Erinnern für alle zugänglich machen

Machine Learning bringt Licht in 110 Millionen Dokumente von KZ-Häftlingen und entschlüsselt damit Schicksale.

Für Arolsen Archives

TWT macht 110 Millionen Objekte der Arolsen Archives online lesbar. Machine Learning und künstliche Intelligenz mehrerer Google-Services helfen, eine maximale Transparenz über die Schicksale von 17,5 Millionen Betroffenen des Nazi-Regimes herzustellen.

Erste Teilprojekte des großen Indexierungsprojektes liefen bereits erfolgreich, obgleich die Herausforderung sehr groß ist: Der Datenbestand der Arolsen Archives besteht aus sehr unterschiedlichen Dokumententypen in sehr verschiedenen Qualitäten. Die Dokumente müssen zunächst sortiert und dann maschinell gelesen werden. Ein Großprojekt, das alle Teilnehmer fordert.

Eine Aufgabe mit unvorstellbar hoher Komplexität

Die Arolsen Archives, bis 2019 als „Internationaler Suchdienst“ bekannt, wollen bis 2025 ihren gesamten Datenbestand über die Opfer des Nationalsozialismus in einer Online-Anwendung für jeden recherchierbar machen. Alle Stationen auf dem Leidensweg der Betroffenen sollen von Interessierten im Internet einsehbar sein.

„Die Schicksalswege sollen möglichst vollständig nachvollzogen werden können“, sagt Michael Hoffmann, Referatsleiter für Projekt- & Qualitätsmanagement bei Arolsen Archives und Leiter des Großprojekts.

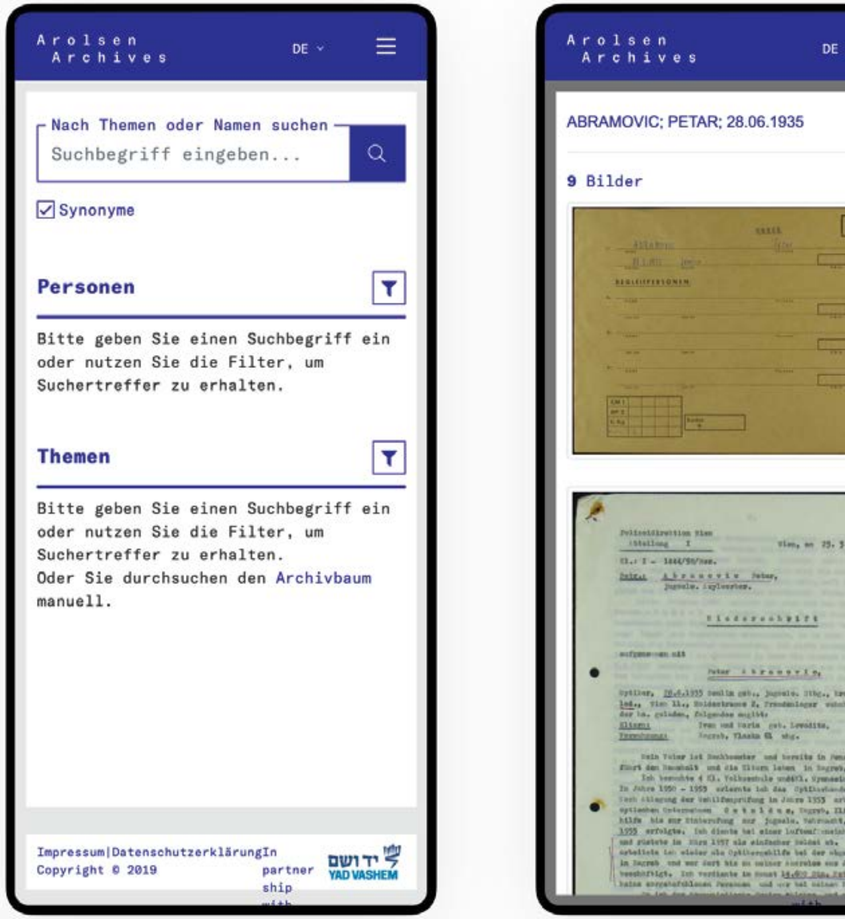
Die Voraussetzungen sind nur teilweise gut: Zwar liegen den Arolsen Archives 90 Prozent aller Dokumente in digitaler Form in einer Datenbank vor, aber die Recherche ist bislang nur den Mitarbeitern der Arolsen Archives möglich. Nur sie wissen, in welchen Dokumentengruppen sie suchen müssen.

„Ohne Vorkenntnisse der Datenstrukturen bei den Arolsen Archives ist eine umfassende Recherche aktuell nicht möglich“, berichtet Michael Hoffmann.

Hinzu kommt: Die insgesamt 110 Millionen Objekte in den Arolsen Archives sind auf viele verschiedene Dokumententypen verteilt. Die Bandbreite reicht von Lagerkarten über Häftlingspersonalakten bis hin zu Krankenblättern.

Die Darstellungsqualität in den vielen verschiedenen Dokumenten geht von lesbar bis hin zu vergilbt, verbrannt oder es fehlen Buchstaben, weil die Schreibmaschine kaputt war. Obendrauf kommen dann noch zahlreiche andere Faktoren, etwa die unterschiedliche Schreibweise von Namen.

So zählen die Arolsen Archives beispielsweise rund 800 verschiedene Versionen des Nachnamens „Abramovich“, berichtet Michael Hoffmann. „Die Komplexität unserer Indexierungs-Aufgabe ist unvorstellbar hoch.“

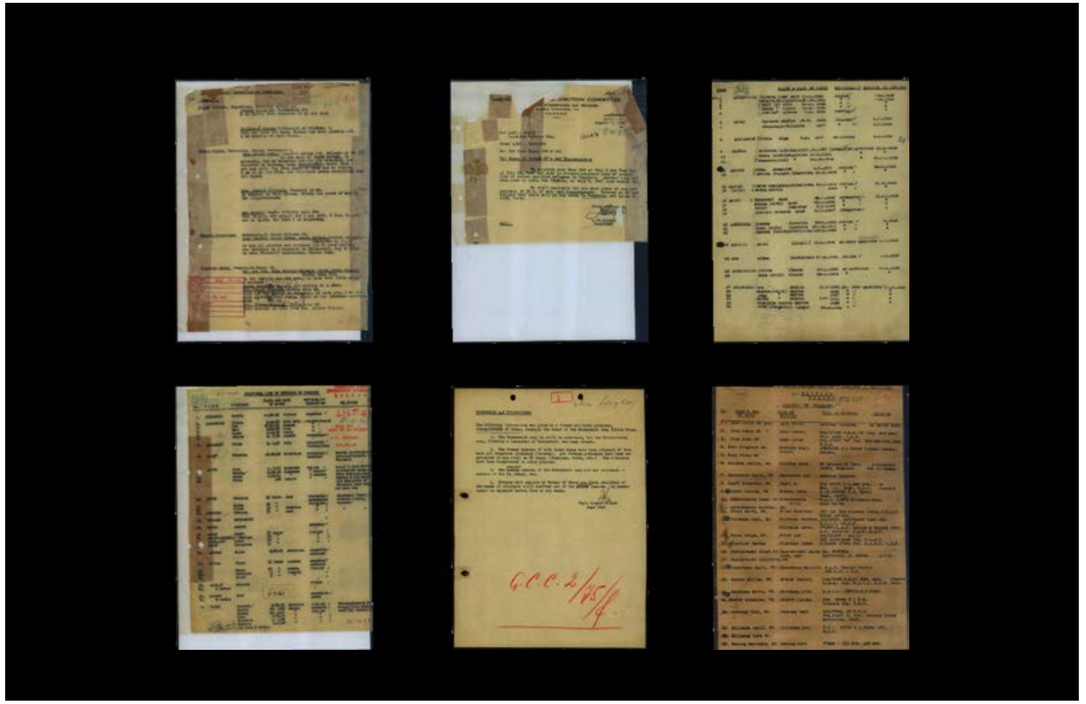


Für dieses ehrgeizige Indexierungs- und Digitalisierungsprojekt hielt Michael Hoffmann nur den größten Technologiekonzern der Welt für den geeigneten Partner. Also rief er 2018 kurzerhand bei Google Europa in Irland an. Dort nannte man ihm eine Reihe von Dienstleistern.

TWT, einziger deutscher zertifizierter Google Cloud Premier Partner für Machine Learning, überzeugte die Arolsen Archives.

Besonders bei Projektleiter Sebastian Butz hatte Michael Hoffmann schnell den Eindruck, „dass er nicht nur Engagement, sondern auch persönliches Interesse“ mitbringt.

Arolsen Archives und TWT waren sich einig, dass es in diesem auf mehrere Jahre angelegten Projekt Sinn macht, iterativ vorzugehen: Jedes Folgeprojekt soll aus dem Vorgängerprojekt lernen und darauf aufbauen. In einem ersten Schritt wurden die unterschiedlichen Dokumententypen von Karteikarten erfasst. „Wir wollten erstmal wissen, wie viele verschiedene Dokumententypen es unter unseren 3,5 Millionen Karteikarten gibt“, berichtet Hoffmann.



„Man bekommt schnell eine Idee dafür, ob ein Partner nur Interesse an einem weiteren Kunden hat, oder ob er wirklich mit Herzblut an diese Aufgabe herangehen wird, mitdenkt und aktiv das Projekt voranbringt.“

Michael Hoffmann zu der Entscheidung,
die schwierige Aufgabe der TWT anzuvertrauen.

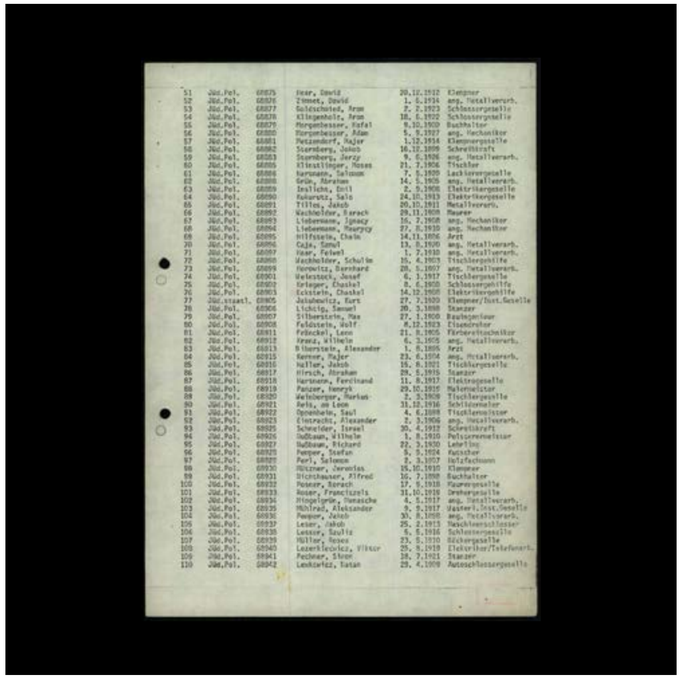
TWT ging dabei folgendermaßen vor: Zunächst wurden manuell rund 30 Dokumententypen erfasst, sie dienten der künstlichen Intelligenz des Google Machine Learning-Service als Trainingsmaterial. Zum Einsatz kam hier der AutoML Vision-Dienst von Google. Auf diese Art trainiert, gelang es TWT und Google, rund 90 Prozent der 3,5 Millionen Dokumente den entsprechenden Typen zu zuordnen. Die restlichen zehn Prozent übernahmen Mitarbeiter und Mitarbeiterinnen von Arolsen Archives sozusagen mit menschlicher Intelligenz und mit Hilfe eines Web-Portals, das ihnen TWT eigens programmiert und zur Verfügung gestellt hat.

Im zweiten Projekt ging es zunächst darum, zum ersten Mal konkrete Namen aus Dokumenten auszulesen: 800 000 Listen – wieder aus allen möglichen Quellen aus dem Zeitraum von 1933 bis 1955 und wieder in sehr vielen Qualitätsvarianten.

TWT clusterte alle Dokumente in Googles AI Platform, um besonders gut geeignete Objekte zu gruppieren.

Nach der Vorverarbeitung der Dokumente erfolgte die OCR-Texterkennung mit Hilfe der Google Vision API. Dann wurden die Datensätze gelabelt und für das Training ein AutoML-Natural-Language-Modell zur Erkennung aller Entitäten (z.B. Nachnamen, Vornamen, Geburtsdatum) jedes Datensatzes geschaffen. Die Daten wurden abschließend validiert und von TWT in das Archivsystem der Arolsen Archives übertragen.

Der Vorteil dieses Vorgehens:
Trotz der stark unteschiedlichen Qualität der Quellen konnten ein Großteil der maschinengeschriebenen Listen und teilweise auch handgeschriebene Listen, erkannt und die Personendaten ausgelesen werden.



„In Handarbeit hätte das Jahre gedauert.“

TWT-Projektleiter Sebastian Butz

Die Verantwortlichen bei Arolsen Archives sind mit dem Vorgehen der KI-Experten der TWT und den bisherigen Resultaten sehr zufrieden. Qualität und Zeit wurden jederzeit eingehalten und auch die Kosten „sind nachvollziehbar und werden sauber erklärt“.

„Wir haben bereits sehr, sehr gute Ergebnisse“, sagt Michael Hoffmann. Es hat sich gezeigt, dass der eingeschlagene Weg genau der richtige ist, indem jedem OCR-Vorgang zunächst ein Clustering der Datentypen vorgeschaltet wird. Das bringe die Qualität der OCR-Prozesse „deutlich nach oben“.

Michael Hoffmann hofft nun, dass man dank der von TWT entwickelten und erprobten Methoden auch bei den weiteren Projekten gut vorankommt und die Arolsen Archives eventuell auch schon vor 2025 verkünden können, dass die Schicksale der Opfer des Nationalsozialismus nun von jedem online nachvollzogen werden können.

Besuchen Sie die Website der Arolsen Archives:
https://arolsen-archives.org/

Kontakt

Mehr zu unseren Kernbereichen Enterprise Search, Geo, Machine Learning, Google Apps und Google Cloud Plattform erfahren Sie auf unserer Website.

Oder schreiben Sie uns an

business-solutions-sales@twt.de