



Erinnern für alle zugänglich machen

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

Für Arolsen Archives

466

COMMANDEMENT EN CHEF FRANÇAIS EN ALLEMAGNE

Carte d'identité de Personnes Déplacées

Nom LIMBERG

Prénoms ELISE - HELENE

Nationalité ESTHONIENNE

Lieu de naissance BELSEN

Lieu de détention TUBINGEN

Date de naissance 1882

Nom de famille LINN

Section des Personnes Déplacées

Date d'émission: 291046

Signature du chef de section des Personnes Déplacées: [Signature]

Signature de la titulaire: B. Limberg

Taille PETITE Nez DROIT

Visage OVALE Yeux BLEUS

Teint CLAIR Cheveux BLANCHES GRIS

Empreintes digitales

Pouce gauche	Pouce droit
[Empreinte]	[Empreinte]

Section des Personnes Déplacées

Section des Personnes Déplacées

Section des Personnes Déplacées

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

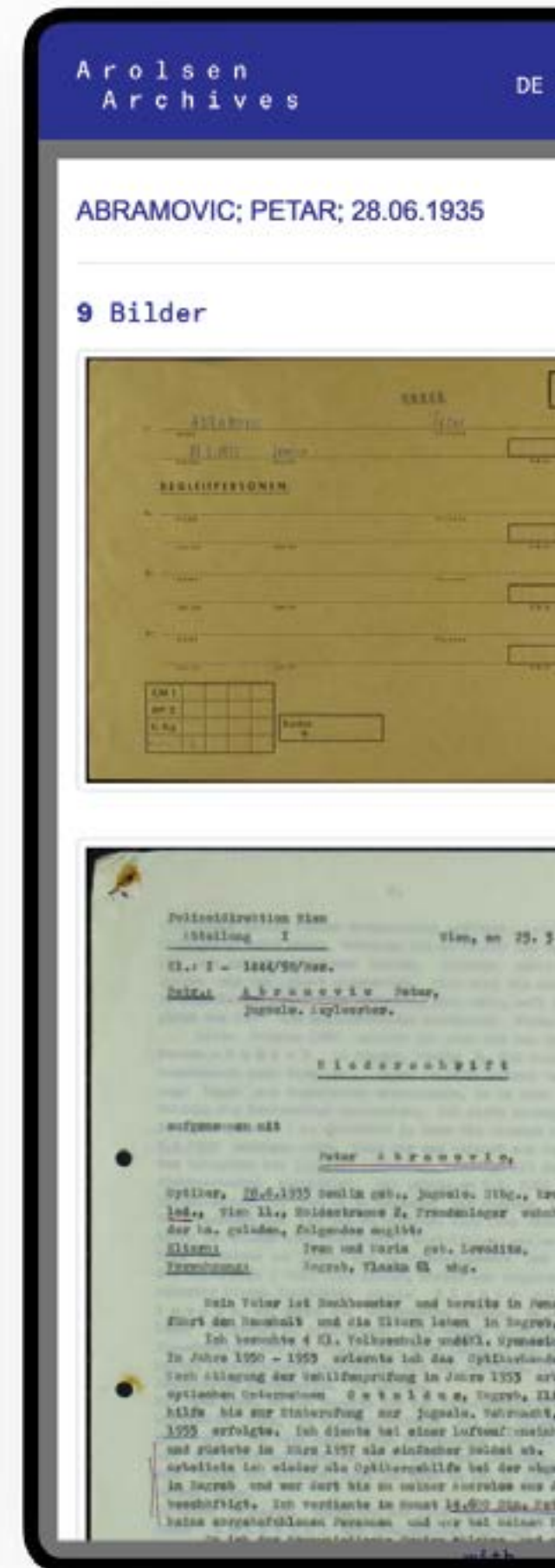
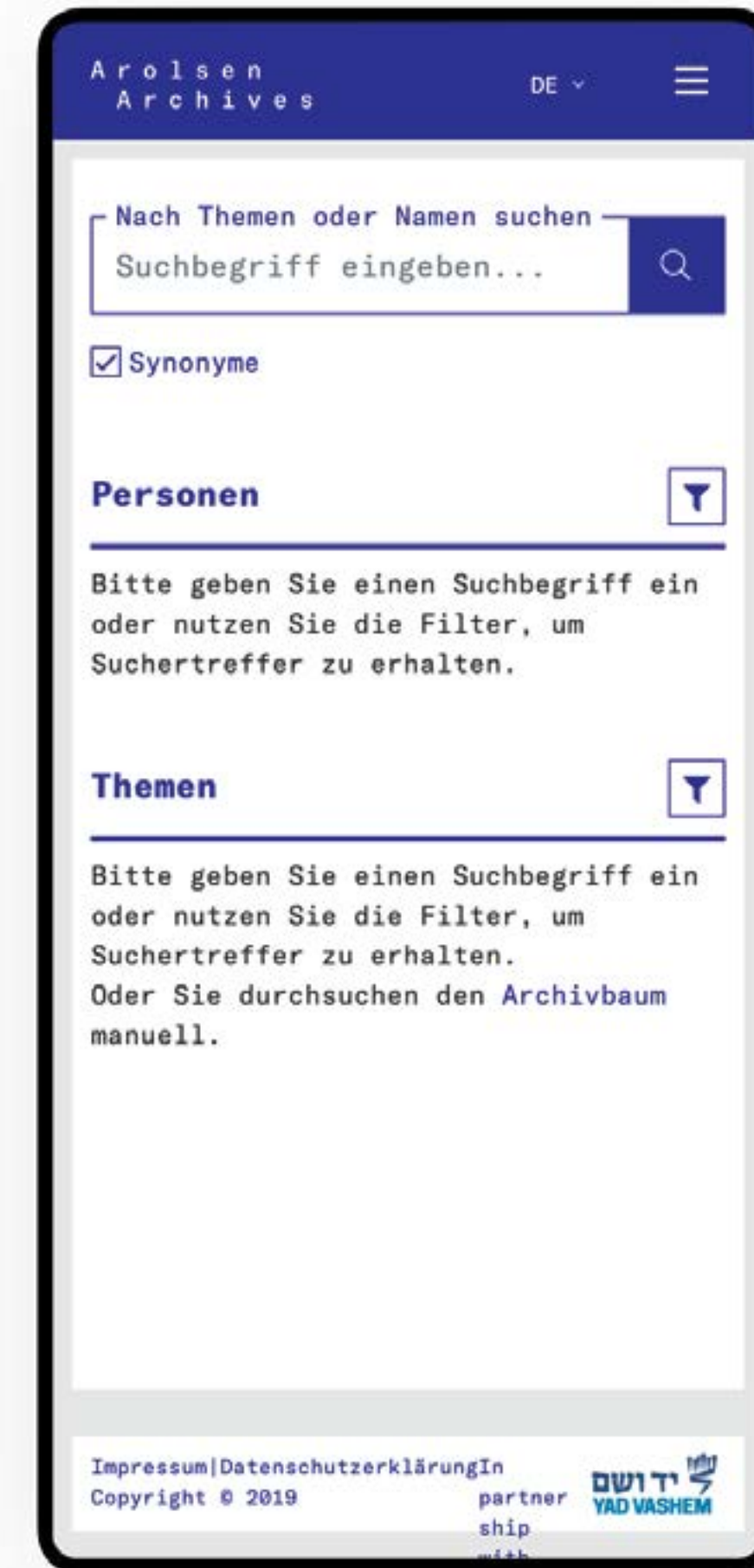
Mitarbeitern der Arolsen Archives möglich. Nur sie wissen, in welchen Dokumentengruppen sie suchen müssen.

„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.“



— AROLSEN ARCHIVES

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



„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

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.

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 zuzuordnen. 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.

AROLSEN ARCHIVES

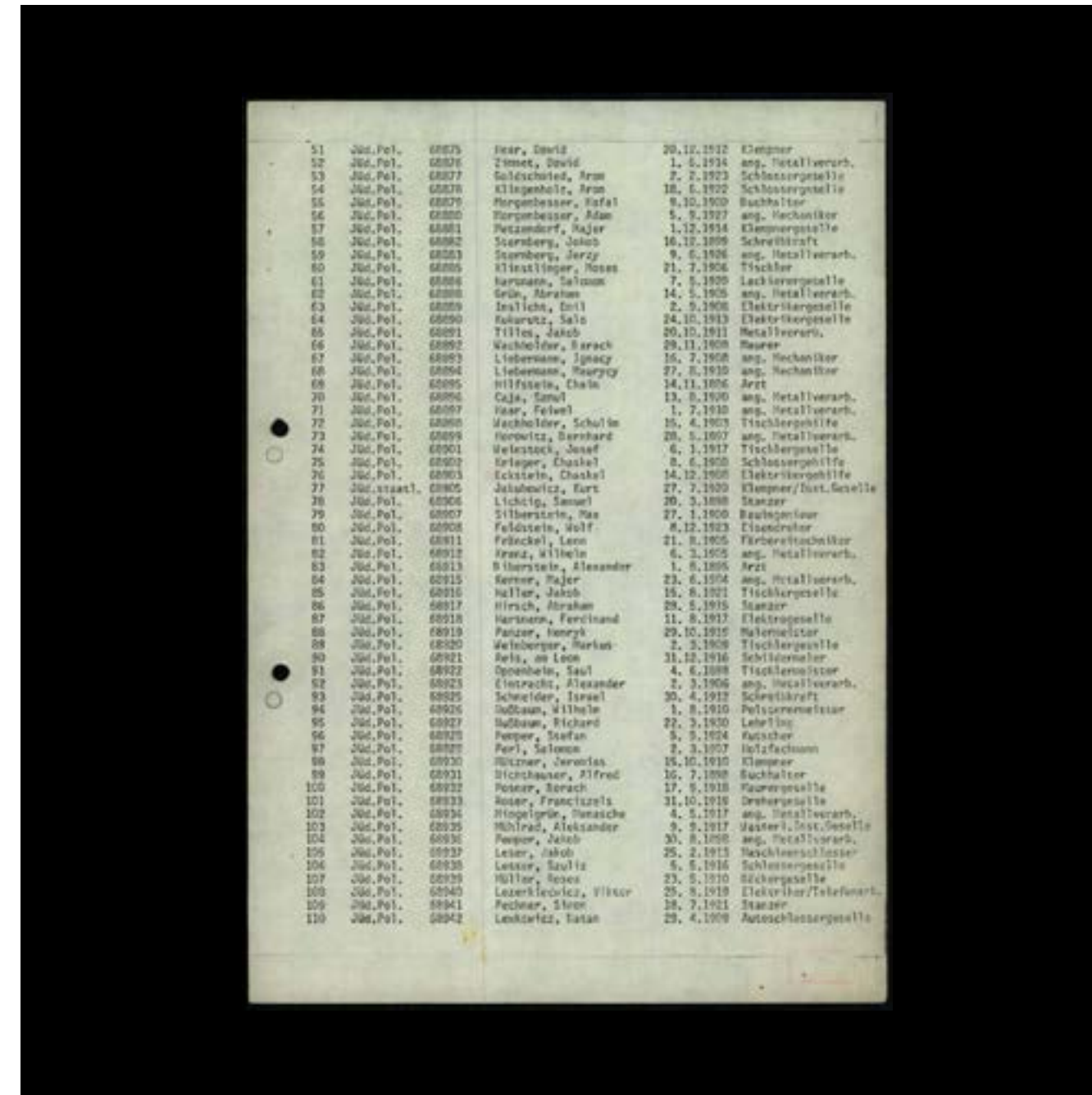
Im zweiten Projekt ging es zunächst darum, zum ersten Mal konkrete Namen aus Dokumenten auszulesen: 800 000 Listen – wieder aus allen möglichen Quellen aus dem Zeitraum von 1933 bis 1955 und wieder in sehr vielen Qualitätsvarianten.

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 unterschiedlichen Qualität der Quellen konnten ein Großteil der maschinengeschriebenen Listen und teilweise auch handgeschriebene Listen, erkannt und die Personendaten ausgelesen werden.



51	Jud. Pol.	00075	Veer, David	20.12.1912	Klempner
52	Jud. Pol.	00076	Zimmer, David	1. 6.1914	ang. Metallverarb.
53	Jud. Pol.	00077	Schleschler, Aron	2. 2.1912	Schlossergeselle
54	Jud. Pol.	00078	Klingenhilf, Aron	10. 6.1912	Schlossergeselle
55	Jud. Pol.	00079	Morgenbesser, Rafael	8. 10.1909	Buchhalter
56	Jud. Pol.	00080	Borgenbesser, Adam	5. 8.1917	ang. Mechaniker
57	Jud. Pol.	00081	Watzinger, Hajer	1.12.1914	Klempnergeselle
58	Jud. Pol.	00082	Sauerberg, Jakob	16.12.1908	Schweißstoff
59	Jud. Pol.	00083	Sauerberg, Jerzy	9. 6.1906	ang. Metallverarb.
60	Jud. Pol.	00084	Kleinlilger, Moses	21. 7.1906	Tischler
61	Jud. Pol.	00085	Karunan, Salomon	7. 6.1909	Lackierergeselle
62	Jud. Pol.	00086	Grub, Abraham	14. 5.1905	ang. Metallverarb.
63	Jud. Pol.	00087	Isakson, Isak	2. 5.1906	Elektrikergeselle
64	Jud. Pol.	00088	Kakurutz, Salo	24.10.1913	Elektrikergeselle
65	Jud. Pol.	00089	Tillen, Jakob	20.10.1911	Metallverarb.
66	Jud. Pol.	00090	Wacholder, Frank	29.11.1909	Stenogr.
67	Jud. Pol.	00091	Liebermann, Ignacy	16. 7.1908	ang. Mechaniker
68	Jud. Pol.	00092	Liebermann, Henryk	27. 6.1910	ang. Mechaniker
69	Jud. Pol.	00093	Hilfsbach, Chaim	14.11.1906	Arzt
70	Jud. Pol.	00094	Caja, Samel	13. 8.1909	ang. Metallverarb.
71	Jud. Pol.	00095	Vaser, Ferenc	1. 7.1910	ang. Metallverarb.
72	Jud. Pol.	00096	Wacholder, Schulze	16. 4.1907	Tischlergeselle
73	Jud. Pol.	00097	Furwitz, Bernhard	20. 5.1907	ang. Metallverarb.
74	Jud. Pol.	00098	Welsch, Josef	6. 1.1917	Tischlergeselle
75	Jud. Pol.	00099	Krieger, Euseb	8. 6.1908	Schlossergeselle
76	Jud. Pol.	00100	Eckstein, Chaskel	14.12.1909	Elektrikergeselle
77	Jud. Straßl.	00101	Jakubowitz, Hart	27. 7.1910	Stenogr./Inst. Geselle
78	Jud. Pol.	00102	Lichtig, Samuel	20. 3.1908	Stenogr.
79	Jud. Pol.	00103	Silberstein, Max	27. 1.1909	Bauingenieur
80	Jud. Pol.	00104	Feldstein, Isak	4.12.1913	Klempner
81	Jud. Pol.	00105	Feldstein, Leon	21. 8.1905	Führer/Mechaniker
82	Jud. Pol.	00106	Kraus, Wilhelm	6. 1.1905	ang. Metallverarb.
83	Jud. Pol.	00107	Bornstein, Alexander	1. 8.1905	Arzt
84	Jud. Pol.	00108	Berner, Hajer	23. 6.1904	ang. Metallverarb.
85	Jud. Pol.	00109	Haller, Jakob	15. 8.1911	Tischlergeselle
86	Jud. Pol.	00110	Hirsch, Mordechai	29. 5.1919	Stenogr.
87	Jud. Pol.	00111	Hartmann, Ferdinand	11. 8.1917	Elektrikergeselle
88	Jud. Pol.	00112	Panzer, Henryk	29.10.1919	Malermeister
89	Jud. Pol.	00113	Welschberger, Moritz	2. 5.1908	Tischlergeselle
90	Jud. Pol.	00114	Pels, von Leon	31.12.1916	Schlossergeselle
91	Jud. Pol.	00115	Oppenheimer, Saul	4. 4.1908	Tischlermeister
92	Jud. Pol.	00116	Liebowitz, Alexander	8. 1.1906	ang. Metallverarb.
93	Jud. Pol.	00117	Schneider, Isakel	30. 4.1912	Schweißstoff
94	Jud. Pol.	00118	Goldman, Wilhelm	1. 8.1910	Polstermeister
95	Jud. Pol.	00119	Goldman, Richard	22. 3.1910	Lehrer
96	Jud. Pol.	00120	Pumper, Isak	5. 5.1904	Kauscher
97	Jud. Pol.	00121	Perl, Selomon	2. 3.1917	Hilfschmied
98	Jud. Pol.	00122	Witzner, Jovanias	16.10.1913	Stenogr.
99	Jud. Pol.	00123	Hirschauer, Friedl	16. 7.1909	Buchhalter
100	Jud. Pol.	00124	Pfeiffer, Ezechiel	17. 8.1913	Flussbergbau
101	Jud. Pol.	00125	Rosen, Franziska	31.10.1918	Druckergeselle
102	Jud. Pol.	00126	Hingelberg, Hansasche	4. 1.1917	ang. Metallverarb.
103	Jud. Pol.	00127	Wolfsberg, Alexander	8. 8.1917	Bauverl. Inst. Geselle
104	Jud. Pol.	00128	Pumper, Jakob	30. 8.1908	ang. Metallverarb.
105	Jud. Pol.	00129	Leiser, Jakob	25. 2.1913	Wacholdermeister
106	Jud. Pol.	00130	Leiser, Sussie	6. 6.1916	Schlossergeselle
107	Jud. Pol.	00131	Hiltner, Hesse	23. 6.1910	Schlossergeselle
108	Jud. Pol.	00132	Laserleukitz, Viktor	25. 8.1919	Elektriker/Telefonverl.
109	Jud. Pol.	00133	Pechner, Simeon	18. 7.1911	Stenogr.
110	Jud. Pol.	00134	Leukowitz, Simeon	29. 4.1909	Aufschlossergeselle

„In Handarbeit hätte das Jahre gedauert.“

TWT-Projektleiter Sebastian Butz

TWT

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

