigen Datenfelder (Attribute) der Bereiche identifiziert. Diese tabellarisch erfassten Attribute wurden anschließend als Grundlage für die Analyse der Dokumente genutzt. Die Dokumente wurden einzeln inhaltlich auf das jeweilige Datenfeld geprüft. Somit konnten jedem Dokument darin befindliche Attribute zugeordnet werden. Auf dieser Basis lassen sich die Dokumentenklassen nach der Quantität der darin vorkommenden Attribute priorisieren und die Kerndokumente für die Anwendungsbereiche identifizieren. Das Vorgehen ist nicht auf diese beiden Anwendungsfälle begrenzt und lässt sich auf weitere beliebige Fälle anwenden.

Der Ansatz in *ML-BAU-DOK* ist also konträr zum Ansatz der Dokumentenklassenbildung. In *ML-BAU-DOK* wurden die Dokumente anhand der benötigten Schlüsselinformationen priorisiert, während im Fachausschuss Dokumentenklassen anhand der Alltagsdokumentation geprüft wurde, welche Daten aus welchen Dokumenten generiert werden können.

Zusammenfassend kann festgehalten werden, dass in *ML-BAU-DOK* die relevanten Attribute aus den Anwendungsfällen abgeleitet und die Dokumente entsprechend klassifiziert und priorisiert wurden. Im allgemeinen Ansatz wurden die Attribute aus den Dokumenten extrahiert und zu Klassen zusammengefasst. Die unterschiedlichen Vorgehensweisen sind auf die unterschiedlichen Zielsetzungen zurückzuführen. Beide Vorgehensweisen haben ihre Berechtigung. Der umgekehrte Ansatz in *ML-BAU-DOK* stellt sicher, dass die relevanten Attribute jeweils aus den am besten geeigneten Dokumenten extrahiert werden.

4.3 Abgrenzung von Klassifizierungssystemen

Standards ermöglichen die marktübergreifende Einführung einheitlicher Dokumentenklassen über nationale Grenzen hinweg. Auch für die Klassifizierung von Dokumenten wurden bereits Standards entwickelt.

Alle Standards verfolgen das Ziel, Dokumente mit ähnlichen Attributen und Eigenschaften in einer Dokumentenklasse zu bündeln und das Vorgehen zu standardisieren. Dabei unterscheiden sich die jeweiligen Standards im gewählten Detaillierungsgrad und damit auch in der Anzahl der Dokumentenklassen, in der Strukturierung bzw. dem Aufbau sowie in der abweichenden Gliederung.

Ein Beispiel hierfür stellen die von der gif veröffentlichten Standards dar. Hier unterscheiden sich die Dokumentenklassenstandards gif-DMS 1.0 und die weiterentwickelte gif-DMS 2.2 sowohl von der Anzahl der Klassen als auch im Aufbau voneinander. Eine weitere Besonderheit ist, dass den nur noch 192 Dokumentenklassen der gif-DMS 2.2 jeweils Attribute zugeordnet werden können. So können innerhalb einer Dokumentenklasse Informationen hinzugefügt werden, ohne die Dokumentenklasse zu verändern.

Der Deutsche Verband für Facility Management (gefma) hat mit GEFMA 198-1 (2013) eine Richtlinie für die analoge Dokumentation im Facility Management veröffentlicht, die Handlungsempfehlungen zur lebenszyklusübergreifenden und ganzheitlichen Dokumentation für Immobilien gibt. Dabei wird ebenfalls eine Ordnungs- und Gliederungsstruktur vorgegeben, die als Dokumentenklassen aufgefasst werden können. Die Gliederung ist gröber und weist nur 95 Dokumentenklassen auf. Im Laufe von *ML-BAU-DOK* hat gefma eine neue Richtlinie 924-12 veröffentlicht, die Dokumenten- und Dateimanagement mit DMS und Attributen behandelt. Somit ist die Entwicklung hin zu einer attributebasierenden Dokumentenklassifizierung auch hier erkennbar.

Einen europäischen Ansatz stellt das open exchange Format, Format d'Inter-echanges de Données Juridiques et Immobilières, oder Financial and Property Data Interchange Format (FIDJI) dar. In diesem Standard gibt es 98 Dokumentenklassen, die ebenfalls hierarchisch strukturiert sind.

Eine Gegenüberstellung der beschriebenen Standards zeigt, dass sich die Dokumentenklassen nur teilweise 1:1 auf andere Standards übertragen lassen. Hierbei kommt es zu m:n-Beziehungen, sodass ggf. kein Mapping zwischen den unterschiedlichen Dokumentenklassen möglich ist.

Neben den bereits veröffentlichten Standards für die Dokumentenklassifikation ist damit zu rechnen, dass weitere Standards von bekannten Organisationen veröffentlicht werden, beispielsweise von der Royal Institution of Chartered Surveyors (RICS), der Society of Property Researchers (SPR), dem Investment Property

Forum (IPF) und der Vereniging Onroerend Goed Onderzoekers Nederland (Vogon). Von diesen werden erste Ansätze verfolgt, jedoch bisher ohne Veröffentlichung.

Eine weitere Klassifikation mit dem Schwerpunkt auf Dokumenten für technische Due Diligence wird durch Müller (2023) veröffentlicht. Der Aufbau dieser Klassifikation ist ebenfalls hierarchisch und unterteilt sich auf 419 Dokumentenklassen. Jeder Dokumententyp stellt eine eigene Dokumentenklasse dar. Durch die feine Gliederung ermöglicht die Klassifikation ein breites Mapping zu FIDJI, gefma und gif-DMS.

4.4 Folgerung: Anpassung des Klassifizierungssystems

Wie ursprünglich vorgesehen, wird für *ML-BAU-DOK* der Dokumentenklassifikationsindex nach Müller (2023) zugrunde gelegt. Der Dokumentenklassifikationsindex wurde an die beiden Anwendungsfälle und für die weitere Übertragbarkeit geringfügig angepasst. Zum einen wurden Dokumentenklassen hinzugefügt, die in der bisherigen Fassung noch nicht enthalten waren und für die weitere Bearbeitung der beiden Anwendungsfälle relevant sind und somit über die Dokumente einer TDD hinausgehen. Zum anderen wurden einzelne Dokumentenklassen entfernt, die im Hinblick auf die spätere automatisierte Klassifikation nicht eineindeutig bestimmbar sind oder über die verschiedenen Kategorien hinweg Dopplungen darstellten.

Die Gesamtaufstellung des angepassten Klassifikationsindexes mit den entsprechenden Anpassungen ist im Anhang (Tabelle 16) einzusehen. Die Anpassungen lassen sich unterteilen in „entfernt“ bzw. „zusammengelegt“ oder in „hinzugefügt“. Die Dokumentenklassen der Kategorien Ankauf und Exit wurden zusammengelegt, da die Inhalte sich nur vom Blickwinkel (Käufer/Verkäufer) unterscheiden und die Dokumentenklassifikation im Anschluss nicht von KI nachvollzogen werden kann. Weitere Dokumentenklassen wurden entfernt, wenn sie ebenfalls Dopplungen darstellten oder zu unspezifisch waren und sich somit nicht für eine automatische Klassifikation geeignet hätten.

Der Dokumentenklassenindex nach Müller (2023) wurde für die beiden Anwendungsfälle Energieeffizienzanalyse und Lebenszyklusanalyse in der Kategorie 15 „Gebäudeversorgung“ erweitert. Hier wurden die Dokumentenklassen „Verbrauchsabrechnung Ver- und Entsorgung“, „Verbrauchsabrechnung Strom“, „Verbrauchsabrechnung Gas“, „Verbrauchsabrechnung Fernwärme“, „Verbrauchsabrechnung Wasser“ und „Verbrauchsabrechnung Daten 'Telekommunikation“ hinzugefügt, da diese eine hohe Relevanz für die beiden Anwendungsfälle aufweisen. Zudem wurden in der Kategorie 6 „Bauplanungsrecht“ die Dokumentenklassen zur Thematik Denkmalschutz an die gif DMS 1.0 angepasst. Hier wird durch die Unterteilung in „Bescheid Denkmalschutz“, „Auskunft Denkmalschutz“, „Gutachten Denkmalschutz“, „Konzept Denkmalschutz“ und „Schriftverkehr zur Erschließung“ feingliedriger unterteilt und somit eine automatische Klassifikation ermöglicht.

5 Automatisierung

Kapitel 5 beschreibt die in *ML-BAU-DOK* entwickelten Automatisierungsprozesse zur Segmentierung und Klassifizierung von Dokumenten. Vorab wird beschrieben, inwieweit eine Priorisierung der Dokumentenklassen – unter Berücksichtigung des *‚data quality assessment framework'* nach *Cai and Zhu (2015)* – möglich ist.

5.1 Priorisierung der Dokumentenklassen

Die Priorisierung der Dokumentenklassen ist ein entscheidender Schritt hin zu einer maschinenbasierten Extraktion von Daten aus Dokumenten. Denn für die Informationsextraktion ist nicht nur entscheidend, welche Information aus einem Dokument zu transferieren ist, sondern auch, in welcher Datenquelle die Information mit der höchsten Qualität enthalten ist. Die Qualität wird von mehreren Faktoren bestimmt. Diese werden im Folgenden berücksichtigt, um Methoden zur Prüfung der Datenqualität und Maschinenlesbarkeit sowie deren Ergebnisse vorzustellen. Neben den Anforderungen an Dokumente im Allgemeinen ist die anwendungsfall-spezifische Datenqualität zu berücksichtigen. Dies geschieht durch die Analyse der für den Anwendungsfall spezifischen Dokumentenklassen und deren inhaltlicher Qualität.

5.1.1 Anforderungen an Dokumente

Wie bereits in Abschnitt 5.1 „Priorisierung der Dokumentenklassen" angedeutet, sind die Anforderungen, die an Daten und Dokumente gestellt werden, sehr vielfältig. Um die Kriterien zu strukturieren, werden in Tabelle 6: „Data Quality assessment framework" nach Cai and Zhu, die Anforderungen an Dokumente übersichtlich dargestellt und beschrieben. Die Anforderungen lassen sich in fünf Hauptkategorien bzw. Dimensionen („Dimensions") unterteilen: Availability (Verfügbarkeit), Usability (Nutzbarkeit), Reliability (Verlässlichkeit), Relevance (Relevanz) und Presentation Quality (Darstellungsqualität).

Jede der fünf Hauptkategorien lässt sich wiederum in feingliedrige Unterkategorien bzw. Elemente („Elements") unterteilen. Diese werden anschließend als Indikatoren („Indicators") beschrieben.

Anhand der Aufschlüsselung durch die Unterkategorien und der Beschreibung der Indikatoren wird in Tabelle 6 beschrieben, welche Anforderungen an Dokumente bestehen.

Für die beiden Anwendungsfälle Energieeffizienzanalyse und Lebenszyklusanalyse sind grundsätzlich alle Kriterien bedeutsam. Als Voraussetzung für die Weiterverarbeitung der Dokumente mit Hilfe digitaler Prozesse sind allerdings besonders „Accessibility" (Zugänglichkeit) und „Readability" (Lesbarkeit) zu nennen. Ohne diese beiden Punkte kann nicht mit den Daten und Informationen aus den Dokumenten gearbeitet werden, da entweder kein Zugriff besteht oder sie nicht gelesen werden können. Aufgrund der hohen Bedeutung wird im folgenden Punkt 5.1.2 ‚Prüfung der Maschinenlesbarkeit' auf diese Thematik genauer eingegangen.

Für beide Anwendungsfälle ist außerdem die Zeitspanne der Datenerhebung und -verarbeitung in der Unterkategorie „Timeliness" (Rechtzeitigkeit) von Relevanz, da sie direkte Auswirkungen auf die Ergebnisse hat. Zudem müssen die Daten den Anforderungen hinsichtlich „Credibility" (Glaubwürdigkeit), „Accuracy" (Korrektheit), „Consistency" (Konsistenz), „Integrity" (Integrität) und „Completeness" (Vollständigkeit) entsprechen, um korrekte und nachvollziehbare Ergebnisse zu erzielen.

Die Relevanz der hervorgehobenen Punkte beschränkt sich nicht nur auf die beiden Anwendungsfälle in *ML-BAU-DOK*, sondern ist allgemeingültig für jegliche Auswertungen und Arbeiten mit digitalisierten Dokumenten.

Tabelle 6: 'Data quality assessment framework' nach Cai and Zhu (2015)[72]

Dimensions	Elements	Indicators
1) Availability	1) Accessibility	- Whether a data access interface is provided - Data can be easily made public or easy to purchase
	2) Timeliness	- Within a given time, whether the data arrive on time - Whether data are regularly updated - Whether the time interval from data collection and processing to release meets requirements
2) Usability	1) Credibility	- Data come from specialized organizations of a country, field, or industry - Experts or specialists regularly audit and check the correctness of the data content - Data exist in the range of known or acceptable values
3) Reliability	1) Accuracy	- Data provided are accurate - Data representation (or value) well reflects the true state of the source information - Information (data) representation will not cause ambiguity
	2) Consistency	- After data have been processed, their concepts, value domains, and formats still match as before processing - During a certain time, data remain consistent and verifiable - Data and the data from other data sources are consistent or verifyable
	3) Integrity	- Data format is clear and meets the criteria - Data are consistent with structural integrity - Data are consistent with content integrity
	4) Completeness	- Whether the deficiency of a component will impact use of the data for data with multi-components - Whether the deficiency of a component will impact data accuracy and integrity
4) Relevance	1) Fitness	- The data collected do not completely match the theme, but they expound one aspect - Most datasets retrieved are within the retrieval theme users need - Information theme provides matches with users' retrieval theme
5) Presentation Quality	1) Readability	- Data (content, format, etc.) are clear and understandable - It is easy to judge that the data provided meet needs - Data description, classification, and coding content satisfy specification and are easy to understand

5.1.2 Prüfung der Maschinenlesbarkeit

Der textliche Inhalt der Dokumente muss inhaltlich gleichbleibend für weiterführende Prozesse extrahiert werden. Die Textqualität ist zudem entscheidend für die angestrebte Klassifizierung und Segmentierung der Dokumente. Demnach ist die Maschinenlesbarkeit digitalisierter Dokumente eines der wichtigsten Parameter für die Anwendung ML-basierter Modelle.[73]

Für die Sicherstellung der Scanqualität wurden bereits in 2.7 Regeln für die Digitalisierung von Dokumenten definiert. Zudem wurden die Auswirkungen von Verschmutzungen oder handschriftlichen Markierungen auf den Dokumenten beschrieben. Abschließend soll die Datenqualität, die durch das Scannen und die Volltextgenerierung entsteht, überprüft werden. Hierzu wurde die Maschinenlesbarkeit nach anerkannter

[72] Vgl. Cai et al. 2015, S. 5.

[73] Vgl. DFG 2016, S. 34.

Stichprobengröße der DFG auf dem in *ML-BAU-DOK* genutzten Datensatz geprüft. Dieser Datensatz wurde auch für die nachfolgende Segmentierung und Klassifizierung verwendet.

Die Qualität der Volltextgenerierung wird gewöhnlicherweise anhand der richtigen Erkennung von Wörtern und Buchstaben eines Datensatzes mit mindestens 500 Elementen geprüft. Die Erkennungsquote wird als Prozentsatz ausgegeben.[74] Die Stichprobengröße umfasst die Erkennung von 506 Buchstaben und die dazugehörige Wortprüfung. Hierzu wurde durch einen Code aus dem angewendeten Datensatz zunächst 506 zufällig generierte Buchstaben ausgewiesen. Diese wurden im Anschluss durch Abgleich mit dem Original auf Richtigkeit geprüft.

Von den 506 zufällig gewählten Buchstaben wurden 478 richtig und 28 falsch erkannt. Daraus ergibt sich ein Prozentsatz der richtigen Buchstabenerkennung von ca. 94 %. Die Analyse der Buchstabenerkennung zeigt, dass falsche Erkennungen auf bekannte Probleme der Texterkennung zurückzuführen sind. So lassen sich falsch erkannte Buchstaben in der Regel auf handschriftliche Notizen, sehr grobe Verschmutzungen des Dokumentes oder einen sehr schlechten Scan zurückführen.

Bei der Prüfung der dazugehörigen Wörter wurden ebenfalls 506 auf Richtigkeit geprüft. Hiervon wurden 452 richtig und 54 falsch erkannt. Die Erkennungsquote der Worterkennung beträgt demnach ca. 89 %.

Es wurde festgestellt, dass die Erkennung von Tabellen- und Textdokumenten keine signifikanten Schwierigkeiten bereitet. Auffällig ist, dass von den 28 falsch erkannten Buchstaben sieben auf großformatige Pläne mit sehr kleiner und teilweiser verdrehter Schrift zurückzuführen sind. Sieben falsch erkannte Wörter stammten von handschriftlichen Notizen, weitere fünf falsche Erkennungen sind auf mangelnde Scanqualität und verschmutzte Seiten zurückzuführen. Somit ließe sich durch eine Steigerung der Scanqualität und die Vermeidung von Handschrift sowie Verschmutzung eine Erfolgsquote bei der Buchstabenerkennung in Höhe von ca. 97 % und bei der Worterkennung in Höhe von ca. 92 % erreichen. Weitere sieben ausgewertete Buchstaben und Wörter sind auf Trennblätter oder inhaltslose Dokumente zurückzuführen, ohne diese ließe sich die Quote der richtigen Erkennung weiter steigen. Im Bereich der Buchstabenerkennung wurden unter idealen Bedingungen also lediglich zwei von 28 Falscherkennungen falsch ausgewertet.

Die Prüfung der Maschinenlesbarkeit ist ein wichtiger Bestandteil von *ML-BAU-DOK*, da weitere Nachnutzbarkeiten hiervon anhängen. Berücksichtigt man, dass die Dokumente in einem Massenscanverfahren digitalisiert wurden und somit geringeren Qualitäts- und Kontrollmechanismen beim Scannen unterlagen, sind die begrenzten Trefferquoten von 97 % bei der Buchstabenerkennung und 92 % bei der Worterkennung positiv zu bewerten.

Die Scanqualität des Datensatzes war außerdem eingeschränkt. Der digitalisierte Massenscan ist auffällig altersbedingt verschmutzt und die Menge an handschriftlichen Vermerken hoch. Vor diesem Hintergrund erscheint das Massenscanverfahren für die Digitalisierung großer Dokumentenmengen sehr gut nutzbar. Unter Einhaltung der in 2.7 beschriebenen Regeln der Digitalisierung von Dokumenten können hinreichend hohe Erkennungsquoten in Buchstaben und Wort erzielt werden.

5.1.3 Methoden der Priorisierung nach Datenqualität

Neben den theoretischen und anwendungsfallspezifischen Anforderungen aus den vorangegangenen Punkten ist die Methodik zur Generierung von priorisierten Dokumenten und Dokumentenklassen entscheidend. Es werden zwei Methoden der Priorisierung vorgestellt, die sich in ihrer Anwendung unterscheiden, jedoch beide praxistauglich sind. Die beiden Methoden sind ein Scoring-Modell und ein Active-Learning-Modell.

5.1.3.1 Scoring-Modell

Das Scoring-Modell gibt jedem Dokument und damit jeder Dokumentenklasse einen Punktwert (Score) bzgl. der Qualität der darin vorhandenen Daten für den betrachteten Anwendungsfall. Auf diese Weise können mehrere Dokumente mit identischem Inhalt miteinander verglichen werden, indem die Qualität in Bezug auf Genauigkeit und Zuverlässigkeit bewertet wird.

[74] Vgl. DFG 2016, S. 35.

Die Bewertung der Informationen erfolgt nach fünf Scores. Die Scores werden für jede Klasse einzeln vergeben. Jedes Dokument wird also vor dem Hintergrund bestimmter vorab definierter Klassen und deren Datenfeldern bewertet. Beispielsweise wird die Dokumentenklasse Energieausweis auf das Datenfeld Primärenergieverbrauch geprüft. Dabei wird zunächst geprüft, ob die Information überhaupt vorhanden ist und dann in welcher Qualitätsstufe (Score). Die Einstufung muss daher für jedes Datenfeld und jede Dokumentenklasse einzeln geprüft werden. Die Scores unterscheiden sich wie folgt:[75]

- 100 %: Daten von Primärdokumenten aus erster Hand mit hoher Aktualität
- 75 %: Daten von Primärdokumenten aus erster Hand mit zyklischer Aktualisierung (z.B. Wartungsintervalle, Abrechnungen)
- 50 %: Daten von Sekundärdokumenten, sekundäre Erfassung (Übertragung aus Primärquelle)
- 25 %: Daten von Sekundärdokumenten, sekundäre Erfassung und zyklische Aktualisierungen
- 0 %: Daten auf geprüftem Dokument nicht verfügbar

Nach der Prüfung einer Dokumentenklasse auf alle Datenfelder eines Anwendungsfalls ergibt sich ein prozentualer Gesamtscore. Dieses Ergebnis kann als Vergleich mit relevanten Dokumentenklassen desselben Anwendungsfalls herangezogen werden. Nach der Prüfung aller relevanten Klassen kann eine Priorisierung der Dokumente nach den Gesamtpunktwerten erfolgen. Durch Festlegung eines prozentualen Grenzwertes ergibt sich eine exklusive Auswahl an Dokumentenklassen, die als Informationsquelle für den spezifischen Anwendungsfall herangezogen werden kann.

Durch die Betrachtung von Datenfeldern anstelle von Dokumentenklassen kann die Dokumentenklasse mit der höchsten Punktzahl eines Datenfeldes und damit der höchsten Informationsqualität für eine vorgegebene Information anhand des Gesamtpunktwertes ausgewählt werden. Hingegen werden bei der Prüfung der Gesamtpunktwerte nach relevanten Dokumentenklassen alle vorab definierten Datenfelder eines Anwendungsfalls geprüft.

Die Ergebnisse werden in einer Matrix erfasst, die eine multidimensionale Auswertung nach Dokumentenklasse oder Datenfeld ermöglicht. Tabelle 7 stellt die Matrix beispielhaft dar. Die Dokumentenklassen sind untereinander in der linken Spalte aufgelistet, anschließend wird die Qualität der Maschinenlesbarkeit angegeben (5.1.2). Daneben wird der errechnete Gesamtpunktwert der jeweiligen Dokumentenklasse auf alle Datenfelder angegeben. Die Spalte Gebäudekategorie ist das erste zu prüfende Datenfeld, in welchem zunächst durch einfache ja/nein-Abfrage die generelle Information geprüft wird. Anschließend sind die Punktwerte als Belastbarkeit des jeweiligen Datenfeldes angegeben.

[75] Die Gewichtungen sind heuristisch entlang von vier gleich großen Intervallen festgelegt. Dies entspricht einer diskreten Notenabstufung von 1 (sehr gut) bis 5 (mangelhaft/ungenügend). Wichtig ist, dass die Abstufungen mit eindeutigen Kriterien voneinander differenzierbar sind.

Tabelle 7: Scoring-Modell[76]

Dokuklassen	Maschinenlesbarkeit	Gesamtpunktwert	Gebäudekategorie	Belastbarkeit Gebäudekategorie [%]	Baujahr	Belastbarkeit Baujahr [%]	NGF	Belastbarkeit NGF [%]	…
01-001	DFG-Score		Ja	100	Ja	75	Nein	0	
01-002	…		Nein	0	Ja	100	Nein	0	
02-001	…		Nein	0	Ja	25	Nein	0	
02-002	…		Ja	75	Ja	25	Nein	0	
03-001	…		Nein	0	Ja	50	Ja	25	
03-002	…		Ja	25	Ja	100	Ja	100	
…	…	…	…	…	…	…	…	…	…

[76] Eigene Darstellung.

5.1.3.2 Active-Learning-Model

Das Active-Learning-Model ist ein Modell zur Anwendung als Open-Source oder in Unternehmen, das eine Bewertung durch Echtzeitabfrage des jeweiligen Dokuments zu einem bestimmten Anwendungsfall erlaubt.

Das Active-Learning-Modell wird in Unternehmen während des Ablageprozesses zur Anwendung gebracht. Somit wird ein digitalisiertes tagesaktuelles Dokument für den jeweiligen Zweck zur Ablage bereitgestellt. Der dahinterstehende Algorithmus gibt bereits eine passende Dokumentenklasse und Auswahl an Datenfelder als Vorschlag. Diese Auswahl kann im Anschluss bestätigt oder korrigiert werden. Insbesondere in permanenten DMS und Datenräumen, die über alle Unternehmensbereiche hinweg genutzt werden, kann das Active-Learning-Modell zur Anwendung kommen. Damit können Datenraumanbieter das Modell der Echtzeitablage integrieren und über Unternehmensgrenzen hinweg die Ablage bestimmter Dokumente und Klassen verfolgen. Voraussetzung ist auch hier ein unternehmensübergreifender Klassenstandard. Auf diese Weise kann ein zielführendes Abbild der Ablagesystematik geschaffen werden.

Durch die akute Anwendung des Active-Learning-Modells in der Praxis kann die Ablage des Dokuments und der darin befindlichen Informationen während der Nutzung in Echtzeit bestimmt werden. Somit wird das Dokument im Sinne des darauf ausgeführten Prozesses bewertet und nicht durch die prozessunabhängige Einschätzung einer einzelnen Person, wie es bei dem Scoring-Modell der Fall wäre.

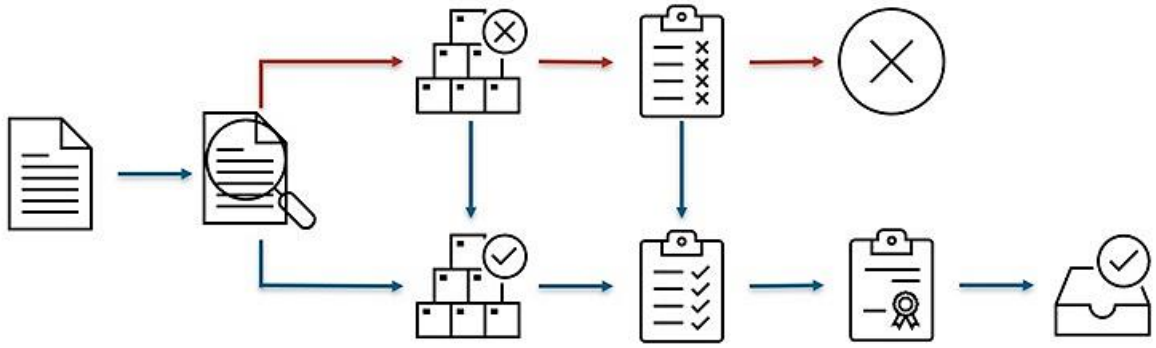


Abbildung 4: Active-Learning-Modell[77]

In Abbildung 4 ist das Active-Learning-Modell dargestellt. Zunächst wird ein Dokument von dem jeweiligen Datenraum ausgelesen und ein Vorschlag zu Ablage unterbreitet. Ist dieser Vorschlag nicht erwartungsgerecht, kann im ersten Schritt die Dokumentenklasse manuell korrigiert werden. Wenn die voreingestellte Auswahl richtig war, kann sie bestätigt werden. Anschließend werden die vordefinierten Datenfelder (Schlüsselinformationen) dargestellt und können manuell bestätigt oder korrigiert werden. Sollte ein Dokument zu keinem Anwendungsfall einen Bezug haben, so kann es für unrelevant erklärt werden. Dokumente, die final einer Dokumentenklasse zugeordnet und in deren Datenfeldern bestätigt wurden, können anschließend durch das Modell verifiziert werden. Die Echtzeitablage wurde vollständig durchgeführt und der Algorithmus kann das Ablageverhalten mit dem bereits bestehenden Ablagemuster der Plattform abgleichen. Dieser mitlernende Algorithmus kann auf Dauer Dokumentenklassen des Immobilienmanagements und deren Inhalte definieren. Die Ordnungssystematik wird gleichermaßen von den Nutzenden bestimmt und kann als Ergebnis der Auswertung großer Datenmengen als standardisiertes Klassifizuerungssystem anerkannt werden.

Durch den Einsatz des Modells über Unternehmensgrenzen hinweg könnte es zudem möglich werden, dass verschiedene Institutionen eine Bewertung zu einzelnen Dokumenten abgeben. Dies würde eine marktgerechte Einstufung von Einzeldokumenten und Dokumentenklassen ermöglichen. Dokumentenklassen könnten so in ihrer Bedeutung objektiv bewertet werden.

[77] Eigene Darstellung.

Das Potenzial eines Active-Learning-Modells ist groß, die Anwendung bedarf aber einiger Voraussetzungen. Es muss in Plattformen integriert werden, die bereits auf dem Markt möglichst breit etabliert sind. Zudem muss die Plattform auch für die tagesaktuelle Ablage genutzt werden, um eine relevante Menge an zugeordneten Dokumenten generieren zu können.

Sowohl das Scoring-Modell als auch das Active-Learning-Modell sind in den Regelbetrieb der Unternehmen integrierbar. Für das Scoring-Modell muss in einem umfangreichen Prozess für jeden Anwendungsfall jede Dokumentenklasse auf Datenfelder geprüft werden. Danach kann das Modell für zukünftige Prozesse angewendet werden. Das Active-Learning-Modell kann durch die Integration in bestehende unternehmensübergreifende Systeme auf eine statistische Validität für die Bewertung von Dokumentenklassen und deren Datenfelder zurückgreifen und ist daher zu bevorzugen.

5.1.4 Folgerung: Anwendungsfallspezifische Dokumentenklassen

Das Scoring-Modell ermöglicht die Auswertung der Dokumentenklassen priorisiert nach ihrem anwendungsspezifischen Nutzen. Die Anwendungsfälle Energieeffizienz- und Lebenszyklusanalyse wurden in Kapitel 3 nach deren Schlüsselinformationen analysiert. Dabei konnten die entscheidenden Datenfelder erkannt werden, anhand derer die Dokumentenklassen innerhalb des Scoring-Modells bewertet werden.

Tabelle 8 zeigt das Ergebnis der mit dem Scoring-Modell priorisierten Dokumentenklassen. Die tatsächliche Bewertung der einzelnen Dokumentenklassen ist in der vollständigen Tabelle 17 im Anhang aufgeführt. Die Vergabe der Scores erfolgte gemäß der Matrix, die in 5.1.3.1 vorgestellt wurde.

Tabelle 8gliedert sich auf der linken Seite in die nummerierten Dokumentenklassen nach Müller (2023). Rechts davon ist das Ergebnis der Maschinenlesbarkeitsprüfung auf dem ausgewählten Datensatz zu sehen, wie in 5.1.2 beschrieben. Daneben sind die Gesamtpunktwerte der einzelnen Dokumentenklassen über die Bewertung aller Datenfelder hinweg aufgelistet. Der Score bildet sich aus dem kumulierten Mittelwert der Punktwertsummen der beiden Anwendungsfälle. Die Sortierung ist absteigend und zeigt in der ersten Zeile die Dokumentklasse mit dem höchsten Informationsgehalt und der höchsten Relevanz. Hier wurde der Energieausweis am höchsten bewertet. Die Bewertung deckt sich mit den Schlüsselinformationen, als deren Quelle der Energieausweis diente. Die Dokumentenklassen mit den höchsten Punktwerten sind des Weiteren (in absteigender Reihenfolge) Gebäudezertifizierungen, Umwelt Due Diligence Gutachten, Umwelt-Zertifikate und Technische Due Diligence Gutachten. Diese Bewertungen setzen sich deutlich von den weiteren Klassen ab. Die Tabelle zeigt die 18 am höchsten bewerteten von insgesamt 332 Dokumentenklassen.

Die Anzahl der informationsrelevanten Dokumentenklassen ist in Relation zur Gesamtzahl an Klassen also sehr gering. Zudem fällt die Auskunftskraft der einzelnen Dokumentenklassen zügig ab.

Die Auswertung ist prinzipiell für alle denkbaren Anwendungsfälle nutzbar. Hierzu müssten lediglich die Matrix um die entsprechenden Datenfelder der anwendungsfallspezifischen Schlüsselinformation erweitert und daraufhin die Dokumentenklassen neu bewertet werden.

Tabelle 8: Priorisierte Dokumentenklassen[78]

Dokumentenklassen nach Mueller et al. 2023		Auswertungen	
		OCR	Energieeffizienz- und Lebenszyklusanalysen
Label	Dokumentenklassen	Maschinenlesbarkeit [%]	Relevanz Anwendungsfälle (Energieeffizienz und Lebenszyklusanalyse) [%]
09-016	Energieausweis	97%	100%
09-019	Gebäudezertifizierung	97%	76%
09-007	Umwelt Due Diligence Gutachten	97%	73%
09-015	Umwelt-Zertifikat	97%	66%
09-002	Technische Due Diligence Gutachten	97%	56%
07-004	Bauantrag	97%	43%
01-001	Exposé	97%	29%
13-036	Nebenkostenabrechnung Mieter	97%	29%
07-006	Baugenehmigungsbescheid 'Ergänzungsbescheid	97%	28%
09-020	Verkehrswertgutachten'Wertgutachten	97%	26%
02-003	Maklerdokumentation	97%	26%
15-009	Verbrauchsabrechnung Strom	97%	20%
12-012	Abnahmeprotokoll 'Prüfprotokoll Heizung	97%	20%
15-010	Verbrauchsabrechnung Gas	97%	20%
15-011	Verbrauchsabrechnung Fernwärme	97%	20%
09-003	Kaufmännische Due Diligence	97%	18%
12-002	Übersicht Gebäudetechnik	97%	18%
12-022	Wartungsprotokoll Heizung	97%	18%

5.2 Automatisierte Dokumentensegmentierung

In *ML-BAU-DOK* wurde zur Digitalisierung die Form des Massenscans gewählt. In der Praxis ist diese eher selten, da eine manuelle Auftrennung der Dokumente in Einzeldokumente zeitintensiv und ineffizient ist. Nachfolgend werden Lösungswege aufgezeigt, wie diese effiziente Form des Scans dennoch zum Einsatz kommen kann. Die entstandenen Algorithmen werden open-source zur Verfügung gestellt.

5.2.1 Theorie

Massenscans erzeugen lange gescannte PDFs, die aus mehreren aufeinander folgenden Dokumenten bestehen. Die Aufgabe ist es, eine solche PDF in einzelne Dokumente zu teilen. Der methodische Ansatz besteht darin, zwischen zwei Seiten zu prüfen, ob sie zum selben Dokument oder zu verschiedenen Dokumenten gehören. Im letzteren Fall wird die PDF zwischen den beiden Seiten getrennt.

Zur Verarbeitung der Daten müssen diese zunächst in sinnvolle Vektorrepräsentationen oder Zahlen umgewandelt werden. Bilder liegen bereits als Matrix vor. So ist ein farbiges Bild bestehend aus 100×200 Pixeln

[78] Eigene Darstellung.

direkt darstellbar als Matrix der Größe 100×200×3, also eine Matrix für jede der drei Grundfarben. Für Text und insbesondere Wörter ist das Finden einer sinnvollen Vektorrepräsentation etwas komplizierter. Hierfür werden im Folgenden mehrere Methoden verwendet: Word2Vec, term Frequency - inverse document frequency (tf-idf) und Convolutional Neural Networks (CNN).

5.2.1.1 Word2Vec

Word2Vec ist eine Methode, um Wörter in Vektoren einheitlicher Größe umzuwandeln. Ziel ist hierbei eine Vektorrepräsentation, bei der ähnliche Wörter ähnliche Repräsentationen zugewiesen werden (siehe Beispiel in Abbildung 5). Das dahinterliegende Modell basiert auf einem neuronalen Netzwerk, das auf Texten trainiert wird. Hierzu werden beispielweise einzelne Wörter aus einem Text entnommen und dem Modell als Lückentextaufgabe präsentiert. Für weitere Informationen wird auf das originale Paper verwiesen.[79]

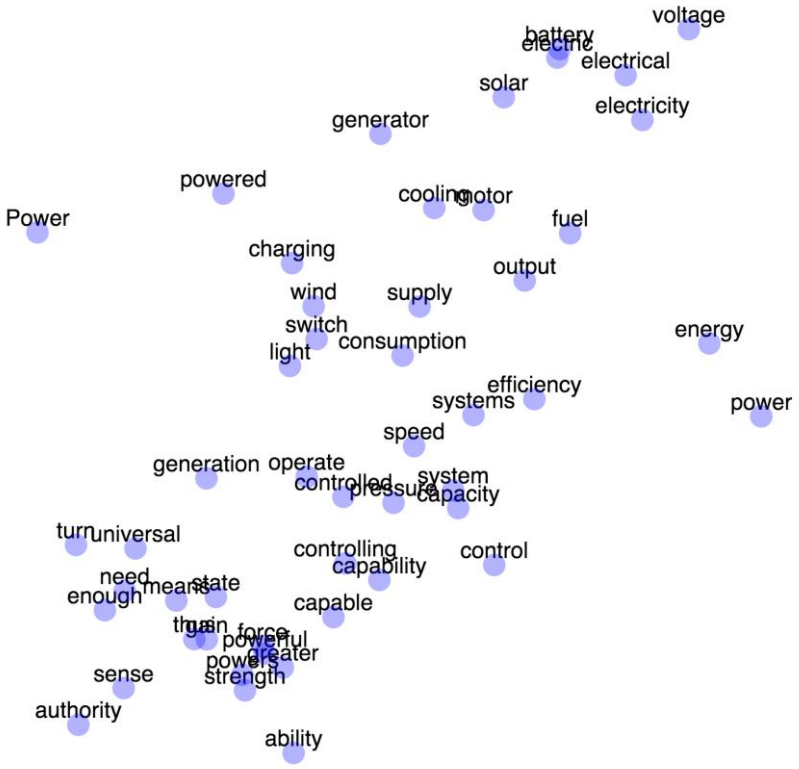


Abbildung 5: Beispiel einer Vektorrepräsentation im Zweidimensionalen[80] [81]

5.2.1.2 Tf-idf

Term Frequency - inverse document frequency (tf-idf)[82] ist eine Methode, um Dokumente oder Texte zu vektorisieren. Es wird davon ausgegangen, dass die Ähnlichkeit zweier Dokumente an der Ähnlichkeit ihrer Wörter gemessen werden kann. Für jedes Wort *w* und Dokument *d* wird die Term Frequency wie folgt berechnet:

[79] Vgl. Mikolov et al. 2013, S. 1–12.
[80] Vgl. Varga 2016.

[81] *Gleichartige Wörter wie etwa "electric", "electrical" und "voltage" haben ähnliche Repräsentationen im Gegensatz zu verschiedenen Wörtern, wie etwa "authority" und "solar". Man beachte, dass diese Vektoren in der Anwendung deutlich höherdimensional sind.*

[82] Vgl. Manning, Raghavan, Schütze 2009, S. 107–110.

$$tf(w,d) = \frac{\text{Anzahl Vorkommnisse des Wortes } w \text{ in Dokument } d}{\text{Anzahl Vorkommnisse des am häufigsten vorkommenden Wortes in } d} \quad (1)$$

Formel 1: Term Frequency

So wird innerhalb eines Dokuments ausgewertet, wie relevant ein Wort *w* in einem Dokument *d* ist. Durch den Nenner in Formel 1 erhält das häufigste Wort die Wertung 1. Weiterhin wird die *Inverse Document Frequency* berechnet:

$$idf(w) = log\frac{\text{Anzahl an Dokumente}}{\text{Anzahl an Dokumente,in denen das Wort w vorkommt}}$$

Formel 2: Inverse document frequency

Dadurch erhalten wir eine Häufigkeitswertung für jedes Wort. Ein in jedem Dokument auftretendes Wort, wie zum Beispiel „ist", erhält eine niedrige Wertung und seltener vorkommende Wörter, wie etwa „Rechnung", erhalten eine sehr hohe Wertung. Der idf ist somit ein Indiz dafür, ob ein Wort relevant ist. Der *tf-idf* setzt sich nun wie folgt zusammen:

$$\text{tf} - \text{idf}(w,d) = \text{tf}(w,d)\text{idf}(w)$$

Formel 3: tf-idf

Sei nun ($w_1, \ldots, w_n$) ein Tupel aller in den Dokumenten vorkommenden Wörter. Jedes Dokument *d* wird folgendermaßen vektorisiert:

$$d \mapsto (\text{tf} - \text{idf}(w_1,d), \ldots, \text{tf} - \text{idf}(w_n,d))$$

Formel 4: Vektorisierung eines Dokuments mittels tf-idf

Insgesamt werden so alle Dokumente in Vektoren einheitlicher Größe umgewandelt.

Zur Verarbeitung der nun vektorisierten Text- und Bilddaten wird ein neuronales Netzwerk eingesetzt. Im Falle der Segmentierung werden dem neuronalen Netzwerk die Bild- und Textdaten zweier aufeinanderfolgender Seiten eingegeben und ein zweidimensionaler Vektor $[p_1, p_2]$ erzeugt. Hierbei beschreibt p_1 die Wahrscheinlichkeit, dass zwischen den Seiten *keine* Trennung vorliegt und umgekehrt p_2 die Wahrscheinlichkeit, *dass* eine Trennung vorliegt.

Für ihre Entscheidungen finden neuronale Netze durch Trainieren eigene Kriterien. Oftmals ist es dennoch hilfreich, handgefertigte Kriterien mitwirken zu lassen. So lassen sich die Ergebnisse z.B. durch die Analyse von Überschriften und Seitenzahlen deutlich steigern. Besonders die Seitenzahlen sind ein gutes Indiz für Dokumentenanfang und -ende.

Ein neuronales Netz kann als Funktion gesehen werden, die von zahlreichen Parametern, auch Gewichte genannt, abhängig ist. Das Ziel ist es, passende Gewichte zu finden, um so das neuronale Netz auf die Aufgabe abzustimmen. Dazu bedarf es Trainingsdaten, also Eingangsdaten, bei denen die gewünschte Ausgabe bereits bekannt ist. So wird bei der Eingabe der Trainingsdaten die Ausgabe des neuronalen Netzwerks mit den gewünschten Ausgaben abgeglichen. Falls diese nicht übereinstimmen, werden die Gewichte entsprechend angepasst. Dieser Prozess ist der Lernprozess eines neuronalen Netzwerks.

Ein effektives neuronales Netzwerk braucht eine effektive Architektur. Bei Daten mit lokaler Korrelation, wie etwa Bilddaten, werden CNN genutzt. Deren Aufbau und Details werden nun erläutert.

5.2.1.3 Convolutional Neural Networks

Ein CNN[83] ist ein künstliches neuronales Netz bestehend aus einem Inputlayer, einem Outputlayer sowie einer Reihe von Hidden Layers, die zumeist eine Kombination aus Convolutional Layers, Pooling Layers, Fully Connected Layers und Normalisierungslayers sind. In diesem Teil werden die wichtigsten Layers einzeln erläutert.

Convolutional Layer: Ein Convolutional Layer wendet diskrete Faltung auf eine 2-dimensionale Matrix $w \times h$ mittels eines Kerns an. Ein Kern ist hierbei eine Matrix $a \times b$, sodass $a \leq w$ und $b \leq h$. Bildlich gesprochen wird hier der Kern auf die Eingangsmatrix gelegt und übereinanderliegende Einträge werden miteinander multipliziert. Die Summe dieser Produkte bildet einen Eintrag der Ausgangsmatrix. Auf diese Weise werden benachbarte Einträge einer Eingangsmatrix in Relation gestellt. Die genaue Rechenweise wird anhand eines Beispiels deutlich.

Beispiel:

Eingangsmatrix des Convolutional Layers: $A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix} \in R^{4\times4}$, Kern: $K = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$

Der Output ist dann:

$$\begin{bmatrix} \mathbf{1} & \mathbf{2} & \mathbf{3} & 4 \\ \mathbf{5} & \mathbf{6} & \mathbf{7} & 8 \\ \mathbf{9} & \mathbf{10} & \mathbf{11} & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix} * \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{2} & \mathbf{3} \\ \mathbf{5} & \mathbf{6} & \mathbf{7} \\ \mathbf{9} & \mathbf{10} & \mathbf{11} \end{bmatrix} * \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} & \begin{bmatrix} 2 & 3 & 4 \\ 6 & 7 & 8 \\ 10 & 11 & 12 \end{bmatrix} * \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \\ \begin{bmatrix} 5 & 6 & 7 \\ 9 & 10 & 11 \\ 13 & 14 & 15 \end{bmatrix} * \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} & \begin{bmatrix} 6 & 7 & 8 \\ 10 & 11 & 12 \\ 14 & 15 & 16 \end{bmatrix} * \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} \begin{bmatrix} \mathbf{1}\times0 & +\mathbf{2}\times1 & +\mathbf{3}\times0 \\ +\mathbf{5}\times0 & +\mathbf{6}\times0 & +\mathbf{7}\times0 \\ +\mathbf{9}\times0 & +\mathbf{10}\times1 & +\mathbf{11}\times0 \end{bmatrix} & \begin{bmatrix} 2\times0 & +3\times1 & +4\times0 \\ +6\times0 & +7\times0 & +8\times0 \\ +10\times0 & +11\times1 & +12\times0 \end{bmatrix} \\ \begin{bmatrix} 5\times0 & +6\times1 & +7\times0 \\ +9\times0 & +10\times0 & +11\times0 \\ +13\times0 & +14\times1 & +15\times0 \end{bmatrix} & \begin{bmatrix} 6\times0 & +7\times1 & +8\times0 \\ +10\times0 & +11\times0 & +12\times0 \\ +14\times0 & +15\times1 & +16\times0 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 2+10 & 3+11 \\ 6+14 & 7+15 \end{bmatrix}$$

$$= \begin{bmatrix} 12 & 14 \\ 20 & 22 \end{bmatrix} =: B$$

Ein Convolutional Layer reduziert die Breite und Höhe der Eingangsmatrix jeweils um $a-1$ bzw. $b-1$. Manchmal ist dies hinderlich, weswegen man auch eine alternative Convolution benutzen kann, bei der zunächst an den Rändern Nullen angefügt werden und dann gefaltet wird.

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix} \star \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 3 & 4 & 0 \\ 0 & 5 & 6 & 7 & 8 & 0 \\ 0 & 9 & 10 & 11 & 12 & 0 \\ 0 & 13 & 14 & 15 & 16 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} =: C \in \mathrm{R}^{4\times4}$$

Anschaulich transformiert ein Kern eine Eingangsmatrix A in die Outputmatrix B oder C wie in den obigen Beispielen.

[83] Vgl. LeCun et al. 1995, S. 1–14.

Beispiele von speziellen Kernen:

1. Identität: Kern $K_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, Eingangsmatrix $A_1 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$

Output:

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix} \star \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$$

Die Anwendung eines Identitätskernes gibt als Output die Eingangsmatrix aus.

2. Verwischen: Kern $K_2 = \frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, Eingangsmatrix $A_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$

Output:

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \star \frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.44 & 0.44 & 0.22 & 0 \\ 0.55 & 0.55 & 0.22 & 0 \\ 0.55 & 0.55 & 0.22 & 0 \\ 0.33 & 0.33 & 0.11 & 0 \end{bmatrix}$$

Dieser Kern verwischt die Matrix, indem es benachbarte Einträge mittelt. (siehe Abbildung 6)

3. Kantenerkennung: Kern $K_3 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, Eingangsmatrix $A_3 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$

Output:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & -3 & 3 & 0 \\ 0 & -3 & 3 & 0 \end{bmatrix}$$

In homogenen Gebieten, also Orten, wo die Einträge der Matrix ungefähr gleich sind, wird dieselbe Zahl achtmal addiert und subtrahiert, sodass das Ergebnis ungefähr 0 beträgt. An Kanten jedoch werden unterschiedliche Werte addiert bzw. subtrahiert, sodass große Beträge entstehen können.

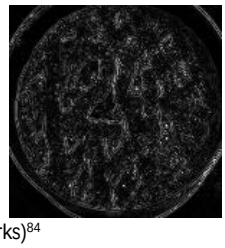
Operation	Eingang	Kern	Output
Identität:		$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Verwischen:		$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Kantenerkennung:		$\begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	

Abbildung 6: Eine Selektion verschiedener Kerne und ihre Auswirkungen (Convolutional Neural Networks)[84]

Ein Convolutional Layer besteht normalerweise nicht nur aus *einem* Kern. Um mehrere Effekte zu analysieren, bedarf es mehrerer Kerne. Gegeben sei nun eine Eingangsmatrix $A \in \mathbb{R}^{w \times h}$ sowie I Kerne $K_1, \dots, K_I \in \mathbb{R}^{a \times b}$. Der Output ist eine Matrix der Größe $(w - a + 1) \times (h - b + 1) \times I$, also *eine* Ausgangsmatrix je Kern.

Durch die Anwendung eines Convolutional Layers wird die Datenmenge ungefähr um ein I-faches vergrößert. Um die Datenmenge und damit die Komplexität des Modells in Grenzen zu halten, werden Pooling Layer verwendet.

Pooling layer: Ein Pooling Layer der Größe (n, m) verkleinert eine Matrix auf ein $\frac{1}{nm}$-tel. Eine sehr verbreitete und effektive Pooling Methode ist das Maxpooling, welches das Maximum einer jeden $n \times m$ Untermatrix extrahiert und alle anderen Werte verwirft. Diese Datenreduktion ermöglicht den Aufbau größerer Modelle, ohne die Rechenkapazitäten zu überlasten.

Beispiel:

Eingangsmatrix= $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$, Pool Größe= (2,2)

Output:

maxpool($\left[\begin{array}{cc|cc} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ \hline 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{array}\right]$, (2,2))= $\begin{bmatrix} max\{1,2,5,6\} & max\{3,4,7,8\} \\ max\{9,10,13,14\} & max\{11,12,15,16\} \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 14 & 16 \end{bmatrix}$

Flattening Layer: Ein Flattening Layer transformiert eine Matrix der Größe $w \times h \times d$ in einen Vektor der Größe whd.

[84] Eigene Darstellung.

Fully Connected Layer: Ein Fully Connected Layer gibt bei einem Eingangsvektor x einen Vektor $y = Ax + b$ aus. Hierbei ist A eine Matrix und b ein Vektor, deren Einträge Gewichte genannt werden. Die Outputdimension ist dabei von der Größe von b. Oftmals folgt darauf eine Aktivierungsfunktion wie etwa ReLU, die positive Einträge belässt, aber negative Einträge auf Null setzt. Diese Aktivierungsfunktion soll biologische Neuronen simulieren, die erst ab einer gewissen Stimulationsschwelle aktiviert werden (Abbildung 7).

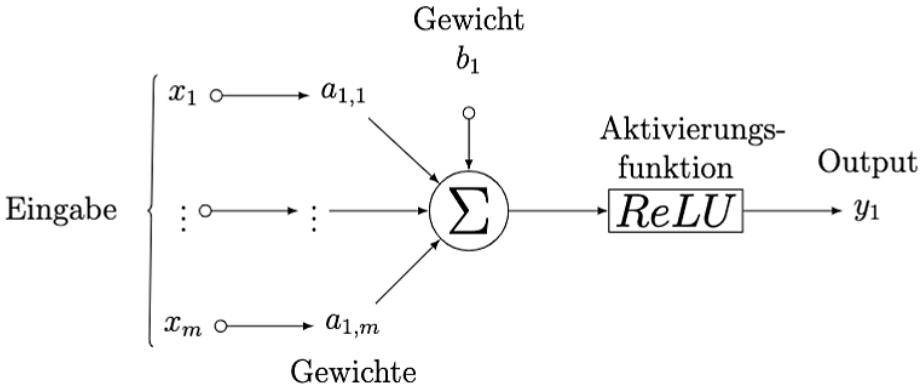


Abbildung 7: Berechnung des ersten Eintrags des Outputvektors eines Fully Connected Layers[85]

Für weitere Informationen wird auf *O'Shea et al 2015* verwiesen.[86]

5.2.1.4 Algorithmus

Bei einem Bildklassifizierungsproblem wird ein Modell durch Kombination der zuvor beschriebenen Layer erstellt. Die Kerne und Gewichte werden zufällig initialisiert und dann später durch Trainieren angepasst. Im Falle der Segmentierung nimmt das neuronale Netzwerk die verarbeiteten Daten zweier aufeinanderfolgenden Seiten auf und gibt einen Wahrscheinlichkeitsvektor $[p_1, p_2]$ aus. Der genaue Prozess, in dem ein solcher Vektor erzeugt wird, wird nun erläutert. Zunächst werden die Bilder zweier Seiten jeweils auf 1100 Pixel in der Höhe und 850 Pixel in der Breite normalisiert, sodass diese als Matrizen der Größe 1100 x 850 x 3 vorliegen. Beide Matrizen werden nun zu einer Matrix der Größe 1100 x 850 x 6 konkateniert und durch das in Tabelle 9beschriebene Modell zu einem Vektor der Größe 1024 verarbeitet.

[85] Eigene Darstellung, in Anlehnung an: vgl. phuong 2013.

[86] Vgl. O'Shea et al. 2015, S. 1–11.

Tabelle 9: Bildverarbeitung[87]

Layer	Dimensionsgröße
Bildeingabe	1100 x 850 x 6
Convolution (3,3), 32 Kerne +ReLU	1098 x 848 x 32
Convolution (3,3), 32 Kerne +ReLU	1096 x 846 x 32
Convolution (3,3), 32 Kerne +ReLU	1094 x 844 x 32
Convolution (3,3), 32 Kerne +ReLU	1092 x 842 x 32
Maxpooling (2,2)	546 x 421 x 32
Convolution (3,3), 64 Kerne +ReLU	544 x 419 x 64
Convolution (3,3), 64 Kerne +ReLU	542 x 417 x 64
Convolution (3,3), 64 Kerne +ReLU	540 x 415 x 64
Convolution (3,4), 64 Kerne +ReLU	538 x 412 x 64
Maxpooling (2,2)	269 x 206 x 64
Convolution (3,3), 64 Kerne +ReLU	267 x 204 x 64
Convolution (3,3), 64 Kerne +ReLU	265 x 202 x 64
Convolution (3,3), 64 Kerne +ReLU	263 x 200 x 64
Convolution (4,3), 64 Kerne +ReLU	260 x 198 x 64
Maxpooling (2,2)	130 x 99 x 64
Convolution (3,3), 128 Kerne +ReLU	128 x 97 x 128
Convolution (3,3), 128 Kerne +ReLU	126 x 95 x 128
Convolution (3,3), 128 Kerne +ReLU	124 x 93 x 128
Convolution (3,4), 128 Kerne +ReLU	122 x 90 x 128
Maxpooling (2,2)	61 x 45 x 128
Convolution (3,3), 128 Kerne +ReLU	59 x 43 x 128
Convolution (3,3), 128 Kerne +ReLU	57 x 41 x 128
Convolution (3,3), 128 Kerne +ReLU	55 x 39 x 128
Convolution (4,4), 128 Kerne +ReLU	52 x 36 x 128
Maxpooling (2,2)	26 x 18 x 128
Convolution (3,3), 128 Kerne +ReLU	24 x 16 x 128
Convolution (3,3), 128 Kerne +ReLU	22 x 14 x 128
Convolution (3,3), 128 Kerne +ReLU	20 x 12 x 128
Flatten	30720
Fully connected, n=1024 + ReLU	1024

Zugleich werden die Textdaten beider Seiten mittels tf-Idf zu je einem Vektor der Größe n transformiert und mit folgendem Modell zu einem Vektor der Größe 1024 verarbeitet (Tabelle 10).

[87] Eigene Darstellung.

Tabelle 10: Textverarbeitung[88]

Layer	Dimensionsgröße
Texteingabe, tf-Idf	2n
Fully connected, n=1024 + ReLU	1024
Fully connected, n=1024 + ReLU	1024
Fully connected, n=1024 + ReLU	1024
Fully connected, n=1024 + ReLU	1024
Fully connected, n=1024 + ReLU	1024
Fully connected, n=1024 + ReLU	1024

Nach der Verarbeitung der Bild und Textdaten werden die Outputs zusammengetragen und schließlich durch das folgende Modell in einen 2-dimensionalen Wahrscheinlichkeitsvektor umgewandelt (Tabelle 11).

Tabelle 11: Gesamtes Modell[89]

Layer	Dimensionsgröße
Eingabe	2 Seiten
Bild- und Textverarbeitung	1024+1024
Fully connected, n=1024 + ReLU	1024
Fully connected, n=512 + ReLU	512
Fully connected, n=100 + ReLU	100
Fully connected, n=2 + Softmax	2

Die Softmaxfunktion[90] wird zum Schluss eingesetzt, um einen zweidimensionalen Vektor (x, y) in eine Wahrscheinlichkeitsverteilung umzuwandeln.

$$softmax(x, y) = \left(\frac{e^x}{e^x + e^y}, \frac{e^y}{e^x + e^y}\right)$$

Formel 5: Softmaxfunktion

Offensichtlich ist die Ausgabe abhängig von den verschiedenen Kernen der Convolutional Layer und Gewichte der Fully Connected Layer. Zunächst werden diese zufällig initialisiert und dann durch einen Trainingsprozess angepasst. Dazu werden Trainingsdaten benötigt, also eine möglichst große Menge an aufeinanderfolgenden Seitenpaaren zusammen mit der Information, ob die jeweiligen Paare zusammengehören oder getrennt werden sollen.

Das in *ML-BAU-DOK* entwickelte Modell gibt für je zwei aufeinanderfolgenden Seiten eine Wahrscheinlichkeitsverteilung $[p_1, p_2]$ aus. p_1 beschreibt die Wahrscheinlichkeit, dass zwischen den Seiten

[88] Eigene Darstellung.

[89] Eigene Darstellung.

[90] Vgl. Heaton 2018, S. 1–3.

keine Trennung vorliegt, umgekehrt beschreibt p_2 die Wahrscheinlichkeit, *dass* eine Trennung vorliegt. Eine Trennung erfolgt genau in dem Fall, dass $p_1 < p_2$.

Um die Ergebnisse weiter zu verbessern, werden vor dem Modell einige eindeutige Dokumente segmentiert. Dazu gehören unter anderem auch Dokumente mit leicht auswertbaren Seitenzahlen, die im Text beispielsweise als „Seite 1 von 4“ vorliegen und somit mittels regulärer Ausdrücke (RegExp) leicht zu extrahieren und zu verwerten sind. Auf ähnliche Weise werden die Seiten auf gleiche Überschriften und sehr ähnlichen Text geprüft.

5.2.2 Auswertung

Als Datensatz wurden 158 PDF-Dateien genutzt, in denen insgesamt 5.683 Dokumente enthalten sind. 19 dieser PDF-Dateien wurden für den Test zufällig entnommen. Die restlichen Dokumente dienen als Trainingsdaten.

Die Auswertung der Testdaten bringt folgende Ergebnisse (Tabelle 12):

Tabelle 12: Ergebnisse der Auswertung des Modells an Testdaten[91] [92]

	Trennung erkannt	Keine Trennung erkannt	Genauigkeit
Trennung	411	68	85,8 %
Keine Trennung	415	3.319	88,89 %
Genauigkeit	49,76 %	98,2 %	88,54 %

Durch das Modell kann bei ungefähr 89 % aller Seitenpaare richtig erkannt werden, ob zwischen ihnen die PDF getrennt werden muss oder nicht. An der Tabelle ist zu erkennen, dass insgesamt 411+68=479 mal getrennt werden sollte, aber 411+415=826 mal von dem Modell getrennt wird. Dadurch entstehen 826-479=347 zusätzliche Dokumente, sodass eine große Zahl der getrennten Dokumente unvollständig sind oder Teile zweier Dokumente enthalten. Die Auswertung der segmentierten Dokumente ist in Tabelle 13 aufgelistet. Wie man sieht, ist der Anteil der vollständig richtig erkannten Dokumente für den praktischen Einsatz noch zu gering, bildet aber eine sehr gute Grundlage für die weitere Optimierung der Open-Source-Algorithmen.

Tabelle 13: Auswertung des Modells anhand der Anzahl richtig segmentierter Dokumente[93]

		Anteil			Anteil
Anzahl Dokumente	498	100 %	Anzahl erkannter Dokumente	845	100 %
Davon richtig getrennt	195	39,16 %	Davon echte Einzeldokumente	195	23,08 %
Davon nicht richtig getrennt	303	60,84 %	Davon fehlerhaft	650	76,92 %

5.2.3 Folgerung: Ausblick

Vor der Textverarbeitung müssen Texte mittels OCR erkannt und extrahiert werden. Leider wurden in dem Datensatz für *ML-BAU-DOK* einige Buchstaben oder Wörter falsch erkannt, was die Qualität der Textvektorisierung und -verarbeitung stark beeinträchtigt. Besonders Texte aus handgeschriebenen und fleckigen Dokumenten sind schwer zu extrahieren.

[91] Eigene Darstellung.

[92] Insgesamt 4.213 Seitenpaare. Es wurde z.B. in 68 Fällen fälschlicherweise keine Trennung erkannt.

[93] Eigene Darstellung.

Es ist zu beachten, dass die vorhandenen Dokumente in Bild und Text unterschiedlich aufgebaut sind. Aufgrund dieser Heterogenität ist es schwierig, Muster und Kriterien zu erkennen, die für eine Segmentierung hilfreich sein könnten. Um eine solche komplexe Aufgabe zu lösen, muss ein entsprechend großes und komplexes Modell gebaut werden. Jedoch erfordert das Trainieren mit zunehmender Komplexität des Problems und des Modells mehr und mehr Trainingsdaten.[94] In unserem Fall wurden die gegebenen Trainingsdaten manuell segmentiert. Eine signifikante Erhöhung der Datenmenge ist nur sehr schwer zu erreichen. Bei ausreichend großer Datenmenge könnten modernere Strukturen wie etwa Transformerarchitekturen[95] in die Algorithmen eingebaut werden, die beispielsweise Eigenschaften aus dem Kontext eines Textes extrahieren können.

5.3 Automatisierte Dokumentenklassifizierung

Nach der Dokumentensegmentierung sind Einzeldokumente in PDF-Format verfügbar. Um mit diesen effektiv arbeiten zu können, muss eine Ordnung geschaffen werden, etwa durch Klassifizierung.

5.3.1 Klassifizierung

Analog zur Segmentierung werden auch hier die Daten zuerst vektorisiert und dann verarbeitet. Zur Klassifizierung werden nun einige Methoden vorgestellt: k-nearest neighbors, Nearest Centroid Classifier und k-means Clustering. Gegeben seien die Klassen $1, \dots, n$. Aufgabe ist es, jedes Dokument einer der n Klassen zuzuordnen. Sämtliche dieser Methoden benötigen Trainingsdaten, also Dokumente, die bereits klassifiziert worden sind. Die Vektorisierungen solcher Dokumente bezeichnen wir mit $x_1, \dots, x_m$ und die Klassen jeweils mit $y_1, \dots, y_m$.

k-nearest neighbors

In k-nearest neighbors[96] erfolgt die Klassifikation eines Vektors x durch Mehrheitsentscheidung aufgrund der k nächsten Trainingsdaten von x. x wird der Klasse mit der größten Anzahl der Objekte dieser k Nachbarn zugewiesen (Abbildung 8).

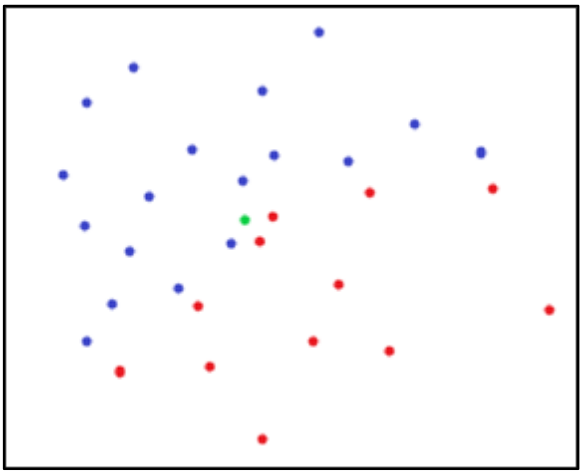


Abbildung 8: Beispiel des k-nearest neighbor Algorithmus[97][98]

[94] Vgl. Althnian, AlSaeed, Al-Baity et al. 2021, S. 1–18.

[95] Vgl. Vaswani et al. 12.06.2017, S. 1–15.

[96] Vgl. Manning, Raghavan, Schütze 2009, S. 273–275.

[97] Eigene Darstellung.

[98] Im Falle k=3 wird der Testdatenpunkt (grüner Punkt) der roten Klasse zugeordnet, da zwei der drei nächsten Nachbarn Teil der roten Klasse sind. Für k=5 wird es der blauen Klasse zugeordnet.

Nearest Centroid Classifier

Der Nearest Centroid Classifier[99] berechnet je Klasse ein Zentrum der Trainingsdaten. Für die i-te Klasse wird das Zentrum c_i als Durchschnitt aller Vektoren x_j, die der Klasse i zugeordnet sind, gesetzt. Für ein neues Dokument mit Vektorisierung x wird die Klasse als diejenige vorausgesagt, deren Zentrum am nächsten ist.

5.3.2 Clustering

Stehen keine oder nur unzureichende Trainingsdaten zur Verfügung, werden Clusteralgorithmen eingesetzt. Hierbei werden ähnliche Datenpunkte, hier Dokumente, zu einer Klasse zusammengefasst ohne im Vorhinein genau zu definieren, welche Klassen existieren. Eine bekannte und verbreitete Methode ist *k-means*, wofür die Dokumente zuerst in Vektoren umgewandelt werden müssen.

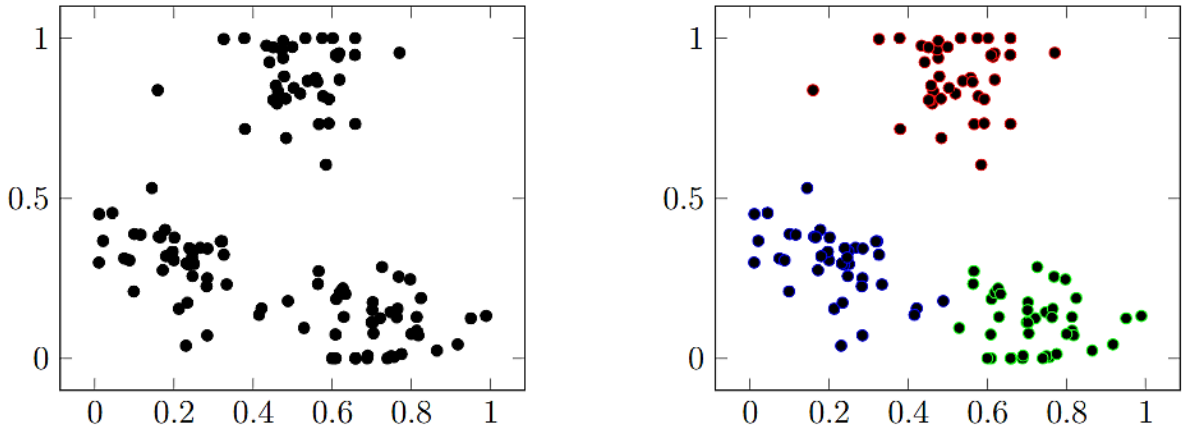


Abbildung 9: Ergebnis der Clustererkennung[100] [101]

k-means

Mittels tf-Idf werden die Dokumente in Vektoren umgewandelt, die wir nun mit $x_1, \ldots, x_n$ bezeichnen. In Abbildung 9 sind beispielhaft diese Vektoren in eine zweidimensionale Ebene eingezeichnet. In *k-means*[102] werden nahe beieinanderliegende Vektoren in eine Klasse zusammengefasst, wie farblich in Abbildung 9 gekennzeichnet. Man beachte hierbei, dass unsere Vektoren nicht zwei-, sondern viel höherdimensional sind. Tabelle 14 fasst den k-means-Algorithmus zusammen.

[99] Vgl. Manning, Raghavan, Schütze 2009, S. 269–273.

[100] Eigene Darstellung.

[101] Links: Vektoren $x_1 \ldots, x_n$, Rechts: Vektoren nach dem Clusteralgorithmus. Die Anzahl an Clustern wurde auf k=3 gesetzt.

[102] Vgl. Manning, Raghavan, Schütze 2009, S. 331–336.

Tabelle 14: Der k-means Algorithmus[103]

Algorithmus	k-means
1: **Input** Vektoren $x_1, \ldots, x_n$	
2: **Initialisierung** Zentren $y_1, \ldots, y_k$	
3: **while** nicht konvergiert **do**	
4: Ordne Vektoren $x_1, \ldots, x_n$ dem nächsten Clusterzentrum y_i zu (entsprechender Vektor liegt somit im Cluster Nummer i)	
5: Aktualisiere Clusterzentren: y_i = Mitte der Vektoren, die dem Cluster i zugehörig sind	
6: **end while**	
7: **Output** Cluster $1, \ldots, k$	

5.3.3 Auswertung

Für die Klassifizierung sind 32 Energieausweise, 38 Stromabrechnungen und 40 Gasabrechnungen als Daten gegeben. Je Klasse werden vier Dokumente als Testdaten entnommen, der Rest wird als Trainingsdatensatz verwertet. Diese wurden zunächst mittels tf-Idf vektorisiert und dann klassifiziert. Auf die Verwertung von Bilddaten für die Klassifizierung wurde verzichtet, sodass die jeweilige Klasse nur durch den Text entschieden wurde. Die Ergebnisse der Testdaten enthält Tabelle 15.

Tabelle 15: Auswertung Klassifizierung[104]

	Genauigkeit
k-nearest neighbors (k=3)	75 %
k-nearest neighbors (k=5)	100 %
Nearest centroid classifier	50 %
Neuronales Netzwerk	66 %

Neuronale Netze sind besonders bei komplexen Aufgaben vorteilhaft, bei denen Eigenschaften berechnet werden müssen, die nur schwer manuell zu implementieren sind. Die Aufgabe der Klassifizierung von Energieausweisen, Stromabrechnungen oder Gasabrechnungen ist jedoch eher eindeutiger und leichter, sodass herkömmliche ML-Methoden wie etwa k-nearest neighbors, sehr gute Ergebnisse erzielen können. Die Fähigkeit, komplexe Probleme zu lösen, verdankt das neuronale Netz seiner flexiblen Struktur. Jedoch bedarf es aufgrund dieser Flexibilität deutlich mehr Trainingsdaten als bei herkömmlichen Methoden, bei denen bereits wesentliche Strukturen vordefiniert sind.

Der segmentierte Großdatensatz lag in nicht-klassifizierter Form dar. Diesen Datensatz manuell zu klassifizieren und daraufhin ein Modell zu trainieren, ist nicht sinnvoll, da aufgrund der Heterogenität zu viele Klassen vertreten sind und jede einzelne Klasse zu wenige Trainingsdokumente hätte. In solchen Fällen empfiehlt sich das Clustern. Hierbei wurden die Dokumente in 13 Cluster aufgeteilt, unter denen die Dokumente gemäß dem Algorithmus ähnlich sind. Nachteil ist hierbei, dass die Cluster nicht vordefiniert sind, sodass Klassen erstellt werden, die womöglich für Anwendung nicht zweckmäßig sind. Zudem muss die Analyse der Cluster manuell durch Analyse der Dokumente pro Klasse vollbracht werden.

5.3.4 Folgerung: Ausblick

Der k-nearest neighbors Algorithmus hat die perfekte Genauigkeit von 100 % auf den zwölf Testdokumenten erreicht. Zu erwarten ist, dass bei mehr Testdaten ein solches Ergebnis nicht zu erzielen ist. Bei einer solchen „leichten" Klassifizierungsaufgabe kann auch mit mehr Daten k-nearest neighbors die genaueste Methode

[103] Eigene Darstellung.

[104] Eigene Darstellung.

sein. Kommen jedoch Klassen hinzu, die schwieriger zu unterscheiden sind, müssen neuronale Netze eingesetzt werden, deren Größe und damit der Bedarf an Trainingsdaten von der Komplexität der Aufgabe abhängt.

Zudem muss der Vektorisierungsalgorithmus an die Komplexität der Aufgabe angepasst werden. In tf-Idf wird ein Dokument als Menge an Wörtern gesehen, Eigenschaften wie etwa Kontext und Reihenfolge der Wörter und Sätze werden ignoriert. Falls Klassen hinzugefügt werden, die nur durch diese Eigenschaften zu unterscheiden sind, muss ein entsprechender Vektorisierungsalgorithmus und dann ein passendes neuronales Netzwerk verwendet werden. Für letzteres eignen sich Long Short-Term Memory Networks (LSTM)[105] und Transformermodelle[106].

[105] Vgl. Goldberg 2017, S. 179–181.

[106] Vgl. Vaswani et al. 12.06.2017, S. 1–15.

6 Dokumentation und Dissemination

Die dokumentierten Ergebnisse werden für die möglichst breite Dissemination in diesem Kapitel als Fazit, Limitation und Ausblick zusammengefasst. Diese folgen jeweils den Zielsetzungen und Forschungsfragen von *ML-BAU-DOK*.

6.1 Fazit

Abschließend ist zu klären, ob und inwieweit die Priorisierung und anschließende automatisierte Klassifizierung von Dokumenten des Immobilienmanagements einen Mehrwert für das Gebäudemanagement und das Dokumentenmanagement bringt. Grundlegend werden die Regeln für den zielführenden Umgang mit Dokumenten und die daraus resultierenden Arbeitsschritte zusammengefasst. Dies geschieht anhand der Dokumentation für die Anwendungsfälle Energieeffizienz- und Lebenszyklusanalysen.

Formulierung von Regeln für die effektive Digitalisierung von Gebäude- und anlagenbezogenen Dokumenten als Grundvoraussetzung für Maschinelles Lernen: Welche Regeln müssen bzw. sollten für die effektive Digitalisierung von gebäude- und anlagenbezogenen Dokumenten eingehalten werden?

Die Regeln der Digitalisierung orientieren sich an den Teilprozessen der Beauftragung. Allerdings sind für die Anwendung einer späteren automatisierten Klassifizierung lediglich die ersten vier Teilprozesse der Digitalisierungsregeln – Übertragung der Aktenordner, Festlegung der Indexierung, Scan der Dokumente, Volltextgenerierung – zu betrachten. Die wichtigsten Regeln im Sinne der späteren Klassifizierung betreffen das ordnungsgemäße Pre-Processing, die Scanqualität und den geordneten Massenscan als Alternative zu dem Einzeldokumentenscan. Darüber hinaus ist die Festlegung eines Index oder dessen automatische Erstellung durch den Klassifikationsalgorithmus von Bedeutung. Die OCR muss eine hohe Qualität aufweisen, unabhängig davon, ob sie bereits bei Scan integriert oder nachträglich bei der Klassifizierung auferlegt wird. Die Regeln sind im Hauptteil in Abbildung 3 zusammengefasst.

Priorisierung der Dokumentenklassen für die definierten Anwendungsbereiche, die Methodik und Kriterien sollen auf andere Anwendungsbereiche übertragbar sein: Wie können die Dokumentenklassen für die spezifizierten Anwendungsbereiche priorisiert werden, um Effizienz und Effektivität einer maschinenbasierten Informationsextraktion sicherzustellen?

Spezifikation der wesentlichen Dokumentenklassen für die Anwendungsbereiche Energieeffizienz und Lebenszyklusanalysen: Welche wesentlichen Dokumentenklassen sind relevant für die spezifizierten Anwendungsbereiche Energieeffizienz und Lebenszyklusanalysen?

Die Energieeffizienz von Gebäuden wird in der Praxis durch den in Deutschland einheitlichen Energieausweis nachgewiesen. Hierdurch können die Schlüsselinformationen für die Feststellung oder Messung von Energieeffizienz bestimmt werden. Die Schlüsselinformationen konnten auf 17 reduziert werden. Die Lebenszyklusanalyse ist ein breit gefächerter und viel diskutierter Begriff in der Immobilienwirtschaft. In *ML-BAU-DOK* umfasst sie die periodische Betrachtung der Ergebnisse aus der Energieeffizienzanalyse. Ansätze für die Definition von Dokumentenklassenstandards sind national und international gegeben (gif, FIDJI, …), allerdings konnte sich bisher kein Standard nachhaltig etablieren. Dies führt dazu, dass eine Übertragbarkeit und einheitliche unternehmensübergreifende Gliederungsform bisher fehlen. Die Einteilung nach Müller (2023) wurde aufgrund der technischen Orientierung gewählt und geringfügig spezifiziert.

Für die Spezifikation wurden zwei Modelle beschrieben, deren Methodik auf alle denkbaren Anwendungsbereiche übertragbar sind. Das Active-Learning Modell und das Scoring-Modell können beide in Unternehmen integriert werden, wobei ersteres einen eigenen Prozess innerhalb der Digitalisierung von Dokumenten darstellt. Das Scoring-Modell ist flexibel erweiterbar und ermöglicht die subjektive Priorisierung einzelner Klassen pro Anwendungsfall. Durch die Anwendung des Scoring-Modells wurde die Anzahl der Klassen auf 18 festgelegt (5.1.4).

Durch eine Priorisierung kann die Anzahl der Klassen hinsichtlich ihrer Aussagekraft weiter eingeschränkt werden. Dadurch können die Algorithmen für ausgewählte Klassen angepasst und die Qualität der Ergebnisse verbessert werden. So kann die Effektivität weiter gesteigert werden, allerdings aufgrund der kleinteiligen Codes ggf. auf Kosten der Effizienz.

Darlegung der Möglichkeiten für (automatisches) Klassifizieren von gebäude- und anlagenbezogenen Dokumenten: Welche Möglichkeiten bestehen für das (automatische) Klassifizieren von gebäude- und anlagenbezogenen Dokumenten?

Durch die Scanweise (Massenscan) der vorliegenden Dokumente in große zusammenhängende PDFs musste zunächst eine Lösung für die Trennung in Einzeldokumente gefunden werden. Dies ist für die Praxis besonders relevant, da das Scannen von Einzeldokumenten prohibitiv aufwändig ist. Anstelle einer manuellen Auftrennung der zahlreichen Dokumente wurde ein Segmentierungsalgorithmus auf Basis neuronaler Netze programmiert. Hierzu werden Merkmale aus Bild und Text für Kriterien der Segmentierung berücksichtigt. Die Genauigkeit dieser Methode, unter Betrachtung der Erkennungsquote einer Trennung zwischen zwei Seiten, beträgt auf den von uns angewendeten Dokumenten ca. 89 %. Allerdings werden bisher lediglich ca. 39 % der Dokumente richtig segmentiert, was auf die begrenzte Anzahl von Trainingsdokumenten und Limitationen der OCR zurückzuführen ist. Der Algorithmus stellt gleichwohl bereits eine sehr gute Basis dar, auf die in Folgeprojekten und durch Unternehmen aufgebaut werden kann.

Zur Klassifizierung wurden mehrere Methoden in Betracht gezogen. **Herkömmliche Klassifizierungsmethoden** wie k-nearest neighbors haben für unsere Anwendungsfälle (Energieausweis, Gas- und Stromabrechnungen) eine Genauigkeit von 100 % erreicht. **Neuronale Netze** erfordern große Mengen Trainingsdaten und sind relevant für komplexe Klassifizierungen. **Clustering** kann alternativ angewandt werden, wenn keine oder unzureichende Trainingsdaten verfügbar sind. Die Auswertung der automatisiert generierten Cluster muss letztendlich wohl auch in Zukunft manuell erfolgen oder zumindest bestätigt werden. Dokumentenmengen können dazu durch Segmentierung automatisiert in Einzeldokumente getrennt werden. Zudem gibt es klassenbezogene Algorithmen für die anschließende Zuordnung der Dokumente in Dokumentenklassen. Allerdings können bei den in *ML-BAU-DOK* vorgestellten Methoden nur die herkömmlichen Klassifizierungsmethoden (wie k-nearest neighbors) oder neuronale Netze auf vorab definierte Klassen angewandt werden.

Abschließend kann festgehalten werden, dass trotz der unterschiedlichen Dokumentenklassen unter Einhaltung definierter Regeln für die Digitalisierung von Dokumenten, eine Klassifizierung sowohl nach einem vorgegebenen Index, als auch in einem automatisiert erstellten Index möglich ist. Insbesondere die Segmentierung der Dokumente und die damit einhergehende Möglichkeit für Massenscans kann die Effektivität und Effizienz im Dokumentenumgang wesentlich steigern. Allerdings müssen die Ergebnisse in nachfolgendem Abschnitt etwas limitiert werden.

6.2 Limitation

Formulierung von Regeln für die effektive Digitalisierung von Gebäude- und anlagenbezogenen Dokumenten als Grundvoraussetzung für Maschinelles Lernen: Welche Regeln müssen bzw. sollten für die effektive Digitalisierung von gebäude- und anlagenbezogenen Dokumenten eingehalten werden?

Die Erkennungsquote und Scanqualität werden in der Praxis durch Verschmutzung, Handschrift und sonstige Rückstände auf Dokumenten beeinträchtigt, besonders bei sehr alten Dokumenten. Hierunter leidet insbesondere die Qualität der angewendeten OCR, was sich auf alle folgenden Textverarbeitungsprozesse auswirkt. Nichtsdestotrotz konnte eine Erkennungsquote von Buchstaben in Höhe von 97 % erreicht werden, bei Wörtern von 92 % (5.1.2). Für optimale Resultate in dieser Größenordnung ist aufwändiges Pre-Processing notwendig.

Priorisierung der Dokumentenklassen für die definierten Anwendungsbereiche, die Methodik und Kriterien sollen auf andere Anwendungsbereiche übertragbar sein: Wie können die Dokumentenklassen für die spezifizierten Anwendungsbereiche priorisiert werden, um Effizienz und Effektivität einer maschinenbasierten Informationsextraktion sicherzustellen?

Die Priorisierung der Dokumentenklassen unter Anwendung des Active-Learning- oder Scoring-Modells ist jeweils subjektiven Einflüssen ausgesetzt. Zudem erfordert die Anwendung beider Systeme die manuelle und daher zeitaufwändige dokumentenspezifische Wertung. Die Übertragbarkeit der beiden Modelle ist gegeben, allerdings müssen die Modelle je nach Anwendungsfall manuell erweitert und spezifisch angepasst werden. Somit sinkt aktuell die Effektivität, Effizienz und Objektivität des angepassten Scoring-Modells. Das Active-Learning-Modell könnte bei Integration in DMS-Anwendungen flächendeckend zum Einsatz kommen und mit einer entsprechend großen und heterogenen Dokumentbasis unternehmensübergreifende Klassenstandards bilden.

Spezifikation der wesentlichen Dokumentenklassen für die Anwendungsbereiche Energieeffizienz und Lebenszyklusanalysen: Welche wesentlichen Dokumentenklassen sind relevant für die spezifizierten Anwendungsbereiche Energieeffizienz und Lebenszyklusanalysen?

Unabhängig davon, welcher Dokumentenklassenindex genutzt wird, ist es möglich, dass die Übertragbarkeit für einzelne Klassen nicht gegeben ist. Durch stetigen Wandel der Immobilienwirtschaft und damit verbundene Weiterentwicklung der Dokumente können neue Dokumentenklassen entstehen, die in dem in *ML-BAU-DOK* genutzten Index nicht enthalten sind. Dies verlangt eine kontinuierliche Anpassung der zugrunde gelegten nationalen oder internationalen Dokumentenklassen. Trotz der spezifizierten Dokumentenklassen können gleichartige Inhalte in mehreren Klassen für die Extraktion enthalten sein. Hieraus ergibt sich weiterhin die Frage, welche Klasse die Primärquelle ist, besonders wenn die Inhalte voneinander abweichen (5.1.3.1).

Darlegung der Möglichkeiten für (automatisches) Klassifizieren von gebäude- und anlagenbezogenen Dokumenten: Welche Möglichkeiten bestehen für das (automatische) Klassifizieren von gebäude- und anlagenbezogenen Dokumenten?

Die Ergebnisse der Dokumentensegmentierung sind unter Betrachtung der tatsächlich gefundenen Dokumente mit ca. 39 % noch zu gering für einen flächendeckenden Einsatz. Dies ist einerseits mit der geringen Menge an Trainingsdaten und andererseits die bei den gegebenen Dokumenten qualitativ schlechte OCR zu erklären. Die Genauigkeit sinkt bereits bei wenigen falsch erkannten Trennungen enorm. Die Nutzung ist erst nach weiterem Training mit größeren Trainingsdatenmengen sinnvoll, dann jedoch aussichtsreich.

Die in *ML-BAU-DOK* entwickelten Modelle ermöglichen zwar die zielgenauere automatisierte Klassifizierung unter Anwendung herkömmlicher Klassifizierungsmethoden oder neuronaler Netze, allerdings nur unter großem zusätzlichem Aufwand und mit dementsprechend reduzierter Zuordnungsgeschwindigkeit. Aufgrund der begrenzten Anzahl von passenden Trainingsdaten musste für die Klassifizierung auf die

Methodik des Clustering zurückgegriffen werden, was die Definition eigener Dokumentenklassen zur Folge hat. Bei der Verwendung von neuronalen Netzen besteht der Nachteil, dass im täglichen Einsatz eine Zuordnung neuer Dokumente zu neuen, bisher nicht verwendeten Dokumentenklassen wiederum erst nach einer ausreichenden Menge von Trainingsdaten und dem damit verbundenen Training möglich ist. Herkömmliche Klassifizierungsmethoden wie k-nearest neighbors sind nur auf leicht klassifizierbaren Problemen erfolgreich. Bei steigender kontextrelevanter Komplexität muss auf neuronale Netze zurückgegriffen werden. Die geringe Anzahl passender Dokumente verhinderte das Experimentieren mit größeren Modellen. Die Grundlagen dafür liegen durch *ML-BAU-DOK* nun vor. Eine größere Datenbasis ermöglicht nicht nur eine höhere Wahrscheinlichkeit für die richtige Zuweisung der Dokumente, sondern auch ein größeres Spektrum an geeigneten Methodiken.

6.3 Ausblick

Formulierung von Regeln für die effektive Digitalisierung von Gebäude- und anlagenbezogenen Dokumenten als Grundvoraussetzung für Maschinelles Lernen: Welche Regeln müssen bzw. sollten für die effektive Digitalisierung von gebäude- und anlagenbezogenen Dokumenten eingehalten werden?

Ein Standard für einheitliche Digitalisierungsregeln für spätere ML-Nutzung ist essenziell, um zukünftig alle Dokumente für die Einbindung von Algorithmen nutzbar zu machen, ohne sie in ihrer Form erneut überarbeiten zu müssen (vgl. DFG 2016, S. 35: wissenschaftliche Nachnutzbarkeit). Erst dadurch wird Rechtssicherheit für Ersetzendes Scannen erreicht und Doppelarchivierung vermieden. Obwohl Handschrifterkennung bereits fortgeschritten ist und fortwirkend verbessert wird, sollten Regeln für den Umgang mit Papieroriginalen beschrieben werden. Auch die Weiterentwicklung der OCR bietet gesteigerte Möglichkeiten (5.1.2). Die automatisierte Segmentierung von Dokumenten ist optimierbar und kann Scangeschwindigkeiten und die automatisierte Ablage wesentlich beschleunigen. Hierzu muss die Datenbasis erweitert, qualitativ verbessert und das Segmentierungsmodell weiterentwickelt werden.

Priorisierung der Dokumentenklassen für die definierten Anwendungsbereiche, die Methodik und Kriterien sollen auf andere Anwendungsbereiche übertragbar sein: Wie können die Dokumentenklassen für die spezifizierten Anwendungsbereiche priorisiert werden, um Effizienz und Effektivität einer maschinenbasierten Informationsextraktion sicherzustellen?

Die Dokumentenklassifikation und die damit verbundenen Klassifizierungssysteme sollten vereinheitlicht und standardisiert werden, ohne zukünftige Erweiterungen einzuschränken. Bei der Auswertung großer Datenmengen ermöglichen die einheitlichen Dokumentenklassen die uneingeschränkte Anwendung ML-basierter Algorithmen. Trotz der Klassenzuordnung bleibt ein Clustering als sinnvolle Alternative möglich. Schlussfolgernd kann festgehalten werden, dass durch die Einbindung von Dokumentenklassen alle Möglichkeiten der weiteren Verarbeitung offengehalten werden. Die zunehmende Digitalisierung wird künftig weitere Lösungen bieten, wie durch den Einsatz von Wissensgraphen.

Spezifikation der wesentlichen Dokumentenklassen für die Anwendungsbereiche Energieeffizienz und Lebenszyklusanalysen: Welche wesentlichen Dokumentenklassen sind relevant für die spezifizierten Anwendungsbereiche Energieeffizienz und Lebenszyklusanalysen?

Die in *ML-BAU-DOK* entwickelten Methodiken zur Spezifizierung der Dokumentenklassen sind über alle Klassen anwendbar. Die beschriebenen Systeme, Scoring oder Active Learning, müssen jedoch flächendeckend eingesetzt werden, um die Spezifikation für alle Anwendungsfälle zu definieren. Durch die Spezifikation von Anwendungsfällen sind die Schlüsselinformationen aus den priorisierten Dokumentenklassen dokumentiert und können für weiterführende Informationsextraktionsprozesse genutzt werden.

Darlegung der Möglichkeiten für (automatisches) Klassifizieren von gebäude- und anlagenbezogenen Dokumenten: Welche Möglichkeiten bestehen für das (automatische) Klassifizieren von gebäude- und anlagenbezogenen Dokumenten?

Die Steigerung der Datenbasis auf einige Millionen von Dokumenten ermöglicht theoretisch die Anwendung modernster Modelle, wie beispielsweise Transformermodelle und LSTM. Diese schaffen zusätzlich zur herkömmlichen Textauswertung insbesondere auch Auswertungen nach Kontext. Mit sehr großen Trainings- und Testdatenmengen kann dann die Methode des Clustering ersetzt werden.

Zusammenfassend kann festgehalten werden, dass die standardisierten Dokumentenklassen, eine ausreichende Zahl an Trainingsdaten und die Erweiterung der Modelle Voraussetzungen für erfolgreiche ML-basierte Klassifizierung sind. Sollte einer der Punkte nicht erfüllt sein, stellt Clustering eine Alternative mit überschaubarem Pre-Processing-Aufwand dar.

Die standardisierte Dokumentenklassifizierung ist Voraussetzung für die effektive und zielführende Informationsextraktion aus Dokumenten der Bau- und Immobilienwirtschaft und bildet die Grundlage für zukünftige ML-basierte Dokumentenauswertungen.

Mitwirkende

Autoren

Prof. Dr. Kurzrock, Björn-Martin (RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau)

Rothenbusch, Jonathan (RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau)

Schütz, Konstantin (RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau)

Huang, Feibai (RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau)

Projektpartner und weitere Fördermittelgeber

Architrave GmbH, Bouchéstraße 12, 12435 Berlin

Stiftung Kloster Eberbach, Kloster Eberbach, 65346 Eltville im Rheingau

Kurzbiographien

Prof. Dr. Björn-Martin Kurzrock
Björn-Martin Kurzrock ist Professor für Immobilienökonomie an der RPTU. Er leitet das Fachgebiet seit 2008 mit dem Fokus auf Immobilienentwicklung und (digitales) Immobilienmanagement. Seit 2015 koordiniert er den B.Sc. und M.Sc. Studiengang IFMT Immobilien und Facilities – Management und Technik. Er ist amtierender Präsident und Board Member der European Real Estate Society (ERES). Stationen davor waren u.a. im Bereich Portfolioanalyse und als Head of Research bei IPD (heute: MSCI).

M. Sc. Jonathan Rothenbusch
Jonathan Rothenbusch ist Wissenschaftlicher Mitarbeiter am Fachgebiet Immobilienökonomie der RPTU. Nach seinem Studium des Facility Management an der RPTU und Stationen im Immobilienmanagement forscht Jonathan Rothenbusch heute im Bereich der automatisierten Erstellung technischer Anlagenregister.

M. Sc. Konstantin Schütz
Konstantin Schütz war bis November 2022 Wissenschaftlicher Mitarbeiter am Fachgebiet Immobilienökonomie der RPTU. Nach seinem Studium des Facility Management an der RPTU und Stationen im Immobilienmanagement arbeitet er fortan wieder im Bereich des Corporate Real Estate Managements.

M. Sc. Feibai Huang
Feibai Huang war bis Dezember 2022 Wissenschaftlicher Mitarbeiter am Fachgebiet Immobilienökonomie und der Arbeitsgruppe Machine Learning der RPTU. Nach seinem Studium der Mathematik mit dem Schwerpunkt Stochastik und Analysis an der RPTU arbeitete Feibai Huang im Versicherungswesen, wo er heute wieder tätig ist.

Literaturverzeichnis

Althnian, Alhanoof/AlSaeed, Duaa/Al-Baity, Heyam/Samha, Amani/Dris, Alanoud Bin/Alzakari, Najla/Abou Elwafa, Afnan/Kurdi, Heba (2021) Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. In: Applied Sciences, Jg. 11, Nr. 2, S. 796.

Baierer, Konstantin/Zumstein, Philipp Verbesserung der OCR in digitalen Sammlungen von Bibliotheken. Online verfügbar unter: https://madoc.bib.uni-mannheim.de/41442/1/Baierer-Zumstein-2016.pdf, zuletzt geprüft am 07.12.2022.

Bawden, D./Robinson, L. (2009) The dark side of information. Overload, anxiety and other paradoxes and pathologies. Online verfügbar unter: https://www.semanticscholar.org/paper/The-dark-side-of-information%3A-overload%2C-anxiety-and-Bawden-Robinson/0d22571135a4cae770ce86c7fe5ec46a3feb0a75, zuletzt geprüft am 01.07.2021.

BMWi (2019) Energieeffizienzstrategie 2050. Online verfügbar unter: https://www.bmwi.de/Redaktion/DE/Publikationen/Energie/energieeffiezienzstrategie-2050.pdf?__blob=publicationFile&v=12, zuletzt geprüft am 11.08.2021.

Brockmann, Tanja; Figl, Hildegund; Huemer-Kals, Veronika; Kusche, Oliver; Kerz, Nicolas; Rössig, Stephan (2019) ÖKOBAUDAT; Grundlage für die Gebäudeökobilanzierung, 2. Aufl., Bonn.

Cai, Li/Zhu, Yangyong (2015) The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Online verfügbar unter: https://pdfs.semanticscholar.org/0fb3/7330a4170ec63d60eec7dbb2b86e6829a3de.pdf, zuletzt geprüft am 07.12.2022.

DFG (2016) DFG-Praxisregeln; "Digitalisierung". Online verfügbar unter: https://www.dfg.de/formulare/12_151/12_151_de.pdf, zuletzt geprüft am 06.07.2021.

DGNB (2018) Leitfaden zum Einsatz der Ökobilanzierung. Online verfügbar unter: https://www.dgnb.de/de/verein/publikationen/bestellung/downloads/DGNB_Report_LCA_Leitfaden.pdf, zuletzt geprüft am 07.12.2022.

Diaz-Bone, Rainer/Weischer, Christoph (2015) Methoden-Lexikon für die Sozialwissenschaften, Wiesbaden.

DIN 14040 DIN EN ISO 14040; Umweltmanagement - Ökobilanz - Grundsätze und Rahmenbedingungen. Fassung vom 02.2021.

Fischer, Matthias Ganzheitliche Bilanzierung. Online verfügbar unter: https://www.ibp.fraunhofer.de/content/dam/ibp/ibp-neu/de/dokumente/broschueren/gabi/abteilungsbroschuere-ganzheitliche-bilanzierung.pdf, zuletzt geprüft am 07.12.2022.

Frischknecht, Rolf (2020) Lehrbuch der Ökobilanzierung, Berlin, Heidelberg.

gif e.V. (2021) Dokumentenmanagement-System (DMS). Online verfügbar unter: https://www.gif-ev.de/glossar/view_contact/473, zuletzt geprüft am 28.07.2021.

Gödert, Winfried/Lepsky, Klaus/Nagelschmidt, Matthias (2012) Informationserschließung und Automatisches Indexieren; Ein Lehr- und Arbeitsbuch, Berlin/Heidelberg.

Goldberg, Yoav (2017) Neural Network Methods for Natural Language Processing. In: Synthesis Lectures on Human Language Technologies, Jg. 10, Nr. 1, S. 1–309.

Götzer, Klaus/Maier, Berthold/Schmale, Ralf/Rehbock, Klaus/Komke, Torsten (2014) Dokumenten-Management; Informationen im Unternehmen effizient nutzen, 5. Aufl., Heidelberg.

Heaton, Jeff (2018) Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. In: Genetic Programming and Evolvable Machines, Jg. 19, 1-2, S. 305–307.

John, Viola/Gut, Silvan/Wallbaum, Holger (2010) Hoch oder quer? Ökologische Lebenszyklusanalyse eines Hochhauses im Vergleich zu einem Riegelbau. Online verfügbar unter: zuletzt geprüft am 08.12.2022.

Keller, Stefan Andreas (2014) Wissensorganisation und -repräsentation mit digitalen Technologien, Berlin.

Kipp, Margaret E.I./Campbell, D. Grant (2006) Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. In: Proceedings of the American Society for Information Science and Technology, Jg. 43, Nr. 1, S. 1–18.

Klöpffer, Walter/Grahl, Birgit (2012) Ökobilanz (LCA); Ein Leitfaden für Ausbildung und Beruf, Weinheim.

König, Holger (2017) Projekt: Lebenszyklusanalyse von Wohngebäuden; Lebenszyklusanalyse mit Berechnung der Ökobilanz und Lebenszykluskosten. Endbericht. Online verfügbar unter: https://www.bauinnung-nuernberg.de/fileadmin/quicklinks/Quick-Link-Nr-98300000-LfU-Inhalt-Lebenszyklusanalyse.pdf, zuletzt geprüft am 07.12.2022.

König, Holger/Kohler, Niklaus/Kreißig, Johannes/Lützkendorf, Thomas (2012) Lebenszyklusanalyse in der Gebäudeplanung; Grundlagen Berechnung Planungswerkzeuge, München.

Krcmar, Helmut (2005) Informationsmanagement, Berlin/Heidelberg.

Krüger, Jochen/Sorge, Christoph/Vogelgesang, Stephanie (2016) Ersetzendes Scannen; Kernelement im Gesamtkonzept einer elektronischen Aktenführung? Online verfügbar unter: https://www.uni-saarland.de/fileadmin/upload/lehrstuhl/sorge/Paper-Downloads/IRIS-2016-Ersetzendes-Scannen.pdf, zuletzt geprüft am 08.07.2021.

Kyocera Document Solutions Deutschland GmbH (2018) Rechtskonform dank DMS; Wie KMU die Vorgaben der DSGVO erfüllen. Online verfügbar unter: https://c.kyoceradocumentsolutions.eu/e/325011/ook-Rechtskonform-dank-DMS-pdf/8llxj/1488611700?h=PyVuUj1yzh1uUvdJWGm_oJHvnFsTEDbTrZ-ZJCE4gw8, zuletzt geprüft am 07.12.2022.

LeCun, Yann/Bengio, Yoshua (1995) Convolutional Networks for Images, Speech, and Time-Series. Online verfügbar unter: http://www.iro.umontreal.ca/~lisa/pointeurs/handbook-convo.pdf, zuletzt geprüft am 16.11.2022.

Lin, Yiming/Wang, Hongzhi/Li, Jianzhong/Gao, Hong (2019) Data source selection for information integration in big data era. In: Information Sciences, Jg. 479, Nr. 1, S. 197–213.

Manning, Christopher D./Raghavan, Prabhakar/Schütze, Hinrich (2009) Introduction to information retrieval, Cambridge.

May, Michael; Krämer, Markus; Schlundt, Maik (2022) BIM im Immobilienbetrieb; Anwendung, Implementierung, Digitalisierungstrends und Fallstudien, Wiesbaden.

May, Sibylle (2011) Das Checklistenbuch; Die wichtigsten Organisationshilfen für das Büromanagement, Wiesbaden.

Mikolov, Tomas/Chen, Kai/Corrado, Greg/Dean, Jeffrey (2013) Efficient Estimation of Word Representations in Vector Space. Online verfügbar unter: http://arxiv.org/pdf/1301.3781v3, zuletzt geprüft am 07.12.2022.

Müller, Philipp Maximilian/Päuser, Philipp/Kurzrock, Björn-Martin (2021) Fundamentals for automating due diligence processes in property transactions. Online verfügbar unter: https://www.emerald.com/insight/content/doi/10.1108/JPIF-09-2019-0130/full/html, zuletzt geprüft am 07.12.2022.

O'Shea, Keiron/Nash, Ryan (2015) An Introduction to Convolutional Neural Networks. Online verfügbar unter: http://arxiv.org/pdf/1511.08458v2, zuletzt geprüft am 07.12.2022.

phuong (2013) Diagram of an artificial neural network. Online verfügbar unter: https://tex.stackexchange.com/questions/132444/diagram-of-an-artificial-neural-network, zuletzt geprüft am 08.12.2022.

Ramseier, Livia/Frischknecht, Rolf (2020) Umweltfußabdruck von Gebäuden in Deutschland; Kurzstudie zu sektorübergreifenden Wirkungen des Handlungsfelds „Errichtung und Nutzung von Hochbauten“ auf Klima und Umwelt. Online verfügbar unter: https://www.bbsr.bund.de/BBSR/DE/veroeffentlichungen/bbsr-online/2020/bbsr-online-17-2020-dl.pdf?__blob=publicationFile&v=3, zuletzt geprüft am 05.08.2021.

Reimann, Norbert (2004) Praktische Archivkunde; Ein Leitfaden für Fachangestellte für Medien- und Informationsdienste, Fachrichtung Archiv, Münster.

RICS (2017) Global Trends in Data Capture and Management in Real Estate and Construction. Online verfügbar unter: https://www.rics.org/globalassets/rics-website/media/knowledge/research/insights/global-trends-in-data-capture-and-management-in-real-estate-and-construction-rics.pdf, zuletzt geprüft am 0712.2022.

Rodeck, Martin/Schulz-Wuulkow, Christian/Hellmuth, ALexander/Seyler, Nicolas Digitalisierungsstudie ZIA und EY Real Estate 2021 Erfolgsfaktoren Automatisierung. Online verfügbar unter: https://zia-deutschland.de/wp-content/uploads/2021/08/2021_EY-Real-Estate-ZIA_Digitalisierungsstudie_Erfolgsfaktor_Automatisierung_final.pdf, zuletzt geprüft am 06.12.2022.

Rössig, Stephan Ergänzung des digitalen Workflow in der Gebäudeplanung. Online verfügbar unter: https://www.bauteileditor.de/docs/publications/BBB_07-08-2018_Bauforschungsserie.pdf, zuletzt geprüft am 15.10.2021.

Schmude, Kathrin/Goebel, Christine/Hambuch, Manuela (2020) Archivisch für Anfänger; 25 Fachbegriffe einfach erklärt. Online verfügbar unter: https://www.bundesarchiv.de/DE/Content/Meldungen/2020-06-26_meldung-archivglossar.html, zuletzt geprüft am 07.12.2022.

Schroeder, Alan T. (2006) Digitizing a real estate document library. In: Records Management Journal, Jg. 16, Nr. 1, S. 34–50.

Sprague, Ralph H. (1995) Electronic Document Management: Challenges and Opportunities for Information Systems Managers. In: MIS Quarterly, Jg. 19, Nr. 1, S. 29–49.

Ugale, Mahendra K./Patil, Shweta J./Musande, Vijaya B. (2017) Document management system: A notion towards paperless office. Online verfügbar unter:

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8122176, zuletzt geprüft am 29.06.2021.

Varga, Daniel (2016) word2vec-web-visualization. Online verfügbar unter: https://camo.githubusercontent.com/fe1a0894679e139a0eb8ebc59d692c496aca44260101c8ed6a44f8c8ed0b675c/687474703a2f2f7777772e72656e79692e68752f7e64616e69656c2f696d616765732f676c6f76652d706f7765722e706e67, zuletzt geprüft am 08.12.2022.

Vaswani, Ashish/Shazeer, Noam/Parmar, Niki/Uszkoreit, Jakob/Jones, Llion/Gomez, Aidan N./Kaiser, Lukasz/Polosukhin, Illia (2017) Attention Is All You Need. Online verfügbar unter: http://arxiv.org/pdf/1706.03762v5, zuletzt geprüft am 07.12.2022.

Zink, Wolfgang/Crantz, Carsten/Grabner, Wiegand/Zöller, Bernhard/Halstenbach, Volker (2015) TR Resiscan: Beschleuniger oder Bremse für E-Government? Online verfügbar unter: https://www.pwc.de/de/offentliche-unternehmen/assets/pwc-tr-resiscan-beschleuniger-oder-bremse-fuer-e-government-2015.pdf, zuletzt geprüft am 09.07.2021.

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

BBSR ..... *Bundesinstitut für Bau-, Stadt- und Raumforschung*
BMWSB..... *Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen*
CAFM ..... *Computer Aided Facility Management*
CNN ..... *Convolutional Neural Network*
DFG..... *Deutschen Forschungsgemeinschaft*
DMS..... *Dokumentenmanagementsystem*
dpi..... *dots per inch*
EDV ..... *elektronische Datenverarbeitung*
FIDJI ..... *Format d'Inter-echanges de Données Juridiques et Immobilières; Financial and Property Data Interchange Format*
gefma ..... *Deutscher Verband für Facility Management*
GEG..... *Gebäudeenergiegesetz*
gif..... *Gesellschaft für immobilienwirtschaftliche Forschung*
GoBD .... *Grundsätze zur ordnungsmäßigen Führung und Aufbewahrung von Büchern, Aufzeichnungen und Unterlagen in elektronischer Form sowie zum Datenzugriff*
HGB ..... *Handelsgesetzbuch*
IPF..... *Investment Property Forum*
KI *Künstliche Intelligenz*
LCA..... *Life Cycle Assessment*
LCC..... *Life Cycle Costs*
LCE..... *Life Cycle Engineering*
LCIA..... *life cycle impact assessment*
LSTM..... *Long Short-Term Memory Networks*
ML..... *Maschinelles Lernen*
NGF..... *Nettogrundfläche*
OCR..... *Optical Character Recognition*
RICS..... *Royal Institution of Chartered Surveyors*
SKE..... *Stiftung Kloster Eberbach*
SPR..... *Society of Property Researchers*
tf-idf..... *term Frequency - inverse document frequency*
Vogon ..... *Vereniging Onroerend Goed Onderzoekers Nederland*

Anhang

Tabelle 16: Dokumentenklassen nach Mueller et al. (2023)[107]

Dokumentenklassen nach Mueller et al. 2023	
Label	**Dokumentenklassen**
01-001	Exposé
01-002	Objektbeschreibung
01-003	Objektfotos
01-004	Luftbild
01-005	Visualisierungen
02-001	An- und Verkaufspräsentation
02-002	An- und Verkaufskalkulation
02-003	Maklerdokumentation
02-004	Q&A
02-005	Grundstückskaufvertrag
02-006	Bezugsurkunde
02-007	Eigentumsübergabe
02-008	Enthaftungserklärung'Vertragserfüllungsbürgschaft
02-009	Weitere Unterlagen Grundstückskauf
02-010	Korrespondenz Anwalt und Notar
02-011	Grunderwerbssteuer
02-012	Kauf- und Verkaufsvertrag
02-013	Notarielle Beurkundung
03-001	Amtlicher Lageplan
03-002	Flurkarte'Liegenschaftskarte
03-003	Liegenschaftsbuch'-kataster
03-004	Grenzbescheinigung'Grenzfeststellung'Grundstücksteilung
03-005	Überbauung'Nutzungsrecht
03-006	Altlastenkataster'Altlastenauskunft
03-007	Altlastensanierung'Beseitigung
03-008	Kampfmittelbeseitigungsunterlagen
03-009	Gründungs- / Baugrundgutachten
03-010	Bodenrichtwertkarte
03-011	Baumbestand
03-012	Baureifmachung
03-013	Entsorgungsbilanz'Abbrucharbeiten
03-014	Hochwasserschutzkonzept
03-015	Investorenvertrag'archäologische Bodendenkmalpflege
04-001	Handelsregisterauszug Eigentümer
04-002	Grundbuchauszug

[107] Eigene Darstellung.

04-003	Gestattungsverträge
04-004	Pfandhaftentlassung
04-005	Erbbauchrechtsvertrag
04-006	Erbbauzins'Zahlungsnachweis
04-007	Nachbarschaftsvereinbarung
05-001	Teileigentum Übersicht
05-002	Teilungserklärung
05-003	Gemeinschaftsordnung
05-004	Aufteilungsplan
05-005	Abgeschlossenheitsbescheinigung
05-006	WEG-protokoll
05-007	Wirtschaftsplan
05-008	Instandhaltungsrücklage der Eigentumsgemeinschaft
05-009	WEG Verwaltervertrag
05-010	Hausordnung
06-001	Flächennutzungsplan
06-002	Bebauungsplan
06-003	Planungsrechtsauskunft
06-004	Bescheid Denkmalschutz
06-005	Auskunft Denkmalschutz
06-006	Gutachten Denkmalschutz
06-007	Konzept Denkmalschutz
06-008	Erschließungskosten'Abgabenbescheid
06-009	Städtebaulicher Vertrag
06-010	Umlegungsgebiet,Sanierungsgebiet,Entwicklungsgebiet
06-011	Erhaltungssatzung, Gestaltungssatzung
06-012	Behördliche Bescheinigungen
06-013	Regenentwässerung
07-001	Bauvoranfrage
07-002	Ausschreibungsunterlagen
07-003	Bauvorbescheid
07-004	Bauantrag
07-005	Landesbauordnung
07-006	Baugenehmigungsbescheid'Ergänzungsbescheid
07-007	Abbruchgenehmigungsbescheid
07-008	Behördliche Bauabnahme
07-009	Nutzungsgenehmigung / Nutzungsänderungsgenehmigung
07-010	Stellplatznachweis,Ablöse,Ausgleich
07-011	Bericht Prüfstatiker,Nachweise, Berechnung
08-001	Baubeschreibung
08-002	Bestandsplan,Lageplan,Abstandsflächen
08-003	Grundrissplan
08-004	Stellplatzplan
08-005	Ansichten'Bestandsplan
08-006	Schnitte'Bestandsplan
08-007	Detailplan
08-008	Bestandsplan Statik

08-009	Bestandsplan Heizung
08-010	Bestandsplan Lüftung
08-011	Bestandsplan Klimatechnik
08-012	Bestandsplan Sanitär
08-013	Bestandsplan Elektro
08-014	Bestandsplan Medien
08-015	Bestandsplan Leitungen
08-016	Schließplan, Schlüssel und Zugangskarten
08-017	Flächenberechnung
08-018	Flächenaufmaß'Pläne
09-001	Sachverständigen-Berichte
09-002	Technische Due Diligence Gutachten
09-003	Kaufmännische Due Diligence
09-004	Rechtliche Due Diligence
09-005	Steuerliche Due Diligence
09-006	Finanzielle Due Diligence
09-007	Umwelt Due Diligence Gutachten
09-008	Gebäude-Gutachten'Beweissicherungsverfahren'Bausubstanzgutachten
09-009	Trinkwassergutachten
09-010	Schallschutzgutachten
09-011	Grundleitungen Gutachten
09-012	Dichtheitsprüfung Abwasserrohrleitungen
09-013	Altlasten-Gutachten, Schadstoffe
09-014	Inspektionsberichte Gebäude
09-015	Umwelt-Zertifikat
09-016	Energieausweis
09-017	Wärmeschutznachweis
09-018	Kühllastermittlung
09-019	Gebäudezertifizierung
09-020	Verkehrswertgutachten'Wertgutachten
10-001	Auflagen Brandschutz
10-002	Brandschutzabnahme'Brandverhütungsschau
10-003	Brandschutzordnung
10-004	Brandschutzkonzept
10-005	Brandschutzgutachten
10-006	Genehmigung Abweichungsantrag 'Protokoll
10-007	Prüfbericht Sprinkleranlage
10-008	Feuerwehrplan
10-009	Fluchtwegeplan
10-010	Interaktionsprüfungen
10-011	Brandfallsteuermatrix
11-001	Planer Fachplaner Vertrag
11-002	GU-GÜ-Vertrag
11-003	Instandhaltungsplan
11-004	Übersicht Baumaßnahmen
11-005	Ausstehende Reparaturen - Mängel
11-006	Geleistete Reparaturen - Mängelbeseitigungsnachweis

11-007	Baurechnungen
11-008	Abnahme'Bauabnahme'Gewerkeabnahme
11-009	Abnahme'Schlussabnahme
11-010	TÜV-Zeugnisse
11-011	Fertigstellungsanzeige
11-012	Gewährleistungsbürgschaft
11-013	Gewährleistungsübersicht
11-014	Fachbauleitererklärung
11-015	Fachunternehmererklärung
11-016	Bauaufsichtliche Zulassung
11-017	Bauprodukte
12-001	Betriebsbeschreibung
12-002	Übersicht Gebäudetechnik
12-003	Anlagenbeschreibung'Bedienungsanleitung Heizung
12-004	Anlagenbeschreibung'Bedienungsanleitung Lüftung
12-005	Anlagenbeschreibung'Bedienungsanleitung Klima
12-006	Anlagenbeschreibung'Bedienungsanleitung Sanitär
12-007	Anlagenbeschreibung'Bedienungsanleitung Brandschutztechnik
12-008	Anlagenbeschreibung'Bedienungsanleitung Aufzug- und Fördertechnik
12-009	Anlagenbeschreibung'Bedienungsanleitung Elektro
12-010	Anlagenbeschreibung'Bedienungsanleitung USV
12-011	Anlagenbeschreibung'Bedienungsanleitung Sonstige
12-012	Abnahmeprotokoll'Prüfprotokoll Heizung
12-013	Abnahmeprotokoll'Prüfprotokoll Lüftung
12-014	Abnahmeprotokoll'Prüfprotokoll Klima
12-015	Abnahmeprotokoll'Prüfprotokoll Sanitär
12-016	Abnahmeprotokoll'Prüfprotokoll Brandschutztechnik
12-017	Abnahmeprotokoll'Prüfprotokoll Aufzug- und Fördertechnik
12-018	Abnahmeprotokoll'Prüfprotokoll Elektro
12-019	Abnahmeprotokoll'Prüfprotokoll USV
12-020	Abnahmeprotokoll'Prüfprotokoll Sonstige
12-021	Wartungsplan
12-022	Wartungsprotokoll Heizung
12-023	Wartungsprotokoll Lüftung
12-024	Wartungsprotokoll Abluftanlage
12-025	Wartungsprotokoll Umluftkühlgeräte
12-026	Wartungsprotokoll Klima
12-027	Wartungsprotokoll Sanitär
12-028	Wartungsprotokoll Brandschutztechnik
12-029	Wartungsprotokoll Brandmeldeanlage
12-030	Wartungsprotokoll Brandschutzklappe
12-031	Wartungsprotokoll Brandschutztüren
12-032	Wartungsprotokoll Brandschutztore
12-033	Wartungsprotokoll Hydranten
12-034	Wartungsprotokoll Feuerlöscher
12-035	Wartungsprotokoll Sprinkleranlage

12-036	Wartungsprotokoll Aufzug- und Fördertechnik
12-037	Wartungsprotokoll Elektro
12-038	Wartungsprotokoll Elektroakustische Anlage
12-039	Wartungsprotokoll Sicherheitsbeleuchtung
12-040	Wartungsprotokoll Blitzschutz
12-041	Wartungsprotokoll USV
12-042	Wartungsprotokoll Notrstromaggregat
12-043	Wartungsprotokoll Gebäudeleittechnik
12-044	Wartungsprotokoll CO Warnanlage
12-045	Wartungsprotokoll Raumlufttechnische Anlage
12-046	Wartungsprotokoll Rauch- und Wärmeabzugsanlage
12-047	Wartungsprotokoll Fettabscheider
12-048	Wartungsprotokoll Motoren
12-049	Wartungsprotokoll Schranke
12-050	Wartungsprotokoll Trinkwasserprüfung
12-051	Wartungsprotokoll Sonstige
13-001	Mieterliste
13-002	Belegungsplan'Mietflächenplan
13-003	Stammdatenblatt
13-004	Mustermietvertrag
13-005	Mieterbaubeschreibung
13-006	Handelsregisterauszüge Gewerbemieter
13-007	Kreditauskünfte Gewerbemieter
13-008	Mieterselbstauskunft
13-009	Bonitätsprüfung'Schufa
13-010	Übersicht Mietkaution
13-011	Kautionsbelege
13-012	Übergabeprotokoll
13-013	Schlüsselübergabe
13-014	Mietvertrag Wohnraum
13-015	Gewerbemietvertrag
13-016	Pachtvertrag
13-017	Untermietvertrag
13-018	Stellplatzmietvertrag
13-019	Vertrag Werbeflächen
13-020	Antennenstellflächen
13-021	Sonstige Nutzungsvereinbarungen
13-022	Sonstige Vereinbarungen
13-023	Nachtrag
13-024	Mietbürgschaft
13-025	Mietrückstände Übersicht
13-026	Mietrückstände Mieter
13-027	Mietanpassung
13-028	Dauermietrecht
13-029	Mietminderung
13-030	Optionsausübung
13-031	Sonstiger Schriftverkehr Mieter

13-032	SEPA Mandat
13-033	Mängelrüge Mieter
13-034	Kündigung
13-035	Kündigungsbestätigung Mietvertrag
13-036	Nebenkostenabrechnung Mieter
13-037	Widerspruch Nebenkostenabrechnung Mieter
14-001	Bewirtschaftungsverträge
14-002	Bewirtschaftungsrechnungen
14-003	Betriebskosten'Übersicht
14-004	Dienstleistungsvertrag Propertymanagement
14-005	Dienstleistungsvertrag Facility Management
14-006	Dienstleistungsvertrag Center Management
14-007	Dienstleistungsvertrag Hausmeister
14-008	Dienstleistungsvertrag Hausverwaltung
14-009	Dienstleistungsvertrag Reinigung
14-010	Dienstleistungsvertrag Sicherheit
14-011	Dienstleistungsvertrag Winterdienst
14-012	Hausverwaltervollmacht
14-013	Wartungs- und Dienstleistungsvertrag Heizung
14-014	Wartungs- und Dienstleistungsvertrag Lüftung
14-015	Wartung- und Dienstleistungsvertrag Abluftanlage
14-016	Wartungs- und Dienstleistungsvertrag Klima
14-017	Wartungs- und Dienstleistungsvertrag Sanitär
14-018	Wartungs- und Dienstleistungsvertrag Brandschutztechnik
14-019	Wartung- und Dienstleistungsvertrag Brandmeldeanlage
14-020	Wartung- und Dienstleistungsvertrag Brandschutzklappe
14-021	Wartung- und Dienstleistungsvertrag Brandschutztüren
14-022	Wartung- und Dienstleistungsvertrag Brandschutztore
14-023	Wartung- und Dienstleistungsvertrag Hydranten
14-024	Wartung- und Dienstleistungsvertrag Feuerlöscher
14-025	Wartung- und Dienstleistungsvertrag Sprinkleranlage
14-026	Wartung- und Dienstleistungsvertrag Rauch- und Wärmeabzugsanlage
14-027	Wartung- und Dienstleistungsvertrag Fluchtwegsteuerung
14-028	Wartungs- und Dienstleistungsvertrag Aufzug- und Fördertechnik
14-029	Wartungs- und Dienstleistungsvertrag Gebäudeleittechnik
14-030	Wartungs- und Dienstleistungsvertrag Elektro
14-031	Wartung- und Dienstleistungsvertrag Elektrotechnische Anlage
14-032	Wartung- und Dienstleistungsvertrag Sicherheitsbeleuchtung
14-033	Wartung- und Dienstleistungsvertrag Blitzschutz
14-034	Wartung- und Dienstleistungsvertrag Stromerzeuger
14-035	Wartungs- und Dienstleistungsvertrag USV
14-036	Wartung- und Dienstleistungsvertrag Notstromaggregat
14-037	Wartung- und Dienstleistungsvertrag Trafostation
14-038	Wartung- und Dienstleistungsvertrag Schrankenanlage
14-039	Wartung- und Dienstleistungsvertrag Sonnenschutz
14-040	Wartung- und Dienstleistungsvertrag Tore/ Rolltor
14-041	Wartung- und Dienstleistungsvertrag Dach

14-042	Wartung- und Dienstleistungsvertrag Sonstige
15-001	Ver- und Entsorgungsvertrag
15-002	Versorgungsvertrag Wärme und Kälte, lufttechnische Versorgung
15-003	Versorgungsvertrag Strom
15-004	Versorgungsvertrag Gas
15-005	Versorgungsvertrag Fernwärme
15-006	Versorgungsvertrag Wasser
15-007	Versorgungsvertrag Daten'Telekommunikation
15-008	Verbrauchsabrechnung Ver- und Entsorgung
15-009	Verbrauchsabrechnung Strom
15-010	Verbrauchsabrechnung Gas
15-011	Verbrauchsabrechnung Fernwärme
15-012	Verbrauchsabrechnung Wasser
15-013	Verbrauchsabrechnung Daten'Telekommunikation
16-001	Übersicht Versicherungen
16-002	Gebäudehaftpflichtversicherung
16-003	Feuerversicherung
16-004	Terrorversicherung
16-005	Mietausfallversicherung
16-006	Umweltschadenversicherung
16-007	Gebäudesachversicherung
16-008	Sonstige Versicherungsunterlagen
16-009	Versicherungsschäden
16-010	Geltend gemachte Versicherungsschäden
17-001	Grundsteuerbescheid
17-002	Einheitswertbescheid
17-003	Straßenreinigungsgebühren
17-004	Abwassergebühr, Oberflächenentwässerung, Schmutzwasserbeseitigung
17-005	Abfallgebühr, Entsorgung
17-006	Sonstige Grundbesitzabgabenbescheide
17-007	Sonstige Kommunalabgaben
17-008	Vorsteuerberichtigung
17-009	Körperschaftssteuer
17-010	Umsatzsteuer
17-011	Umsatzsteuerschlüssel
17-012	Freistellungsbescheinigung Bauleistungen
17-013	Bescheinigung in Steuersachen
17-014	Mieter-Vorsteuer
17-015	Verkäufer-Vorsteuer
17-016	Optionssätze, Flächenschlüssel
18-001	Rechtsstreitigkeiten
18-002	Objektbezogene Arbeitsverträge
18-003	Zusatzinformationen für Spezialimmobilien
18-004	Sonstige Dokumente, Verträge, Vereinbarungen
19-001	Marktanalyse
19-002	Standortanalyse
19-003	Infrastruktur und Anbindung

19-004	Umfeldanalyse
19-005	Marketingmaßnahmen
19-006	Pressemitteilungen
19-007	Mietermarketing Maßnahmen
19-008	Bilder
20-001	Controlling Bericht
20-002	Ausgangsrechnungen
20-003	Eingangsrechnungen
20-004	Baubudget Kontrolle
20-005	Liquiditätsabrufe
21-001	Forecast Kalkulation
21-002	Projektupdates
21-003	Monatsreport
21-004	Quartalsreport
21-005	Financial Report
21-006	Management Report

Tabelle 17: Gesamtauswertung Dokumentenklassen für den Anwendungsfall Energieeffizienz- und Lebenszyklusanalyse[108]

	OCR	Energieeffizienz- und Lebenszyklusanalysen
Label	Maschinenlesbarkeit [%]	Relevanz Anwendungsfälle (Energieeffizienz und Lebenszyklusanalyse) [%]
01-001	97%	29%
01-002	97%	11%
01-003	97%	0%
01-004	97%	0%
01-005	97%	0%
02-001	97%	15%
02-002	97%	10%
02-003	97%	26%
02-004	97%	0%
02-005	97%	0%
02-006	97%	0%

[108] Eigene Darstellung.

02-007	97%	0%
02-008	97%	0%
02-009	97%	1%
02-010	97%	0%
02-011	97%	0%
02-012	97%	14%
02-013	97%	0%
03-001	97%	0%
03-002	97%	0%
03-003	97%	0%
03-004	97%	0%
03-005	97%	0%
03-006	97%	0%
03-007	97%	0%
03-008	97%	0%
03-009	97%	0%
03-010	97%	0%
03-011	97%	0%
03-012	97%	0%
03-013	97%	0%
03-014	97%	0%
03-015	97%	0%
04-001	97%	0%
04-002	97%	5%
04-003	97%	0%
04-004	97%	0%
04-005	97%	10%
04-006	97%	0%
04-007	97%	0%
05-001	97%	4%
05-002	97%	3%
05-003	97%	0%
05-004	97%	0%
05-005	97%	0%
05-006	97%	9%
05-007	97%	5%
05-008	97%	0%
05-009	97%	1%
05-010	97%	0%
06-001	97%	0%
06-002	97%	0%
06-003	97%	0%
06-004	97%	5%
06-005	97%	1%
06-006	97%	5%
06-007	97%	1%
06-008	97%	0%

06-009	97%	0%
06-010	97%	0%
06-011	97%	0%
06-012	97%	0%
06-013	97%	0%
07-001	97%	5%
07-002	97%	6%
07-003	97%	5%
07-004	97%	43%
07-005	97%	0%
07-006	97%	28%
07-007	97%	0%
07-008	97%	6%
07-009	97%	13%
07-010	97%	0%
07-011	97%	0%
08-001	97%	11%
08-002	97%	0%
08-003	97%	3%
08-004	97%	0%
08-005	97%	0%
08-006	97%	0%
08-007	97%	0%
08-008	97%	0%
08-009	97%	13%
08-010	97%	3%
08-011	97%	5%
08-012	97%	6%
08-013	97%	3%
08-014	97%	0%
08-015	97%	0%
08-016	97%	0%
08-017	97%	10%
08-018	97%	8%
09-001	97%	1%
09-002	97%	56%
09-003	97%	18%
09-004	97%	10%
09-005	97%	10%
09-006	97%	15%
09-007	97%	73%
09-008	97%	0%
09-009	97%	0%
09-010	97%	3%
09-011	97%	0%
09-012	97%	0%
09-013	97%	0%

09-014	97%	0%
09-015	97%	66%
09-016	97%	100%
09-017	97%	10%
09-018	97%	4%
09-019	97%	76%
09-020	97%	26%
10-001	97%	3%
10-002	97%	0%
10-003	97%	0%
10-004	97%	3%
10-005	97%	3%
10-006	97%	0%
10-007	97%	0%
10-008	97%	0%
10-009	97%	0%
10-010	97%	0%
10-011	97%	0%
11-001	97%	1%
11-002	97%	5%
11-003	97%	13%
11-004	97%	0%
11-005	97%	0%
11-006	97%	0%
11-007	97%	0%
11-008	97%	14%
11-009	97%	11%
11-010	97%	0%
11-011	97%	0%
11-012	97%	0%
11-013	97%	0%
11-014	97%	0%
11-015	97%	0%
11-016	97%	0%
11-017	97%	0%
12-001	97%	0%
12-002	97%	18%
12-003	97%	13%
12-004	97%	8%
12-005	97%	10%
12-006	97%	11%
12-007	97%	0%
12-008	97%	0%
12-009	97%	8%
12-010	97%	0%
12-011	97%	0%
12-012	97%	20%

12-013	97%	8%
12-014	97%	15%
12-015	97%	11%
12-016	97%	0%
12-017	97%	0%
12-018	97%	8%
12-019	97%	0%
12-020	97%	0%
12-021	97%	0%
12-022	97%	18%
12-023	97%	8%
12-024	97%	4%
12-025	97%	8%
12-026	97%	15%
12-027	97%	11%
12-028	97%	0%
12-029	97%	0%
12-030	97%	0%
12-031	97%	0%
12-032	97%	0%
12-033	97%	0%
12-034	97%	0%
12-035	97%	0%
12-036	97%	0%
12-037	97%	8%
12-038	97%	0%
12-039	97%	0%
12-040	97%	0%
12-041	97%	0%
12-042	97%	0%
12-043	97%	0%
12-044	97%	0%
12-045	97%	8%
12-046	97%	0%
12-047	97%	0%
12-048	97%	0%
12-049	97%	0%
12-050	97%	0%
12-051	97%	0%
13-001	97%	0%
13-002	97%	0%
13-003	97%	0%
13-004	97%	0%
13-005	97%	0%
13-006	97%	0%
13-007	97%	0%

13-008	97%	0%
13-009	97%	0%
13-010	97%	0%
13-011	97%	0%
13-012	97%	0%
13-013	97%	0%
13-014	97%	1%
13-015	97%	1%
13-016	97%	0%
13-017	97%	0%
13-018	97%	0%
13-019	97%	0%
13-020	97%	0%
13-021	97%	0%
13-022	97%	0%
13-023	97%	0%
13-024	97%	0%
13-025	97%	0%
13-026	97%	0%
13-027	97%	0%
13-028	97%	0%
13-029	97%	0%
13-030	97%	0%
13-031	97%	0%
13-032	97%	0%
13-033	97%	0%
13-034	97%	0%
13-035	97%	0%
13-036	97%	29%
13-037	97%	0%
14-001	97%	0%
14-002	97%	0%
14-003	97%	9%
14-004	97%	4%
14-005	97%	4%
14-006	97%	4%
14-007	97%	1%
14-008	97%	4%
14-009	97%	4%
14-010	97%	1%
14-011	97%	1%
14-012	97%	0%
14-013	97%	14%
14-014	97%	5%
14-015	97%	3%
14-016	97%	10%
14-017	97%	9%

14-018	97%	0%
14-019	97%	0%
14-020	97%	0%
14-021	97%	0%
14-022	97%	0%
14-023	97%	0%
14-024	97%	0%
14-025	97%	0%
14-026	97%	0%
14-027	97%	0%
14-028	97%	0%
14-029	97%	0%
14-030	97%	5%
14-031	97%	0%
14-032	97%	0%
14-033	97%	0%
14-034	97%	0%
14-035	97%	0%
14-036	97%	0%
14-037	97%	0%
14-038	97%	0%
14-039	97%	0%
14-040	97%	0%
14-041	97%	0%
14-042	97%	0%
15-001	97%	0%
15-002	97%	10%
15-003	97%	8%
15-004	97%	5%
15-005	97%	5%
15-006	97%	0%
15-007	97%	0%
15-008	97%	0%
15-009	97%	20%
15-010	97%	20%
15-011	97%	20%
15-012	97%	0%
15-013	97%	0%
16-001	97%	3%
16-002	97%	10%
16-003	97%	10%
16-004	97%	10%
16-005	97%	10%
16-006	97%	10%
16-007	97%	10%
16-008	97%	1%
16-009	97%	0%

16-010	97%	0%
17-001	97%	1%
17-002	97%	1%
17-003	97%	0%
17-004	97%	0%
17-005	97%	0%
17-006	97%	0%
17-007	97%	0%
17-008	97%	0%
17-009	97%	0%
17-010	97%	0%
17-011	97%	0%
17-012	97%	0%
17-013	97%	0%
17-014	97%	0%
17-015	97%	0%
17-016	97%	3%
18-001	97%	0%
18-002	97%	0%
18-003	97%	0%
18-004	97%	0%
19-001	97%	0%
19-002	97%	0%
19-003	97%	0%
19-004	97%	0%
19-005	97%	0%
19-006	97%	0%
19-007	97%	0%
19-008	97%	0%
20-001	97%	10%
20-002	97%	0%
20-003	97%	0%
20-004	97%	0%
20-005	97%	0%
21-001	97%	0%
21-002	97%	0%
21-003	97%	1%
21-004	97%	1%
21-005	97%	4%
21-006	97%	5%

Tabelle 18: Beispielhafter Auszug der Prüfung der Maschinenlesbarkeit[109]

Seite	Buchstabe	Satz vor und nach Zielbuchstabe	Wort	Buch-stabe	Erkennung Wort	Erkennung Buchstabe	Fehlerquelle
20	219	/1 19 mm, dunkelbraun 1 Stk. 46,00 5065 Klemmfix Podestverbinder 1 Stk. 5,00 . -..--------. .---- 5066 Ersat	Podestverbinder	v			
74	1065	alternativ eingeplant. Weitere Details sind dem LV zu entnehmen. geschätzte Vergabesumme ca. 50.000 € Öfentliche A	entnehmen	t			
253	735	mit größter Sorgfalt, Vorsicht und weitestgehender Erhaltung der Bausubstanz durchzuführen,	weitestgehender	w			
272	13	qoY�� 7 /r _ l 3 Posteingang mit Be erbungsschreibent	Posteingang	o			verschmiert und mit Handschrift markiert
98	638	heit-Wandgerät, mit infrarot Fernbedienung, einschl. Wandhalterung Kühlleistung: max. 2,5 kW Schallleistungspegel: max	Wandhalterung	l			
196	1013	wid dskla ist sig. Die Wände müssen im Anschlussbereich eben sein. Geg enfall sind A eichsmaßnahmen mit nichtbrenn	eben	b			
148	23	2 7 Hnzipdarstellungen Wandkonstruktion F30 . F120 Anlage 14 zur Gutachterlichen Recht	Wandkonstruktion	o			
226	235	o o ur CD 60x27 0,25 ja nein a (paarweise) Co Anlage 3 _ .hängte Deckenkonstruktion zum ABP-Nr. ra n	Anlage	n			
198	1630	0 mm; bzw. der Feuerwiderstandsklasse „F 90"d>_20m Bei Deckensp en b' r R` cke ist das Anschlussprofil mit geeigne	Deckensp	k			diagonale Schrift, Erkennung richtig bis Störfaktor
118	184	Bauherr: Stiftung Kloster Eberbach Foto Nr.:57 Altes Hospital, Südseite, Apsisanbau, Westfassade,	Stiftung	t			
93	660	g Haustechnik, Herr Michel - hbm, Herr Denich - hbm, Herr Martin - Sigeko, Frau Langner - Gündling Ingenieure Datum: 31	Martin	r			
128	1665	t optional Flächen für die Anordnung von Entsorgungseinrichtungen und eines Streusalzbehälters. Der Gärtnerhof erstreck	Entsorgungs-einrichtungen	t			
322	150	V: 2012-07-10 NK . Gerüstbauarbeiten Titel Bezeichnung Seite 30. Baukonstruktion 8 30.10. Baustelleneinrichtu	Seite	S			
56	320	fung Grundangaben Vergabeunterlagen: ok 2-3 (Datum / Handz. / Stz.) Rückgabe zur Bearbeitung: (Datum / Handz. / St	Handz.	n			
312	716	30.4.20. Schuttmaterial, belastet 10,000 m' 30.4.30. Schuttmaterial, 5,000 m3 höherbelastet 30.4.40. Umsichtige	Schuttmaterial	h			
2	396	mit Vergabelaufzetel u. Nachtragsangebot 10 13. Auftragsschreiben 14. Vergabevorgang	Auftragsschreiben	g			
269	130	Baumaßnahme: A.0428.059105, H- Kloster Eberbach -San. der Klosterur. Vergabenummer: VG-A0428-2015-1290	Vergabenummer	e			

[109] Eigene Darstellung.

272	1222	Betriebskosten für die gesamte Bauzeit bis zur Verkehrsfreigabe. 1.2.30. Der für die Beschilderung	Verkehrsfreigabe	k			
108	117	hrling h • • Gesamt Material Unterschrift des Auftraggebers Sie erkennen mit Ihrer Unterschrift die oben aufgeführ	Auftraggebers	e			
172	979	il may contain confidential and/or privileged information. If you are not the intended cipient (or have received this e-	If	f			
148	649	Anhänge enthalten vertrauliche Informationen, welche ausschließlich für den/die oben erwähnten Empfänger bestimmt sind.	ausschließlich	l			
52	759	g mit dem Auftraggeber 7. Mitwirken beim Vertreten der Planungskonzeption mit bis zu 5 Erläuterungs- und/oder ' Erör	Planungs-konzeption	a			
84	2492	11). Zu denken wäre zwar an eine entsprechende Anwendung der Inhouse-Grundsätze auch auf invers-vertikale Beauftragunge	der	r			
180	0	0 • z	z	z			Trennblatt ohne Inhalt