

Materialdaten für die Umstellung von Erdgastransportleitungen auf Wasserstoff

Künstliche Intelligenz für die Auswertung von Materialzeugnissen

Von Devin Schumacher, Andreas Schwörer, Sebastian Wolters und Daniel S. Bick

Die Analyse von Materialzeugnissen mittels einer umfangreichen Toolchain bestehend aus Preprocessing, Spracherkennung, Outputchannel, Postprocessing sowie der Bereinigung der Daten und der Datenanalyse lieferte wichtige Erkenntnisse. Über 1,6 Mio. DIN A4-Seiten wurden analysiert, wobei eine Extraktionsrate von über 50 % ab dem Jahr 1955 erreicht wurde. Clusteranalysen identifizierten verschiedene Kugelhahn- und Ventil-Cluster. Die Toolchain bietet technische Vorteile wie Effizienz, Zeitersparnis und Skalierbarkeit, was die Verwaltung und Nutzung von Informationen in Unternehmen erleichtert. Sie stellt eine leistungsstarke Lösung dar, um Geschäftsprozesse zu vereinfachen und Zertifizierungsverfahren in der Gasinfrastruktur zu unterstützen.

Einleitung

Die Gasinfrastruktur spielt eine entscheidende Rolle bei der Umsetzung der Energiewende in Deutschland. Durch den Transport von Biomethan und Wasserstoff über die bestehende Infrastruktur kann eine nachhaltige und kohlenstofffreie Energieversorgung in der Zukunft ermöglicht werden. Um dieses Ziel zu erreichen, ist es erforderlich, die Eignung der betroffenen Infrastruktur zu prüfen. Während die grundsätzliche Eignung der Pipelines für den Wasserstofftransport bereits geklärt ist, konzentriert sich die Prüfung hauptsächlich auf den Nachweis der verwendeten Komponenten und Materialien. Die wesentlichen Informationen über die Komponenten im Hochdruckbereich liegen in den Materialprüfzeugnissen vor. Eine manuelle Überprüfung einer einzelnen Leitung ist immer noch möglich, erfordert jedoch entweder erheblichen personellen Aufwand oder eine technische Lösung, um eine umfassende Systemprüfung aller relevanten Dokumente durchzuführen. Der Einsatz von künstlicher Intelligenz bietet sich an, um die zeitaufwändige manuelle Auswertung dieser Dokumente zu vermeiden. Durch den Einsatz von KI kann der manuelle Aufwand auf ein Minimum reduziert und gleichzeitig die Fehleranfälligkeit bei der Auswertung minimiert werden.

In diesem Praxisbericht wird der Einsatz einer trainierten Sprachintelligenz, bereitgestellt von der Firma Planet AI und angeboten als Dienstleistung von der Firma Adesso, in Verbindung mit anderen Algorithmen zur Muster- und Spracherkennung beschrieben. Die Gesamtheit der angewendeten digitalen Schritte zur Prozessierung der Materialprüfzeugnisse wird „Toolchain“ genannt. Diese Toolchain führt in ihrer Gesamtheit von dem PDF-Dokument bis zu einem Datenpunkt, der im richtigen Format und an der richtigen Stelle in einer Datenbank sitzt. Aus den Materialprüfzeugnissen werden bis zu 31 relevante Informationen entnommen. Nach der Extraktion werden die ausgelesenen Daten anhand verschiedener Regeln bereinigt, um nur plausible Werte beizubehalten. Diese Bereinigung dient dazu, Fehler und Ungenauigkeiten zu minimieren. Anschließend werden die bereinigten Daten genutzt, um Cluster und Baugruppen zu identifizieren und zu definieren. Durch die

Identifikation und Definition von Clustern und Baugruppen ist es möglich, nicht nur einzelne Armaturen oder Komponenten zu bewerten, sondern eine umfangreiche Anzahl von Elementen mit identischen Werkstoffen und Eigenschaften direkt einzubeziehen. Diese Cluster und Baugruppen dienen zukünftig als Grundlage für die Zertifizierung der Materialien in Bezug auf ihre Eignung für den Einsatz mit Wasserstoff.

Durch die Verwendung dieser KI-Toolchain können die manuellen Arbeitsaufwände signifikant reduziert werden und die Effizienz und Genauigkeit bei der Auswertung der Materialprüfzeugnisse steigt. Der Einsatz dieser Technologie ist somit ein wichtiger Schritt in Richtung einer nachhaltigen und effektiven Nutzung von Wasserstoff im Gassektor.

Im folgenden Abschnitt wird detailliert auf die Methoden eingegangen, die bei der Entwicklung und Implementierung der Toolchain verwendet wurden. Darüber hinaus wird die Funktionsweise der Toolchain sowie ihre Anwendung in einem realen Szenario genauer betrachten. Anschließend werden die technischen Voraussetzungen und Vorteile der Toolchain erläutert. Zum Abschluss werden eine Zusammenfassung und Fazit gegeben.

Es wird angenommen, dass die vorliegende Toolchain einen erheblichen Mehrwert für Unternehmen in der Gasbranche darstellen kann. Daher wird der Leser in diesem Bericht dazu eingeladen, weitere Informationen über die Funktionen, Vorteile und Anwendungsmöglichkeiten dieser Toolchain zu erhalten.

KI-Toolchain

Die Herausforderung besteht darin, die Eignung des 12.000 km langen OGE-Leitungssystems für den Transport von Wasserstoff zu bewerten. Dabei müssen die Materialien, die in allen Komponenten der Gasinfrastruktur verwendet werden, berücksichtigt werden. Besonders Bauteilgruppen wie Armaturen mit unterschiedlichen Materialien müssen in die Bewertung einfließen. Eine zentrale Rolle spielen dabei die Armaturen wie Kugelhähne und Schieber, ergänzt um Details wie Dichtungen und Federn, sofern verfügbar. Es existieren im Bestand der OGE über 100.000 Dokumente (Material-

zeugnisse) mit insgesamt 1,6 Mio. DIN A4-Seiten, die Informationen über Materialien enthalten. Allerdings sind diese Informationen nur sporadisch und in unstrukturierter Form als digitale Dokumente verfügbar. Um dieser Herausforderung zu begegnen und eine systematische Bewertung des Assets sicherzustellen, wurde die KI-Toolchain entwickelt.

Der Schwerpunkt liegt darauf, die Daten aus den Materialzeugnissen in einer digital strukturierten Form zu erfassen. In **Tabelle 1** werden die relevanten Parameter präsentiert, die mithilfe der Toolchain extrahiert werden. Dabei steht die präzise Identifizierung des Bauteils sowie die Erfassung der relevanten mechanischen und chemischen Eigenschaften im Vordergrund.

Nach der Extraktion der Parameter ist das Hauptziel, identische Bauteilgruppen zu identifizieren, um eine umfassende Bewertung der Materialverwendung für ganze Baugruppen zu ermöglichen. Zusätzlich wird angestrebt, verschiedene Datenstrukturen wie Rohrbücher und andere Datensätze zu integrieren, indem sie mithilfe der KI-Ergebnisse verknüpft werden (Datenintegration). Darüber hinaus wird dieses Verfahren als Dienstleistung für andere Infrastrukturbetreiber angeboten, die ähnlichen Herausforderungen gegenüberstehen. Die Toolchain wurde dahingehend entwickelt, auf den Ergebnissen des OGE-Assets aufzubauen und eine hohe Erweiterbarkeit zu gewährleisten. Hierfür ist es notwendig, neue Klassen und die entsprechenden Key-Value-Paare zu definieren und einzufügen, die von den bestehenden Kategorien des OGE-Assets abweichen. Dies wird insbesondere relevant, wenn Zuliefererunternehmen Seitenlayouts verwenden, die nicht im OGE-Asset vorhanden sind. Durch diese Vorgehensweise kann die Toolchain kontinuierlich erweitert werden, um den sich wandelnden Anforderungen gerecht zu werden. Eine detailliertere Erläuterung dieses Prinzips erfolgt in den folgenden Abschnitten.

Aufbau und Funktionsweise der Toolchain

Der schematische Aufbau der Toolchain ist in **Bild 1** dargestellt. Das Verfahren besteht aus den sechs Schritten Preprocessing, der Planet AI-Umgebung, dem Output, Postprocessing sowie der Bereinigung der Daten (Cleansing) und der Analyse, die das Ziel verfolgt, Baugruppen zu finden. Diese Schritte werden im Folgenden genauer erläutert.

Tabelle 1: Extrahierte Eigenschaften aus den Materialprüfzeugnissen

Prüfobjektbezeichnung	Anzahl	Komponente
Hersteller	Seriennummer	Streckgrenze R_{eH}
Bestellnummer	Fabriknummer	Streckgrenze $R_{p0,2}$
Werkstoff	Werksnummer	Zugfestigkeit
Nennweite	Probe Nr.	Kerbschlagarbeit
Nenndruck	Abmessung	Härte
Druckklasse	Außendurchmesser	Kohlenstoff
Baujahr	Länge	Mangan
Typ	Wanddicke	Phosphor
Datum	Prüfdruck	Schwefel

1. Preprocessing

Der erste Schritt des Prozesses umfasst das Preprocessing, das das Einlesen der PDF-Dokumente in die Planet AI-Software und die Vorklassifizierung der Klassen beinhaltet. Die Seitenklassifikation ist von entscheidender Bedeutung für die anschließende Key-Value-Extraktion, da die Extraktionsmodelle speziell für bestimmte Klassen trainiert werden. Aufgrund der unbekanntenen Anzahl von Klassen im OGE-Dokumentenbestand müssen die Seitenklassen manuell definiert werden. Zur Bewältigung dieser Herausforderung wurde ein neuronales Netzwerk (Siamesisches Netzwerk) mithilfe vorgruppiertes Schuck-Dokumente trainiert, die zunächst von PDF-Form in Bilddateien umgewandelt wurden. Das neuronale Netzwerk erlernt Muster aus dieser Trainingsmenge, um die Dokumentenbilder optimal zu klassifizieren. Die Ergebnisse der Mustererkennung werden in hochdimensionale RGB-Informationen des Bildes umgewandelt und als Zahlenvektoren (Embeddings) repräsentiert. Dies ermöglicht den inhaltlichen Vergleich der Bilder und ihre Einordnung in Klassen. Im weiteren Verlauf ist eine nachträgliche Prüfung der Klassen und eine manuelle nach Sortierung erforderlich, um die Ergebnisse zu verwenden.

2. Planet AI-Umgebung

Der zweite Schritt umfasst die Verwendung der Planet AI-Software, die es ermöglicht, für jedes Dokument einen Textlayer und eine

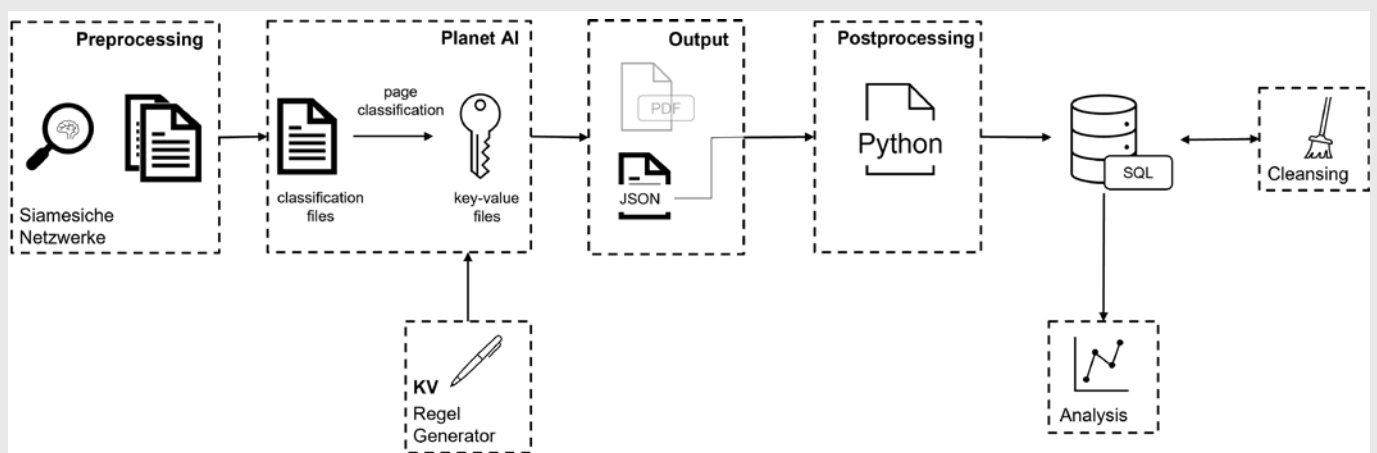


Bild 1: Schematische Darstellung der Toolchain

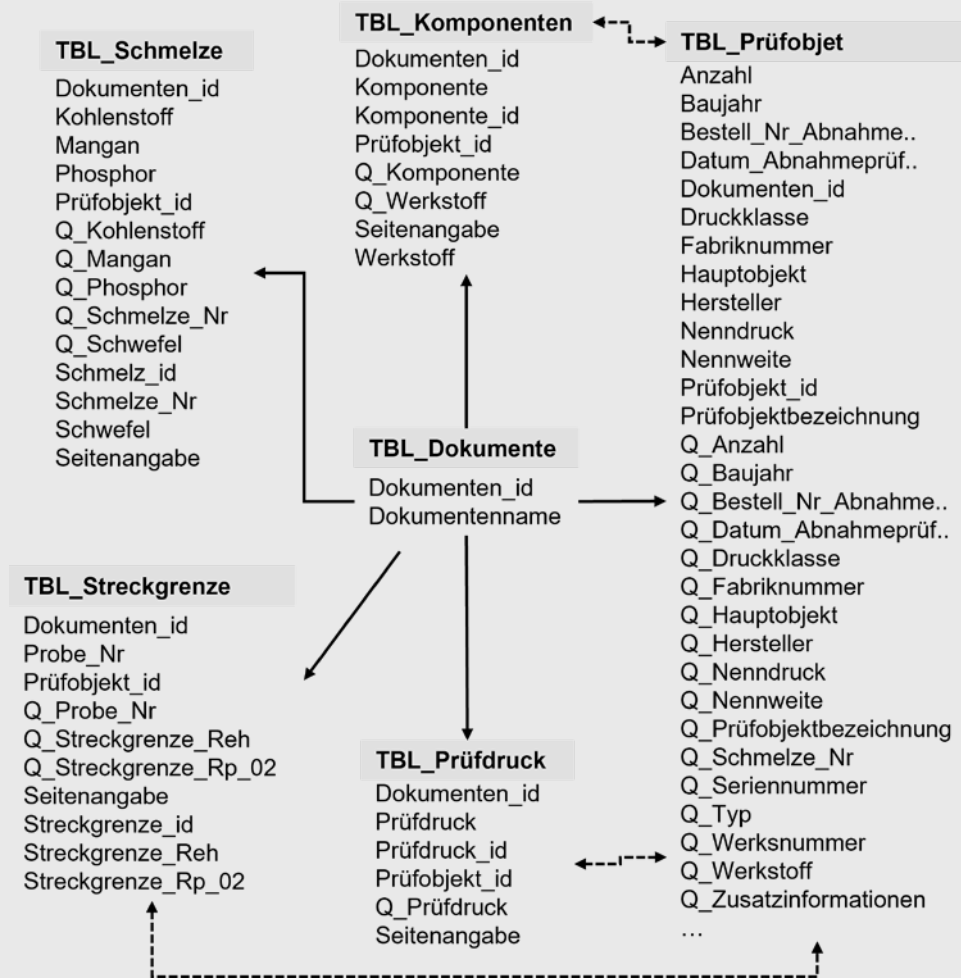


Bild 2: Schematische Darstellung der Datenstruktur der Ergebnisse und der zugrundeliegenden Beziehungen zwischen den einzelnen Tabellen. Die Pfeile verdeutlichen die Abhängigkeit von sowohl der Dokumenten-ID als auch der Prüfobjekt-ID

Perception-Matrix mit hohen Erkennungsraten für maschinengeschriebenen und handschriftlichen Text zu generieren. Dadurch wird der gesamte Dokumentenbestand durchsuchbar gemacht. Dieser Prozess ermöglicht die Implementierung der identifizierten Klassen mit den Konfigurationsdateien der Key-Value-Extraktion. Die Key-Value-Extraktion erfolgt anhand vordefinierter Regeln, bei denen sogenannte „Keys“ definiert werden, um den Abstand und den Bereich zu bestimmen, in dem der zugehörige „Value“ gefunden werden kann. Aufgrund der großen Anzahl an Dokumenten und der zu extrahierenden Eigenschaften wurde festgestellt, dass es nicht praktikabel ist, für jede mögliche Klasse eine spezifische Key-Value-Datei zu erstellen. Aus diesem Grund wurden zusätzlich zu den 110 speziellen Klassen mehrere allgemeine Key-Value-Dateien erstellt, die es ermöglichen, auch aus nicht klassifizierten Dokumenten Eigenschaften abzuleiten. Dieses Vorgehen kann jedoch Ungenauigkeiten verursachen, die in späteren Schritten bereinigt werden. Um diese Schritte umzusetzen, ist ein gewisser manueller Labeling-Aufwand erforderlich, der in einem Dashboard durchgeführt wird, um die proportionale Abhängigkeit zwischen dem Key und dem Value zu definieren. Anschließend werden diese Informationen in die entsprechenden Konfigurationsdateien umgewandelt und der Durchlauf kann durchgeführt werden.

3. Output und Postprocessing

Der dritte Schritt beinhaltet die Ausgabe der Ergebnisse als JSON-Datei und das Postprocessing mithilfe von Python-Skripten, um die Ergebnisse in einer strukturierten Datenbank zu speichern. Die Python-Skripte wurden als Python-Bibliothek entwickelt, so dass ihre Funktionen für verschiedene (Seiten-)Klassen importiert werden können. Die Hauptfunktion besteht darin, die JSON-Datei von Planet AI in ein Pandas-Datenframe umzuwandeln. Dies bietet den Vorteil, dass Anpassungen bei einem Software-Update von Planet AI leicht durchführbar sind, ohne die grundlegende Architektur ändern zu müssen. In einem Konfigurationsfile kann festgelegt werden, welche Funktionen für die jeweils erkannte Klasse ausgeführt werden sollen. Dabei stehen verschiedene Toolboxes zur Verfügung, die es ermöglichen, eine hohe Flexibilität sicherzustellen, um später ohne großen Aufwand neue Klassen festzulegen. Die Ergebnisse werden in einer (SQL-)Datenbank geschrieben und in einem nachgelagerten Schritt bearbeitet. Die festgelegte Datenstruktur ist in **Bild 2** dargestellt.

Alle Tabellen sind über einen eindeutigen Identifikator miteinander verknüpft (Dokumenten-ID oder Prüfobjekt-ID). Die Tabellen gliedern sich in die Bereiche: Dokumente (Metadaten), Prüfobjekt (Eigenschaften wie Hersteller, Nenndruck oder Baujahr), Kom-

ponenten (Bestandteile des Bauteils und dessen Eigenschaften), Schmelze (chemische Elemente), Prüfdruck (Prüfdruck) und Streckgrenzen (R_{eH} , $R_{p0.2}$). Der Q-Wert repräsentiert die Wahrscheinlichkeit, mit der das Schriftbild eines Buchstabens durch die Sprach-KI bewertet wird. Er liefert Erkenntnisse darüber, wie sicher die KI die tatsächliche Existenz eines spezifischen Buchstabens einschätzt.

4. Bereinigung der Daten

Nach dem Auslesen der Dokumente durch die KI liegen die erfassten Daten in unveränderter Form in der Datenbank vor. Abhängig von der Scanqualität der Dokumente und der Güte der KI-Filter sind viele Daten allerdings noch nicht nutzbar. Daher ist der KI-Erfassung ein Cleansing-Prozess nachgeschaltet, der die Daten bereinigen und vereinheitlichen soll. Mithilfe sogenannter „Dataflows“ und „Pipelines“ können in der Azure Data Factory die Daten aus ihrer Quelle abgerufen werden, um diese dann in hintereinander- oder parallelgeschalteten Prozessen zu bearbeiten, umzuwandeln oder zu entfernen. So können beispielsweise aus weniger eindeutigen Datenbankeinträgen überflüssige Zeichen entfernt oder interessante Zeichenfolgen extrahiert werden. Sofern der Q-Wert der Daten unterhalb von 85 % liegt, erfolgt eine Aussortierung der Daten. Außerdem wurden umfangreiche Dictionaries für mehrere Attribute (Hersteller, Nennweite, Werkstoff usw.) erarbeitet, mit denen die Daten abgeglichen und vereinheitlicht werden. Zusätzlich zu der Bereinigung der Daten werden in diesem Schritt Werkstoffgruppen gebildet, diese basieren auf der DIN CEN ISO 15608 und der chemischen und mechanischen Eigenschaften der Werkstoffe. Um zwei Beispiele zu nennen, Gruppe 1.2 sind Stähle mit Mindeststreckgrenze $275 \text{ N/mm}^2 < R_{eh} < 360 \text{ N/mm}^2$ und Gruppe 3.1 Vergütete Feinkornbaustähle mit Mindeststreckgrenze $360 \text{ N/mm}^2 < R_{eh} < 690 \text{ N/mm}^2$. Dieser Schritt ist von wesentlicher Bedeutung, um die Zusammenfassung verschiedener Werkstoffe zu ermöglichen, da sich gezeigt hat, dass insgesamt 514 unterschiedliche Werkstoffbezeichnungen identifiziert wurden. Die generierten Ergebnisse werden in eine

neue bereinigte Datenbank übertragen, auf der die anschließende Datenanalyse und Bauteilgruppenbildung basiert.

5. Definieren der Bauteilcluster

Der finale Schritt besteht aus der Durchführung der Datenanalyse, die das Ziel verfolgt, Gruppen von Bauteilen/Serien zu identifizieren und zusammenzufassen. Hierbei liegt zunächst der Fokus auf Kugelhähnen, wobei die Eigenschaften Hersteller, Komponente, Werkstoffe, Nenndruck, Nennweite und Baujahr berücksichtigt werden. Die mechanischen Eigenschaften werden indirekt durch ihre Zuordnung zu Werkstoffgruppen berücksichtigt. Mithilfe der Definition relevanter Eigenschaften und der Datenextraktion können Cluster-Algorithmen auf die Daten angewendet werden. Ein weiterer Aspekt der Bauteilserien besteht in der automatischen Korrektur bzw. Ergänzung von Einträgen, die von der KI-Extraktion nicht erkannt wurden. Das Vorgehen umfasst die folgenden drei Voraussetzungen:

- Das Zielcluster muss mindestens zehn Einträge aufweisen.
- Das Zielcluster unterscheidet sich maximal in einer Eigenschaft.
- Das Ursprungscluster besitzt mehr als nur eine Eigenschaft.

In der weiteren Analyse werden auch Stellventile und Schieber berücksichtigt, die anhand der Eigenschaften Hersteller, Werkstoff und Baujahr geclustert werden. Die Ergebnisse werden im Folgenden an den Ergebnissen des OGE-Assets detaillierter beschrieben und vorgestellt.

Ergebnisse und Auswertung der Toolchain

Die umfangreiche Datensammlung des OGE-Assets besteht aus über 110.000 Dokumenten mit etwa 1,6 Mio. Seiten, die sich über den Zeitraum von 1927 bis 2020 erstrecken. Um einen Überblick über die vorhandenen Daten zu gewinnen, zeigt **Bild 3** die Verteilung der digitalisierten Dokumente in Form von PDF-Dateien für die einzelnen Jahre. Diese Dokumente enthalten detaillierte

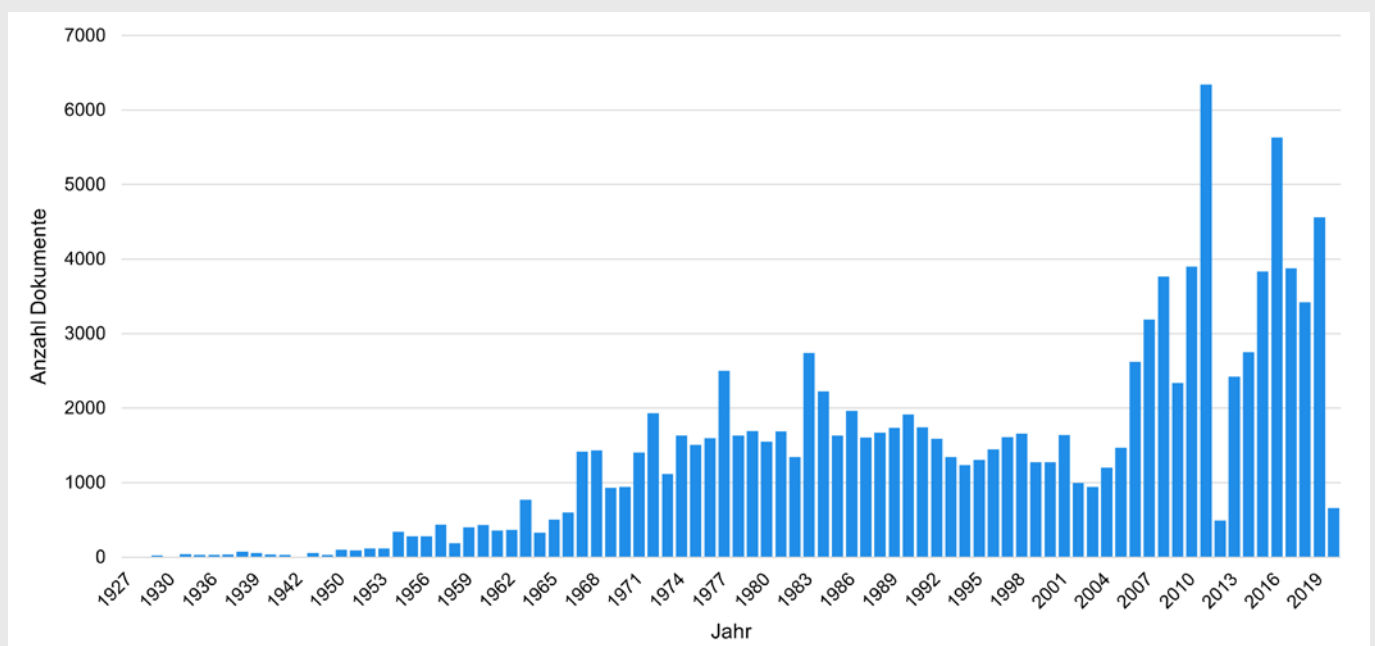


Bild 3: Überblick über die Dokumentenverteilung des OGE-Assets von 1927 bis 2020

Informationen über die verschiedenen Leitungsabschnitte des OGE-Netzwerks. Die Analyse zeigt, dass die Extraktion von Informationen bis zum Jahr 1955 weniger erfolgreich verlief, was auf eine Vielzahl von Faktoren zurückzuführen ist. Einer dieser Faktoren ist das fortgeschrittene Alter der Dokumente, die durch Verfärbungen und Alterserscheinungen gekennzeichnet sind, was wiederum zu einer beeinträchtigten Scanqualität führt. Darüber hinaus beeinflusst das Jahr des Scannens die Qualität, wobei ältere Scans vor 2010 im Vergleich zu aktuellen Scans von geringerer Qualität sind. Zusätzlich erschwert die Verwendung von handschriftlichen Informationen die Lesbarkeit sowohl für menschliche Leser als auch für automatisierte Verfahren. An dieser Stelle ist es wichtig zu betonen, dass die Handschrifterkennung von PlanetAI bemerkenswert effektiv ist, jedoch klar erkennbare Schrift voraussetzt. Die Qualität der Handschrift beeinflusst direkt die Fähigkeit des Systems, eine präzise Erkennung durchzuführen. Es ist daher entscheidend, dass die Schrift deutlich und gut lesbar ist, um optimale Ergebnisse zu erzielen.

Nach dem Jahr 1955 konnte eine Extraktionsrate von über 50 % aus den vorliegenden Dokumenten festgestellt werden. Dabei ist Extraktionsrate bezogen auf die Extraktion von mindestens drei extrahierten Parametern, wie Prüfobjektbezeichnung, Komponente und Streckgrenze. Diese Beobachtung verdeutlicht, dass nicht jedes einzelne Dokument die erforderlichen Eigenschaften aufweist, um für die Analyse herangezogen zu werden. Insbesondere Dokumente, die lediglich Bestellbestätigungen enthalten, wurden in dieser Untersuchung nicht berücksichtigt, da der Fokus auf Materialzeugnissen lag. Zudem ist wichtig zu beachten, dass

erst ab Mitte der 1960er Jahre stumpfnahgeschweißte Rohre für Wasserstoffumstellungen überhaupt relevant werden. Die hohe Extraktionsrate nach 1955 deutet darauf hin, dass in diesem Zeitraum eine verbesserte Dokumentation und eine größere Verfügbarkeit von relevanten Materialdatenzugnissen vorliegen. Im Weiteren werden wir uns auf die extrahierten Ergebnisse der Materialzeugnisse konzentrieren und deren Eigenschaften genauer vorstellen.

Ergebnisse der KI-Extraktion

Tabelle 2 gibt einen Überblick über die Anzahl der extrahierten Einträge für verschiedene Parameter. Die Anzahl der Einträge variiert je nach Prüfobjektbezeichnung, Werkstoff, Nennweite, Nenndruck, Komponenten, Prüfdruck, Streckgrenzen und Schmelzwerten. Diese Zahlen geben einen ersten Einblick in

Tabelle 2: Ergebnisse der Extraktion

Extrahierte Parameter	Anzahl Einträge
Prüfobjektbezeichnung	805.000
Werkstoff	332.000
Nennweite	700.000
Nenndruck	115.500
Komponenten	51.900
Prüfdruck	107.000
Streckgrenzen	2.475.260
Schmelzwerte	7.800.000

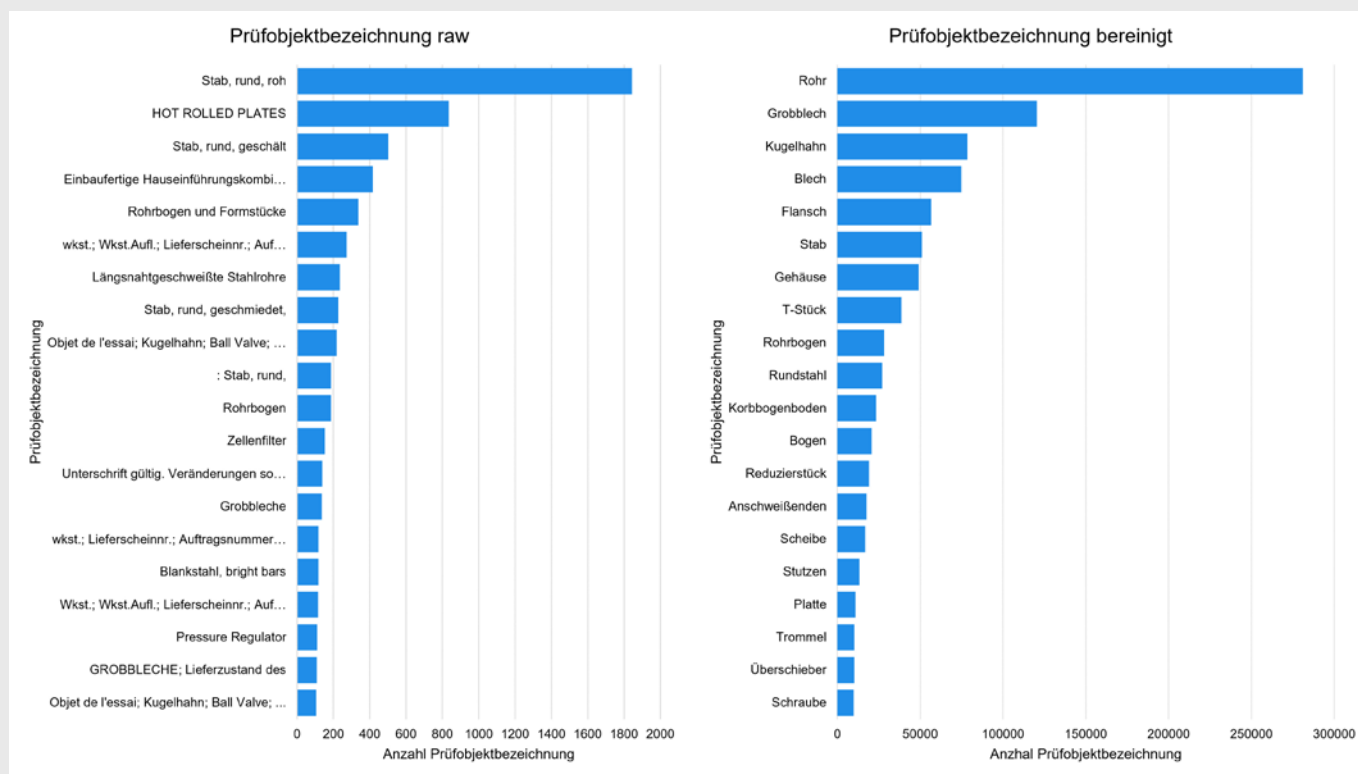


Bild 4: Darstellung der erhaltenen Ergebnisse für die Prüfobjektbezeichnung vor und nach der Bereinigung. Links sind die 20 am häufigsten extrahierten Daten direkt aus den Ergebnissen dargestellt, während rechts die 20 am häufigsten extrahierten Daten nach der Bereinigung (Cleansing) aufgeführt sind

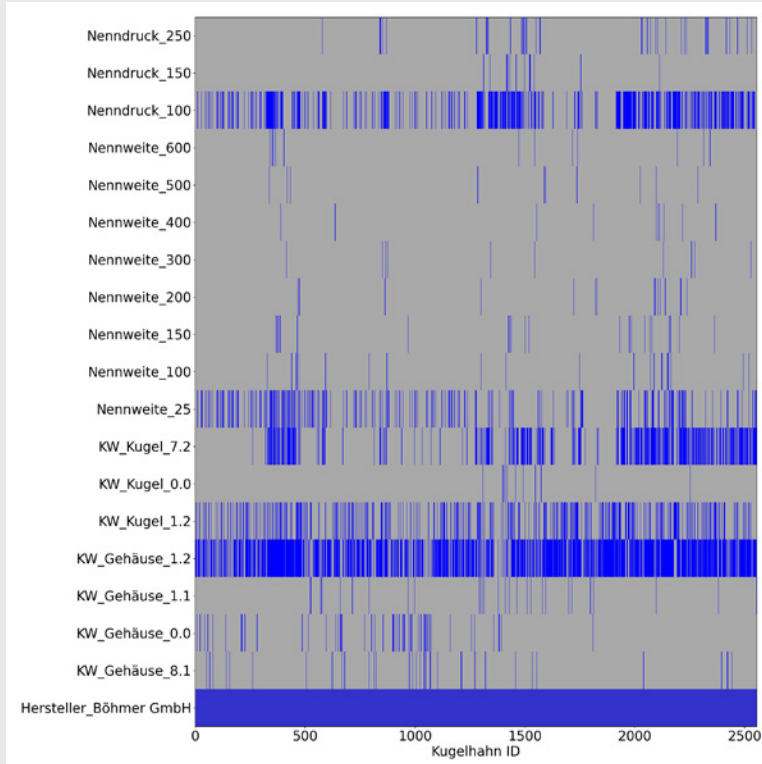


Bild 5: Ungeclusterte Datenbankeinträge (blau) am Beispiel Böhmer GmbH Kugelhähne, betrachtet werden Nenndrücke, Nennweiten, Kugel- und Gehäusewerkstoffgruppen

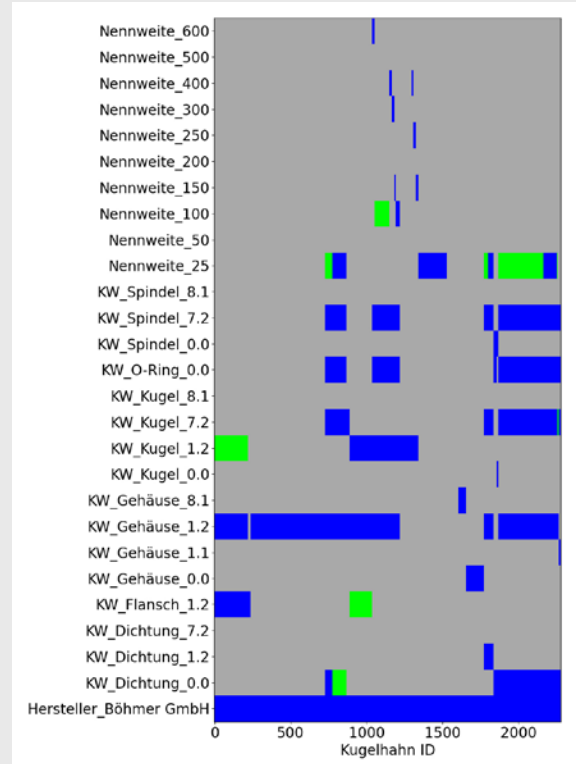


Bild 6: Geclusterte (blau) und ergänzte (grün) Daten vom Hersteller Böhmer GmbH, betrachtet werden Nenndrücke, Nennweiten, Kugel- und Gehäusewerkstoffgruppen

die Leistungsfähigkeit der KI-Toolchain. So konnten Prüfobjektbezeichnungen 805.000-mal extrahiert werden, was der Hälfte der Seitenanzahl entspricht.

Auffällig ist die signifikante Anzahl von Streckgrenzen und Schmelzwerten, die die Gesamtseitenzahl der extrahierten Dokumente übersteigt. Dieses Phänomen ist auf die Tatsache zurückzuführen, dass viele Informationen in tabellarischer Form vorliegen und mehrere Werte für ein Prüfobjekt extrahiert werden. Zusätzlich sind die Schmelzwerte für Kohlenstoff, Mangan, Phosphor und Schwefel zusammengefasst dargestellt.

In **Bild 4** werden die extrahierten Ergebnisse im Vergleich zu den bereinigten Daten dargestellt. Die Bereinigung erfolgt durch die gezielte Suche nach bestimmten Wörtern bzw. Bezeichnungen in der unbereinigten Tabelle und ordnet es zu einer vereinheitlichten Prüfobjektbezeichnung gemäß einer vordefinierten Zuordnungstabelle (Dictionary) zu. Dieser Ansatz wird auch angewendet, um beispielsweise zusätzliche Informationen aus der Spalte „Zusatzinformationen“ zu extrahieren und die Spalte „Prüfobjektbezeichnung“ weiter zu ergänzen.

Baugruppen-Auswertung

In der vorliegenden Auswertung liegt der Fokus auf der Analyse von Kugelhähnen, da diese im gesamten Projekt eine hohe Priorität haben. Darüber hinaus erfolgt eine Beschreibung der Clusterergebnisse für die Ventile.

1. Kugelhahn-Cluster

Die Ergebnisse der Clusterbildung von Kugelhähne umfassen die folgenden Informationen:

- » Es wurden insgesamt 16 verschiedene Hersteller im OGE-Asset identifiziert, die Kugelhähne für die Gasinfrastruktur liefern.
- » Innerhalb dieser Hersteller wurden insgesamt 145 Cluster von Kugelhähnen identifiziert. Ein Cluster repräsentiert eine Gruppe von Kugelhähnen mit ähnlichen Werkstoffen und Eigenschaften.
- » Dem Hersteller „Deutsche Audco GmbH“ werden 45 Kugelhahn-Cluster zugeteilt, was auf eine bedeutende Präsenz und Lieferkapazität in der Gasinfrastruktur hinweist.
- » Der Hersteller „Böhmer GmbH“ wird in 25 Kugelhahn-Clustern eingeteilt.
- » „Schuck Group GmbH“ ist ein weiterer relevanter Hersteller, der in 18 Kugelhahn-Cluster eingeteilt wird.
- » Der Hersteller „RMA Rheinau GmbH“ wird in sechs Kugelhahn-Cluster eingeteilt, was auf eine begrenzte, aber dennoch vorhandene Präsenz hinweist.

Diese Ergebnisse verdeutlichen die Vielfalt der Hersteller und die Anzahl der verfügbaren Kugelhahn-Cluster, die im OGE-Asset eingesetzt werden.

Bild 5 zeigt die Ergebnisse in ungeclusterter Form für den Hersteller Böhmer GmbH, die nach der Bereinigung vorliegen. Jeder blaue Strich steht dabei für einen erkannten Eintrag in der angegebenen Zeile. Es werden die Komponenten, Werk-

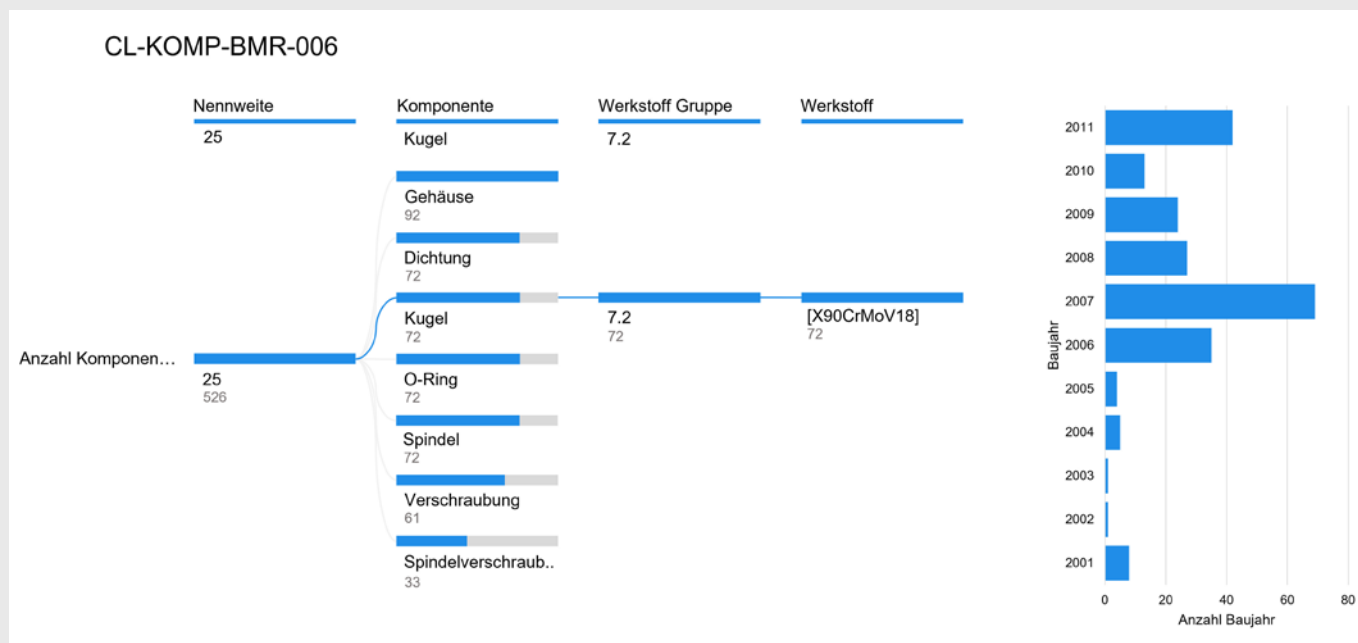


Bild 7: Darstellung eines Beispielclusters der Firma Böhmer GmbH für Kugelhähne, unter Betrachtung des Werkstoffs der Kugel aus Cluster 6 sowie der jährlichen Verteilung des Gesamtclusters

stoffgruppen sowie Nennweite und Nenndruck betrachtet. Diese Ergebnisse verdeutlichen die Schwierigkeiten bei der Übersichtlichkeit der vorliegenden Daten und betonen die Wichtigkeit einer Clusteranalyse zur Strukturierung und Kategorisierung.

Im Vergleich zu **Bild 5** präsentiert **Bild 6** die geclusterte Darstellung der Ergebnisse. Die grünen Markierungen stellen die automatische Korrektur der Daten dar, die gemäß den oben beschriebenen Voraussetzungen angewendet wurden.

Die Clusterung der Ergebnisse erweist sich als äußerst vorteilhaft und bietet eine Vielzahl von Nutzen. Erstens ermöglicht sie eine gemeinsame Auswertung und bietet Projektleitern und Gutachtern einen besseren Überblick über die Daten. Durch die Strukturierung und Kategorisierung der Ergebnisse wird es einfacher, Zusammenhänge und Muster zu erkennen. Zweitens liefert die Clusteranalyse Informationen darüber, welcher Hersteller zu welchem Zeitpunkt bestimmte Ventiltypen geliefert hat und ob es Änderungen in Bezug auf die Bauart oder die Verwendung von Materialien gab. Dies ist besonders relevant für die zeitliche Nachverfolgung von Entwicklungen und ermöglicht eine genauere Analyse der historischen Daten.

Des Weiteren unterstützt die Clusterung bei der Ersatzbeschaffung fehlender Dokumentation. Durch die Zusammenfassung und Kategorisierung von ähnlichen Bauteilen können Informationen über fehlende Dokumente aus anderen Clustergruppen extrapoliert und ergänzt werden. Dies hilft, Lücken in den vorhandenen Daten zu schließen und eine umfassendere Datengrundlage zu schaffen.

Schließlich ermöglicht die Clusteranalyse die Bestimmung beschreibender Eigenschaften für Baugruppen. Durch die Identifizierung von gemeinsamen Merkmalen und spezifischen Eigenschaften innerhalb der Clustergruppen können

präzise Beschreibungen und Charakterisierungen erstellt werden. Dies trägt zur Verbesserung der Dokumentation und der Erstellung detaillierter Berichte bei.

Zur detaillierten Veranschaulichung des Clusters wird in **Bild 7** ein eingehender Einblick präsentiert. Konkret handelt es sich um das Cluster 006 der Firma Böhmer GmbH für Kugelhähne mit einer Nennweite von DN 25. Die einzelnen Komponenten sowie die zugehörige Werkstoffgruppe und Werkstoffe werden dargestellt. In diesem Fall erfolgt eine nähere Betrachtung der Kugel, die der Werkstoffgruppe 7.2 angehört. Diese Werkstoffgruppe umfasst nichtrostende Stähle mit einem bestimmten Anteil an Chrom, Nickel oder Molybdän. Es ist anzumerken, dass die Kugel einen einzigen Werkstoff aufweist, obwohl in der Werkstoffgruppe auch verschiedene Werkstoffe zulässig sind. Des Weiteren wird das Baujahr der bestellten Kugelhähne aufgezeigt, was darauf schließen lässt, dass dieser Kugelhahn von 2001 bis 2011 gefertigt wurde.

2. Ventil-Cluster

Die Ventil-Cluster umfassen 12 Haupthersteller, denen 47 Cluster zugeordnet werden. In **Bild 8** sind die verschiedenen Ventiltypen und die zugehörigen Werkstoffgruppen dargestellt. Die vorherrschende Werkstoffgruppe bei den verbauten Axialsitzventilen ist 1.1, die unlegierte Stähle mit einer Mindeststreckgrenze von 185 N/mm² beschreibt. Diese Werkstoffgruppe ist ebenfalls die dominierende bei Rückschlagventilen, Gradsitzventilen, Sicherheitsabsperrenten (SAV) und Druckreglern. Bei Sicherheitsabblaseventilen (SBV) stellt die Werkstoffgruppe 7.2, die sich auf nichtrostende Stähle bezieht, den größten Anteil dar.

An dieser Stelle wird deutlich, dass die Gruppierung von Ventilen einen erheblichen Vorteil bietet, um eine umfassende Analyse der Eignung für den Wasserstoffbetrieb zu ermöglichen.

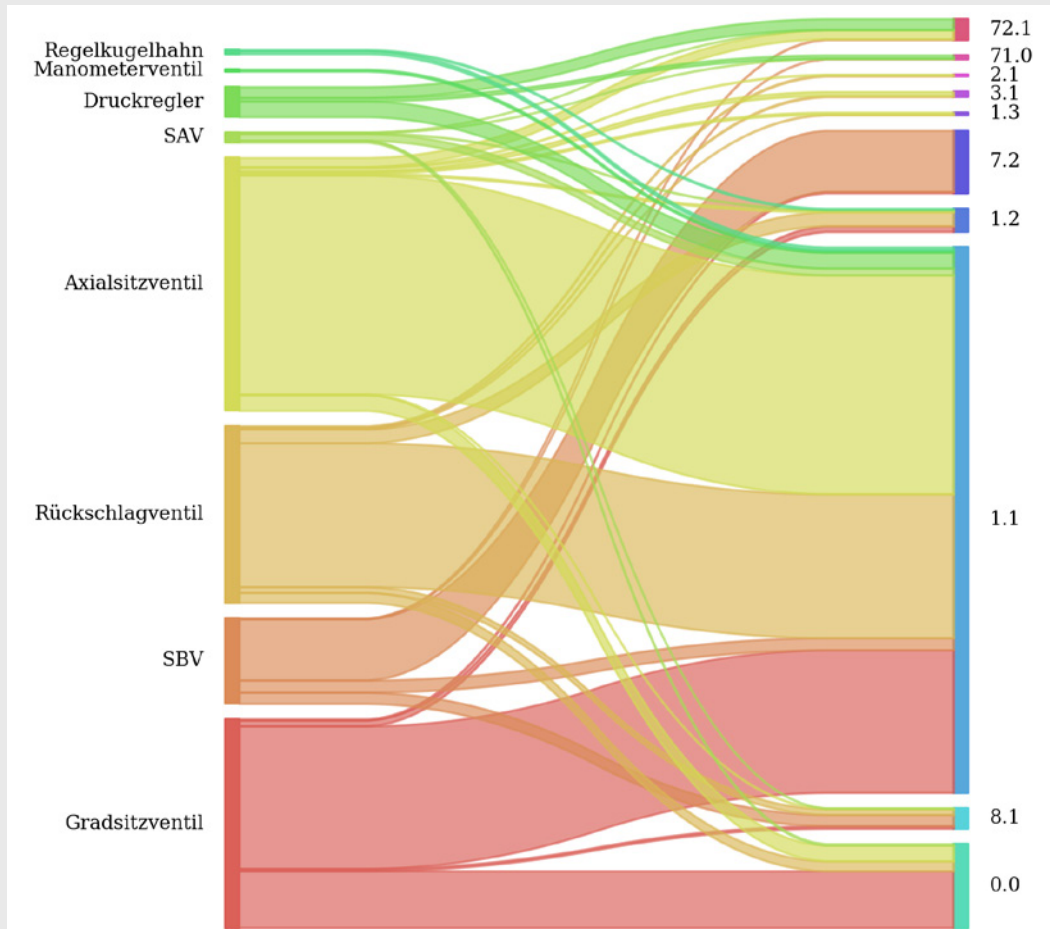


Bild 8: Geclusterte Ventile nach Werkstoffgruppen

Dadurch ist es nicht erforderlich, alle Bauteile individuell zu überprüfen, sondern es können repräsentative Komponenten aus einer bestimmten Werkstoffgruppe geprüft werden.

Technische Voraussetzung und Vorzüge der Toolchain

Die technischen Anforderungen für Infrastrukturbetreiber sind minimal und beinhalten lediglich die Bereitstellung der zu analysierenden Dokumente im PDF-Format. Dabei spielt die Qualität der Scans und PDFs eine entscheidende Rolle für die Datenqualität. Eine umfangreiche Bereitstellung von Metadaten und/oder eine gut strukturierte Organisation der Dokumente erleichtern die Klassifizierung. Darüber hinaus ist die Verwendung einer (SQL-) Datenbank erforderlich, insbesondere wenn die Datenmengen die Kapazität herkömmlicher Tabellenkalkulationsprogramme von etwa 1 Mio. Zeilen überschreiten.

Der primäre Nutzen liegt darin, dass die verarbeiteten PDF-Dokumente in einer strukturierten Form vorliegen, wobei die korrekte Zuordnung an den entsprechenden Positionen gewährleistet ist. Dies führt zu einer Steigerung der Effizienz und Zeitersparnis im Unternehmen und ermöglicht die weitere Identifizierung von Baugruppen zur Zertifizierung ihrer Wasserstofftauglichkeit. Die Toolchain bietet zudem weitere Vorteile. Ein bedeutender Aspekt ist ihre Flexibilität, da

sie an die spezifischen Anforderungen und Bedürfnisse von Unternehmen angepasst werden kann. Dies ermöglicht die Modifikation und Integration der Funktionen entsprechend den spezifischen Anwendungsfällen. Ein weiterer Aspekt ist die Skalierbarkeit, da sie problemlos in Unternehmen unterschiedlicher Größenordnungen implementiert werden kann und in der Lage ist, mit wachsenden Anforderungen und Datenmengen Schritt zu halten. Dadurch wird sichergestellt, dass die Toolchain langfristig den Bedürfnissen der Unternehmen gerecht wird. Die Ergebnisse bieten zudem umfangreiche Möglichkeiten für das Datenmanagement. Sie ermöglicht eine strukturierte Suche und Analyse von Daten, was zu einer effektiven Verwaltung und Nutzung von Informationen führt. Unternehmen können ihre Datenbestände optimal nutzen, wertvolle Erkenntnisse gewinnen und die Integration verschiedener Datenstrukturen vorantreiben. Insgesamt stellt die Toolchain anderen Unternehmen eine leistungsstarke Lösung mit vielfältigen Vorteilen bereit. Sie ermöglicht eine effiziente und flexible Nutzung der Tools zur Vereinfachung von Geschäftsprozessen und zur Unterstützung von Zertifizierungsverfahren. Dadurch bietet die Toolchain eine zuverlässige und fortschrittliche Lösung, um den Anforderungen von Unternehmen gerecht zu werden und ihren Geschäftsbetrieb zu optimieren.

Zusammenfassung und Fazit

Die Toolchain zur Extraktion und Analyse von Materialdatenzugnissen im Rahmen des OGE-Assets ist ein effektives Verfahren, das auf maschinellem Lernen und KI-Algorithmen basiert. Die Toolchain besteht aus mehreren Schritten, darunter Preprocessing, Planet AI-Umgebung, Output und Postprocessing, Datenbereinigung und Analyse. Jeder Schritt erfüllt eine spezifische Funktion, um die extrahierten Informationen zu optimieren und nutzbar zu machen.

Im Preprocessing-Schritt werden die PDF-Dokumente eingelesen und in Klassen vorklassifiziert. Dies geschieht mithilfe eines neuronalen Netzwerks, das auf Bilddaten trainiert wurde. Durch die Nutzung der Planet AI-Umgebung kann eine effektive Generierung von Textlayern und Perception-Matrices erreicht werden, wodurch eine zuverlässige Erkennung von maschinen- und handschriftlichem Text ermöglicht wird. Die Extraktion von Key-Value-Paaren erfolgt mithilfe festgelegter Regeln, die sowohl spezifische als auch allgemeine Key-Value-Dateien einbeziehen. Die erzielten Resultate werden in einer Datenbank strukturiert abgespeichert. Nachdem die KI die Extraktion durchgeführt hat, werden die Daten einer Bereinigung und Standardisierung unterzogen. Dies umfasst das Entfernen überflüssiger Zeichen, die Extraktion interessanter Zeichenfolgen und den Abgleich der Daten mit umfangreichen Dictionaries. Zusätzlich werden Werkstoffgruppen gebildet, um die Zusammenfassung verschiedener Werkstoffe zu ermöglichen. Der Cleansing-Prozess erfolgt mithilfe von Dataflows und Pipelines in der Azure Data Factory. Der finale Schritt besteht in der Durchführung der Datenanalyse, um Bauteilcluster zu identifizieren und zusammenzufassen. Dabei werden relevante Eigenschaften berücksichtigt und Cluster-Algorithmen angewendet. Die Ergebnisse ermöglichen die Identifizierung von Bauteilen und Serien sowie die automatische Korrektur oder Ergänzung von Einträgen.

Die Analyse der Materialzeugnisse für die Wasserstoffinfrastruktur hat wichtige Erkenntnisse geliefert. Die Extraktionsrate der relevanten Informationen aus den über 110.000 analysierten Dokumenten war nach dem Jahr 1955 über 50 %. Mithilfe einer Clusteranalyse konnten die Eigenschaften von Ventilen und Kugelhähnen eindeutig sortiert, geordnet und eindeutig beschrieben werden.

Dieses Ergebnis verdeutlicht die Vielfalt der Komponenten in Bestandssystemen und die breite Verwendung von Kugelhahn- und Ventil-Clustern im OGE-Asset. Die Anwendung der Toolchain bietet technische Vorteile wie Effizienz, Zeitersparnis und Skalierbarkeit innerhalb von H₂-Umstellungsprojekten. Durch eine strukturierte Suche und Analyse von Daten ermöglicht sie eine effektive Verwaltung und Nutzung von Informationen für den Gasnetzbetreiber.

Zusammenfassend lässt sich feststellen, dass die angewendete Toolchain eine wirkungsvolle Methode zur Extraktion und Ana-

lyse von Materialdatenzugnissen im Kontext des OGE-Assets darstellt. Sie erweist sich als effizientes Verfahren zur Verarbeitung umfangreicher Datensätze und erzielt präzise Resultate. Die Integration von maschinellem Lernen, KI-Algorithmen und Datenanalyse ermöglicht eine ganzheitliche Auswertung der Materialdaten, was zu einer verbesserten Dokumentation und erhöhten Verfügbarkeit relevanter Informationen führt. Darüber hinaus wird erwartet, dass diese Toolchain in Zukunft eine entscheidende Rolle bei der Bewertung der Wasserstofftauglichkeit von Leitungsnetzen spielen kann.

SCHLAGWÖRTER: Materialdaten, Analyse, Wasserstoff, Künstliche Intelligenz, Wasserstofftauglichkeitsbewertung, Zertifizierung

AUTOREN



M.Sc. **DEVIN SCHUMACHER**
PSvdL Engineering GmbH, Essen
devin.schumacher@psvdl.com



Dr. **ANDREAS SCHWÖRER**
Adesso SE, Köln
andreas.schworer@adesso.de



M.Sc. **SEBASTIAN WOLTERS**
PLEDoc GmbH, Essen
sebastian.wolters@pledoc.de



Dr. **DANIEL S. BICK**
Open Grid Europe GmbH, Essen
daniel.bick@oge.net



Hier geht's lang!

3R INFO

Der Newsletter für Rohrleitungssysteme

