BERD
@NFDI

Focused Tutorial on Capturing, Enriching, Disseminating Research Data Objects

# Extracting research data from historical documents with eScriptorium and Python

Jan Kamlah, Thomas Schmidt and Renat Shigapov
Universitätsbibliothek Mannheim

24.11.2022

1. Einführung & Digitalisierung
2. Layoutsegmentierung und OCR via eScriptorium
3. Datenextraktion und –strukturierung via Python
4. Fazit

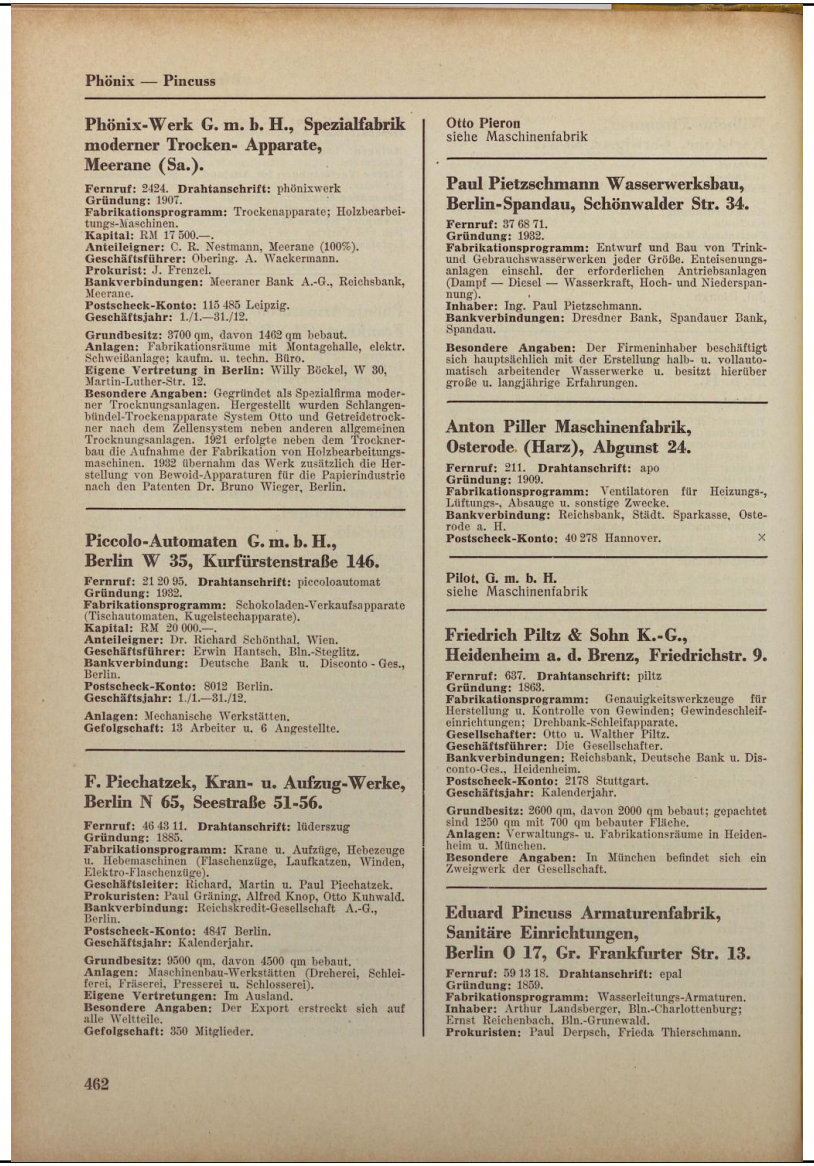
# 1. Einführung & Digitalisierung

# 1. Einführung & Digitalisierung

- Anfrage zur Extraktion von Forschungsdaten aus "**Die Maschinen-Industrie im Deutschen Reich von 1937**" durch Prof. Jochen Streb (Professur für Wirtschaftsgeschichte @ Uni Mannheim)

- Kooperationsprojekt von BERD@NFDI und OCR-D Modulprojekt “Werkspezifisches Training” @ UB Mannheim

- Digitalisierung (654 Seiten) durch Digitalisierungswerkstatt der UB Mannheim



Phönix — Pincuss

**Phönix-Werk G. m. b. H., Spezialfabrik moderner Trocken- Apparate, Meerane (Sa.).**

**Fernruf:** 2424. **Drahtanschrift:** phönixwerk
**Gründung:** 1907.
**Fabrikationsprogramm:** Trockenapparate; Holzbearbeitungs-Maschinen.
**Kapital:** RM 17 500.—.
**Anteileigner:** C. R. Nestmann, Meerane (100%).
**Geschäftsführer:** Obering. A. Wackermann.
**Prokurist:** J. Frenzel.
**Bankverbindungen:** Meeraner Bank A.-G., Reichsbank, Meerane.
**Postscheck-Konto:** 115 485 Leipzig.
**Geschäftsjahr:** 1./1.—31./12.

**Grundbesitz:** 3700 qm, davon 1462 qm bebaut.
**Anlagen:** Fabrikationsräume mit Montagehalle, elektr. Schweißanlage; kaufm. u. techn. Büro.
**Eigene Vertretung in Berlin:** Willy Böckel, W 30, Martin-Luther-Str. 12.
**Besondere Angaben:** Gegründet als Spezialfirma moderner Trocknungsanlagen. Hergestellt wurden Schlangenbündel-Trockenapparate System Otto und Getreidetrockner nach dem Zellensystem neben anderen allgemeinen Trocknungsanlagen. 1921 erfolgte neben dem Trocknerbau die Aufnahme der Fabrikation von Holzbearbeitungsmaschinen. 1932 übernahm das Werk zusätzlich die Herstellung von Bewoid-Apparaturen für die Papierindustrie nach den Patenten Dr. Bruno Wieger, Berlin.

**Piccolo-Automaten G. m. b. H., Berlin W 35, Kurfürstenstraße 146.**

**Fernruf:** 21 20 95. **Drahtanschrift:** piccoloautomat
**Gründung:** 1932.
**Fabrikationsprogramm:** Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstechapparate).
**Kapital:** RM 20 000.—.
**Anteileigner:** Dr. Richard Schönthal, Wien.
**Geschäftsführer:** Erwin Hantsch, Bln.-Steglitz.
**Bankverbindung:** Deutsche Bank u. Disconto - Ges., Berlin.
**Postscheck-Konto:** 8012 Berlin.
**Geschäftsjahr:** 1./1.—31./12.

**Anlagen:** Mechanische Werkstätten.
**Gefolgschaft:** 13 Arbeiter u. 6 Angestellte.

**F. Piechatzek, Kran- u. Aufzug-Werke, Berlin N 65, Seestraße 51-56.**

**Fernruf:** 46 43 11. **Drahtanschrift:** lüderszug
**Gründung:** 1885.
**Fabrikationsprogramm:** Krane u. Aufzüge, Hebezeuge u. Hebemaschinen (Flaschenzüge, Laufkatzen, Winden, Elektro-Flaschenzüge).
**Geschäftsleiter:** Richard, Martin u. Paul Piechatzek.
**Prokuristen:** Paul Gräning, Alfred Knop, Otto Kuhwald.
**Bankverbindung:** Reichskredit-Gesellschaft A.-G., Berlin.
**Postscheck-Konto:** 4847 Berlin.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 9500 qm, davon 4500 qm bebaut.
**Anlagen:** Maschinenbau-Werkstätten (Dreherei, Schleiferei, Fräserei, Presserei u. Schlosserei).
**Eigene Vertretungen:** Im Ausland.
**Besondere Angaben:** Der Export erstreckt sich auf alle Weltteile.
**Gefolgschaft:** 350 Mitglieder.

Otto Pieron
siehe Maschinenfabrik

**Paul Pietzschmann Wasserwerksbau, Berlin-Spandau, Schönwalder Str. 34.**

**Fernruf:** 37 68 71.
**Gründung:** 1932.
**Fabrikationsprogramm:** Entwurf und Bau von Trink- und Gebrauchswasserwerken jeder Größe. Enteisenungsanlagen einschl. der erforderlichen Antriebsanlagen (Dampf — Diesel — Wasserkraft, Hoch- und Niederspannung).
**Inhaber:** Ing. Paul Pietzschmann.
**Bankverbindungen:** Dresdner Bank, Spandauer Bank, Spandau.

**Besondere Angaben:** Der Firmeninhaber beschäftigt sich hauptsächlich mit der Erstellung halb- u. vollautomatisch arbeitender Wasserwerke u. besitzt hierüber große u. langjährige Erfahrungen.

**Anton Piller Maschinenfabrik, Osterode. (Harz), Abgunst 24.**

**Fernruf:** 211. **Drahtanschrift:** apo
**Gründung:** 1909.
**Fabrikationsprogramm:** Ventilatoren für Heizungs-, Lüftungs-, Absauge u. sonstige Zwecke.
**Bankverbindung:** Reichsbank, Städt. Sparkasse, Osterode a. H.
**Postscheck-Konto:** 40 278 Hannover.

Pilot, G. m. b. H.
siehe Maschinenfabrik

**Friedrich Piltz & Sohn K.-G., Heidenheim a. d. Brenz, Friedrichstr. 9.**

**Fernruf:** 637. **Drahtanschrift:** piltz
**Gründung:** 1863.
**Fabrikationsprogramm:** Genauigkeitswerkzeuge für Herstellung u. Kontrolle von Gewinden; Gewindeschleifeinrichtungen; Drehbank-Schleifapparate.
**Gesellschafter:** Otto u. Walther Piltz.
**Geschäftsführer:** Die Gesellschafter.
**Bankverbindungen:** Reichsbank, Deutsche Bank u. Disconto-Ges., Heidenheim.
**Postscheck-Konto:** 2178 Stuttgart.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 2600 qm, davon 2000 qm bebaut; gepachtet sind 1250 qm mit 700 qm bebauter Fläche.
**Anlagen:** Verwaltungs- u. Fabrikationsräume in Heidenheim u. München.
**Besondere Angaben:** In München befindet sich ein Zweigwerk der Gesellschaft.

**Eduard Pincuss Armaturenfabrik, Sanitäre Einrichtungen, Berlin O 17, Gr. Frankfurter Str. 13.**

**Fernruf:** 59 13 18. **Drahtanschrift:** epal
**Gründung:** 1859.
**Fabrikationsprogramm:** Wasserleitungs-Armaturen.
**Inhaber:** Arthur Landsberger, Bln.-Charlottenburg; Ernst Reichenbach, Bln.-Grunewald.
**Prokuristen:** Paul Derpsch, Frieda Thierschmann.

462

# 1. Einführung & Digitalisierung

**Forschungsfrage:**

- Identifizierung von Rechtsformwechseln der gelisteten Unternehmen
- Hierfür Extraktion des **Firmennamens** sowie der **Rechtsformen** notwendig

**Piccolo-Automaten G. m. b. H.,**
**Berlin W 35, Kurfürstenstraße 146.**

**Fernruf:** 21 20 95. **Drahtanschrift:** piccoloautomat
**Gründung:** 1932.
**Fabrikationsprogramm:** Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstechapparate).
**Kapital:** RM 20 000.—.
**Anteileigner:** Dr. Richard Schönthal, Wien.
**Geschäftsführer:** Erwin Hantsch, Bln.-Steglitz.
**Bankverbindung:** Deutsche Bank u. Disconto - Ges., Berlin.
**Postscheck-Konto:** 8012 Berlin.
**Geschäftsjahr:** 1./1.—31./12.

**Anlagen:** Mechanische Werkstätten.
**Gefolgschaft:** 13 Arbeiter u. 6 Angestellte.

# 1. Einführung & Digitalisierung

**Forschungsfrage:**

- Identifizierung von Rechtsformwechseln der gelisteten Unternehmen
- Hierfür Extraktion des **Firmennamens** sowie der **Rechtsformen** notwendig

**Piccolo-Automaten G. m. b. H.,**
**Berlin W 35, Kurfürstenstraße 146.**

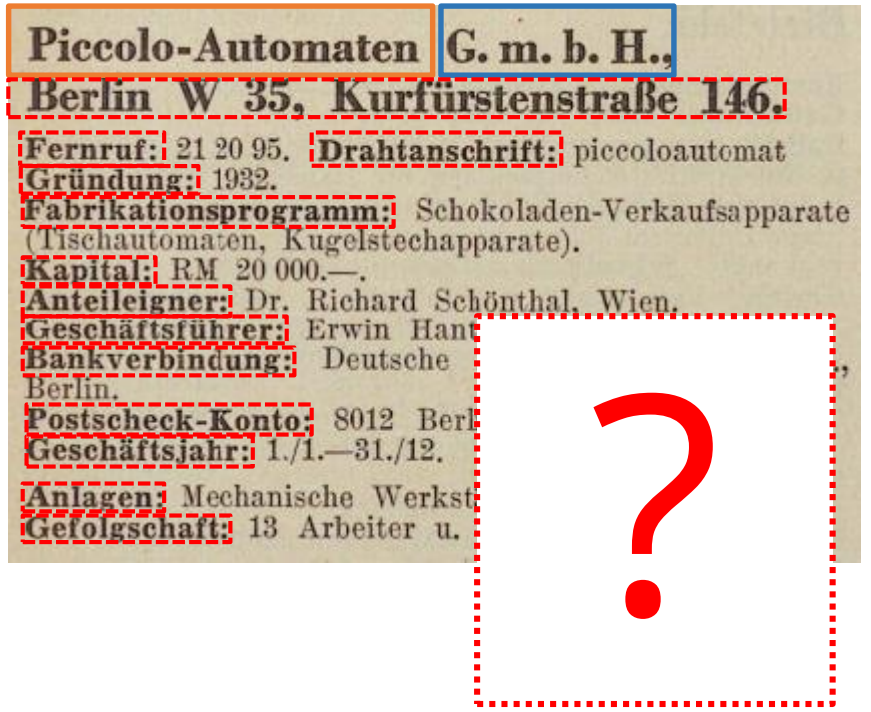
**Fernruf:** 21 20 95. **Drahtanschrift:** piccoloautomat
**Gründung:** 1932.
**Fabrikationsprogramm:** Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstechapparate).
**Kapital:** RM 20 000.—.
**Anteileigner:** Dr. Richard Schönthal, Wien.
**Geschäftsführer:** Erwin Hantsch, Bln.-Steglitz.
**Bankverbindung:** Deutsche Bank u. Disconto - Ges., Berlin.
**Postscheck-Konto:** 8012 Berlin.
**Geschäftsjahr:** 1./1.—31./12.

**Anlagen:** Mechanische Werkstätten.
**Gefolgschaft:** 13 Arbeiter u. 6 Angestellte.

# 1. Einführung & Digitalisierung

**Forschungsfrage:**

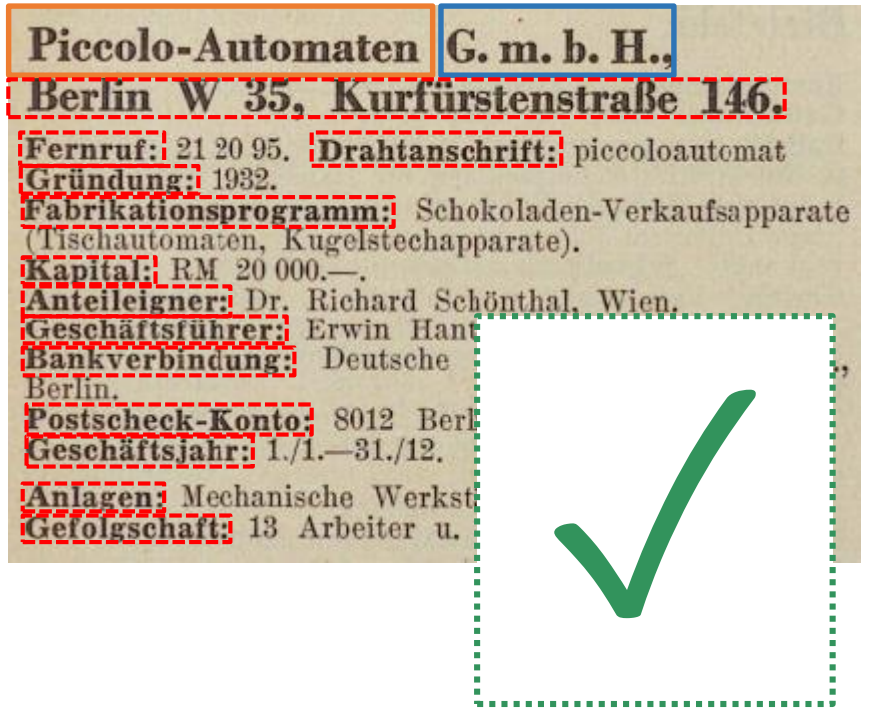
- Identifizierung von Rechtsformwechseln der gelisteten Unternehmen
- Hierfür Extraktion des **Firmennamens** sowie der **Rechtsformen** notwendig



# 1. Einführung & Digitalisierung

**Projektansatz und –ziel:**

- Daten über die Forschungsfrage hinaus extrahieren und strukturieren (realistischer Mehraufwand)
- Wiederverwendbarkeit der Daten für andere Forschungszwecke sicherstellen
- Testworkflow, um Erfahrungen für vergleichbare Projekte zu sammeln



## Projektansatz und –ziel:

### JPG

**Piccolo-Automaten G. m. b. H.,**
**Berlin W 35, Kurfürstenstraße 146.**

**Fernruf:** 21 20 95. **Drahtanschrift:** piccoloautomat
**Gründung:** 1932.
**Fabrikationsprogramm:** Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstechapparate).
**Kapital:** RM 20 000.—.
**Anteileigner:** Dr. Richard Schönthal, Wien.
**Geschäftsführer:** Erwin Hantsch, Bln.-Steglitz.
**Bankverbindung:** Deutsche Bank u. Disconto - Ges., Berlin.
**Postscheck-Konto:** 8012 Berlin.
**Geschäftsjahr:** 1./1.—31./12.

**Anlagen:** Mechanische Werkstätten.
**Gefolgschaft:** 13 Arbeiter u. 6 Angestellte.

Transformation

### Strukturierte Forschungsdaten

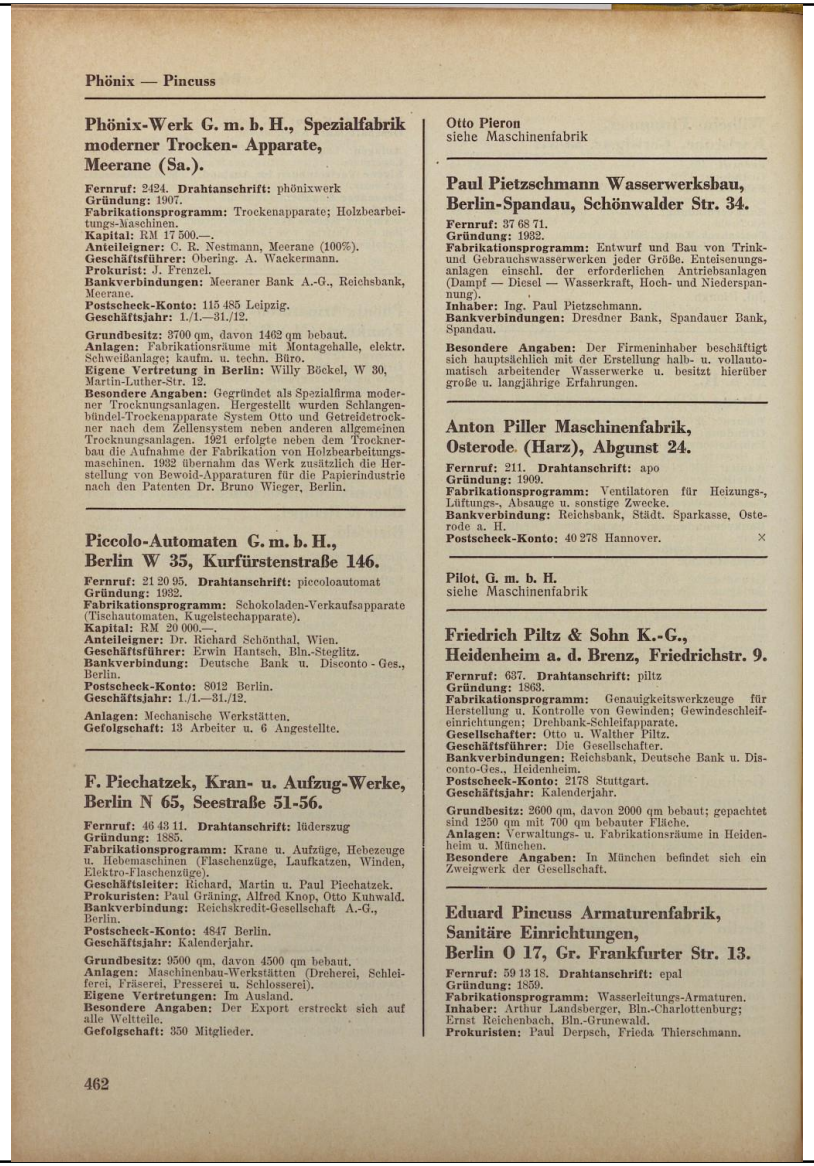
| | |
|---|---|
| Firmenname | Piccolo-Automaten |
| Rechtsform | G.m.b.H. |
| Sitz | Berlin W 35 Kurfürstenstraße 146 |
| Fernruf | 212095 |
| Drahtanschrift | piccoloautomat |
| Gründung | 1932 |
| Produkt | Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstechapparate) |

...

# 1. Einführung & Digitalisierung

**Herausforderungen (OCR):**

- Hohe Anforderungen an Qualität für Forschungsdaten
- Aufwand für OCR Nachtraining schwer abzuschätzen
- Layout (zweispaltig → Reading Order)



Phönix — Pincuss

**Phönix-Werk G. m. b. H., Spezialfabrik moderner Trocken- Apparate, Meerane (Sa.).**

**Fernruf:** 2424. **Drahtanschrift:** phönixwerk
**Gründung:** 1907.
**Fabrikationsprogramm:** Trockenapparate; Holzbearbeitungs-Maschinen.
**Kapital:** RM 17 500.—.
**Anteileigner:** C. R. Nestmann, Meerane (100%).
**Geschäftsführer:** Obering. A. Wackermann.
**Prokurist:** J. Frenzel.
**Bankverbindungen:** Meeraner Bank A.-G., Reichsbank, Meerane.
**Postscheck-Konto:** 115 485 Leipzig.
**Geschäftsjahr:** 1./1.—31./12.

**Grundbesitz:** 3700 qm, davon 1462 qm bebaut.
**Anlagen:** Fabrikationsräume mit Montagehalle, elektr. Schweißanlage; kaufm. u. techn. Büro.
**Eigene Vertretung in Berlin:** Willy Böckel, W 30, Martin-Luther-Str. 12.
**Besondere Angaben:** Gegründet als Spezialfirma moderner Trocknungsanlagen. Hergestellt wurden Schlangenbündel-Trockenapparate System Otto und Getreidetrockner nach dem Zellensystem neben anderen allgemeinen Trocknungsanlagen. 1921 erfolgte neben dem Trocknerbau die Aufnahme der Fabrikation von Holzbearbeitungsmaschinen. 1932 übernahm das Werk zusätzlich die Herstellung von Bewoid-Apparaturen für die Papierindustrie nach den Patenten Dr. Bruno Wieger, Berlin.

**Piccolo-Automaten G. m. b. H., Berlin W 35, Kurfürstenstraße 146.**

**Fernruf:** 21 20 95. **Drahtanschrift:** piccoloautomat
**Gründung:** 1932.
**Fabrikationsprogramm:** Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstechapparate).
**Kapital:** RM 20 000.—.
**Anteileigner:** Dr. Richard Schönthal, Wien.
**Geschäftsführer:** Erwin Hantsch, Bln.-Steglitz.
**Bankverbindung:** Deutsche Bank u. Disconto - Ges., Berlin.
**Postscheck-Konto:** 8012 Berlin.
**Geschäftsjahr:** 1./1.—31./12.

**Anlagen:** Mechanische Werkstätten.
**Gefolgschaft:** 13 Arbeiter u. 6 Angestellte.

**F. Piechatzek, Kran- u. Aufzug-Werke, Berlin N 65, Seestraße 51-56.**

**Fernruf:** 46 43 11. **Drahtanschrift:** lüderszug
**Gründung:** 1885.
**Fabrikationsprogramm:** Krane u. Aufzüge, Hebezeuge u. Hebemaschinen (Flaschenzüge, Laufkatzen, Winden, Elektro-Flaschenzüge).
**Geschäftsleiter:** Richard, Martin u. Paul Piechatzek.
**Prokuristen:** Paul Gräning, Alfred Knop, Otto Kuhwald.
**Bankverbindung:** Reichskredit-Gesellschaft A.-G., Berlin.
**Postscheck-Konto:** 4847 Berlin.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 9500 qm, davon 4500 qm bebaut.
**Anlagen:** Maschinenbau-Werkstätten (Dreherei, Schleiferei, Fräserei, Presserei u. Schlosserei).
**Eigene Vertretungen:** Im Ausland.
**Besondere Angaben:** Der Export erstreckt sich auf alle Weltteile.
**Gefolgschaft:** 350 Mitglieder.

Otto Pieron
siehe Maschinenfabrik

**Paul Pietzschmann Wasserwerksbau, Berlin-Spandau, Schönwalder Str. 34.**

**Fernruf:** 37 68 71.
**Gründung:** 1932.
**Fabrikationsprogramm:** Entwurf und Bau von Trink- und Gebrauchswasserwerken jeder Größe. Enteisenungsanlagen einschl. der erforderlichen Antriebsanlagen (Dampf — Diesel — Wasserkraft, Hoch- und Niederspannung).
**Inhaber:** Ing. Paul Pietzschmann.
**Bankverbindungen:** Dresdner Bank, Spandauer Bank, Spandau.

**Besondere Angaben:** Der Firmeninhaber beschäftigt sich hauptsächlich mit der Erstellung halb- u. vollautomatisch arbeitender Wasserwerke u. besitzt hierüber große u. langjährige Erfahrungen.

**Anton Piller Maschinenfabrik, Osterode (Harz), Abgunst 24.**

**Fernruf:** 211. **Drahtanschrift:** apo
**Gründung:** 1909.
**Fabrikationsprogramm:** Ventilatoren für Heizungs-, Lüftungs-, Absauge u. sonstige Zwecke.
**Bankverbindung:** Reichsbank, Städt. Sparkasse, Osterode a. H.
**Postscheck-Konto:** 40 278 Hannover.

Pilot, G. m. b. H.
siehe Maschinenfabrik

**Friedrich Piltz & Sohn K.-G., Heidenheim a. d. Brenz, Friedrichstr. 9.**

**Fernruf:** 637. **Drahtanschrift:** piltz
**Gründung:** 1863.
**Fabrikationsprogramm:** Genauigkeitswerkzeuge für Herstellung u. Kontrolle von Gewinden; Gewindeschleifeinrichtungen; Drehbank-Schleifapparate.
**Gesellschafter:** Otto u. Walther Piltz.
**Geschäftsführer:** Die Gesellschafter.
**Bankverbindungen:** Reichsbank, Deutsche Bank u. Disconto-Ges., Heidenheim.
**Postscheck-Konto:** 2178 Stuttgart.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 2600 qm, davon 2000 qm bebaut; gepachtet sind 1250 qm mit 700 qm bebauter Fläche.
**Anlagen:** Verwaltungs- u. Fabrikationsräume in Heidenheim u. München.
**Besondere Angaben:** In München befindet sich ein Zweigwerk der Gesellschaft.

**Eduard Pincuss Armaturenfabrik, Sanitäre Einrichtungen, Berlin O 17, Gr. Frankfurter Str. 13.**

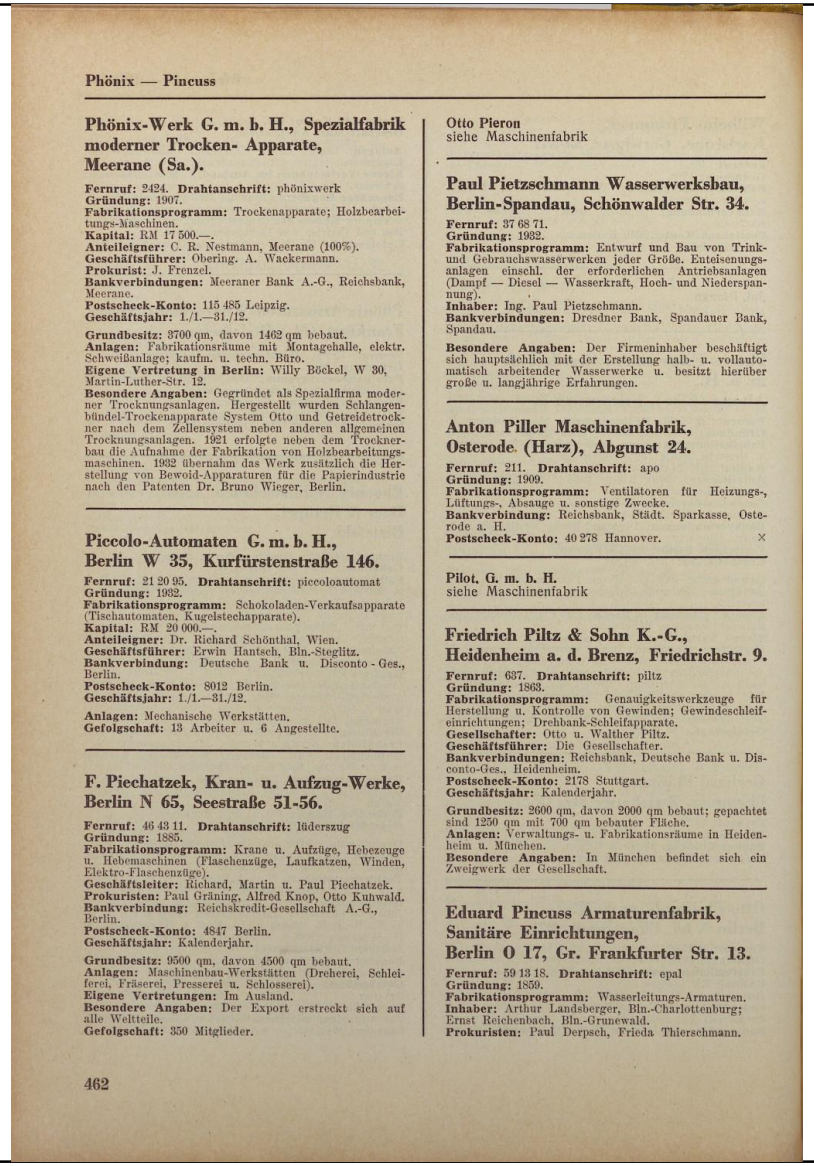
**Fernruf:** 59 13 18. **Drahtanschrift:** epal
**Gründung:** 1859.
**Fabrikationsprogramm:** Wasserleitungs-Armaturen.
**Inhaber:** Arthur Landsberger, Bln.-Charlottenburg; Ernst Reichenbach, Bln.-Grunewald.
**Prokuristen:** Paul Derpsch, Frieda Thierschmann.

462

# 1. Einführung & Digitalisierung

**Herausforderungen (Datenstrukturierung):**

- Keine “all-in-one”-Lösung vorhanden: Software muss neu geschrieben werden
- Hohe Anforderungen an Qualität von Forschungsdaten
- Profiltrennung
- Inkonsistente Attributbezeichnungen:
  - *Postscheck-Konto, Postschekkonto, Postcheck-Konto*
  - *Geschäftsführer, Geschäftsleiter, Direktor, Betriebsführer*
  - *…*



Phönix — Pincuss

**Phönix-Werk G. m. b. H., Spezialfabrik moderner Trocken- Apparate, Meerane (Sa.).**

**Fernruf:** 2424. **Drahtanschrift:** phönixwerk
**Gründung:** 1907.
**Fabrikationsprogramm:** Trockenapparate; Holzbearbeitungs-Maschinen.
**Kapital:** RM 17 500.—.
**Anteileigner:** C. R. Nestmann, Meerane (100%).
**Geschäftsführer:** Obering. A. Wackermann.
**Prokurist:** J. Frenzel.
**Bankverbindungen:** Meeraner Bank A.-G., Reichsbank, Meerane.
**Postscheck-Konto:** 115 485 Leipzig.
**Geschäftsjahr:** 1./1.—31./12.

**Grundbesitz:** 3700 qm, davon 1462 qm bebaut.
**Anlagen:** Fabrikationsräume mit Montagehalle, elektr. Schweißanlage; kaufm. u. techn. Büro.
**Eigene Vertretung in Berlin:** Willy Böckel, W 30, Martin-Luther-Str. 12.
**Besondere Angaben:** Gegründet als Spezialfirma moderner Trocknungsanlagen. Hergestellt wurden Schlangenbündel-Trockenapparate System Otto und Getreidetrockner nach dem Zellensystem neben anderen allgemeinen Trocknungsanlagen. 1921 erfolgte neben dem Trocknerbau die Aufnahme der Fabrikation von Holzbearbeitungsmaschinen. 1932 übernahm das Werk zusätzlich die Herstellung von Bewoid-Apparaturen für die Papierindustrie nach den Patenten Dr. Bruno Wieger, Berlin.

**Piccolo-Automaten G. m. b. H., Berlin W 35, Kurfürstenstraße 146.**

**Fernruf:** 21 20 95. **Drahtanschrift:** piccoloautomat
**Gründung:** 1932.
**Fabrikationsprogramm:** Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstechapparate).
**Kapital:** RM 20 000.—.
**Anteileigner:** Dr. Richard Schönthal, Wien.
**Geschäftsführer:** Erwin Hantsch, Bln.-Steglitz.
**Bankverbindung:** Deutsche Bank u. Disconto - Ges., Berlin.
**Postscheck-Konto:** 8012 Berlin.
**Geschäftsjahr:** 1./1.—31./12.

**Anlagen:** Mechanische Werkstätten.
**Gefolgschaft:** 13 Arbeiter u. 6 Angestellte.

**F. Piechatzek, Kran- u. Aufzug-Werke, Berlin N 65, Seestraße 51-56.**

**Fernruf:** 46 43 11. **Drahtanschrift:** lüderszug
**Gründung:** 1885.
**Fabrikationsprogramm:** Krane u. Aufzüge, Hebezeuge u. Hebemaschinen (Flaschenzüge, Laufkatzen, Winden, Elektro-Flaschenzüge).
**Geschäftsleiter:** Richard, Martin u. Paul Piechatzek.
**Prokuristen:** Paul Gräning, Alfred Knop, Otto Kuhwald.
**Bankverbindung:** Reichskredit-Gesellschaft A.-G., Berlin.
**Postscheck-Konto:** 4847 Berlin.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 9500 qm, davon 4500 qm bebaut.
**Anlagen:** Maschinenbau-Werkstätten (Dreherei, Schleiferei, Fräserei, Presserei u. Schlosserei).
**Eigene Vertretungen:** Im Ausland.
**Besondere Angaben:** Der Export erstreckt sich auf alle Weltteile.
**Gefolgschaft:** 350 Mitglieder.

Otto Pieron
siehe Maschinenfabrik

**Paul Pietzschmann Wasserwerksbau, Berlin-Spandau, Schönwalder Str. 34.**

**Fernruf:** 37 68 71.
**Gründung:** 1932.
**Fabrikationsprogramm:** Entwurf und Bau von Trink- und Gebrauchswasserwerken jeder Größe. Enteisenungsanlagen einschl. der erforderlichen Antriebsanlagen (Dampf — Diesel — Wasserkraft, Hoch- und Niederspannung).
**Inhaber:** Ing. Paul Pietzschmann.
**Bankverbindungen:** Dresdner Bank, Spandauer Bank, Spandau.

**Besondere Angaben:** Der Firmeninhaber beschäftigt sich hauptsächlich mit der Erstellung halb- u. vollautomatisch arbeitender Wasserwerke u. besitzt hierüber große u. langjährige Erfahrungen.

**Anton Piller Maschinenfabrik, Osterode. (Harz), Abgunst 24.**

**Fernruf:** 211. **Drahtanschrift:** apo
**Gründung:** 1909.
**Fabrikationsprogramm:** Ventilatoren für Heizungs-, Lüftungs-, Absauge u. sonstige Zwecke.
**Bankverbindung:** Reichsbank, Städt. Sparkasse, Osterode a. H.
**Postscheck-Konto:** 40 278 Hannover. ×

Pilot, G. m. b. H.
siehe Maschinenfabrik

**Friedrich Piltz & Sohn K.-G., Heidenheim a. d. Brenz, Friedrichstr. 9.**

**Fernruf:** 637. **Drahtanschrift:** piltz
**Gründung:** 1863.
**Fabrikationsprogramm:** Genauigkeitswerkzeuge für Herstellung u. Kontrolle von Gewinden; Gewindeschleifeinrichtungen; Drehbank-Schleifapparate.
**Gesellschafter:** Otto u. Walther Piltz.
**Geschäftsführer:** Die Gesellschafter.
**Bankverbindungen:** Reichsbank, Deutsche Bank u. Disconto-Ges., Heidenheim.
**Postscheck-Konto:** 2178 Stuttgart.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 2600 qm, davon 2000 qm bebaut; gepachtet sind 1250 qm mit 700 qm bebauter Fläche.
**Anlagen:** Verwaltungs- u. Fabrikationsräume in Heidenheim u. München.
**Besondere Angaben:** In München befindet sich ein Zweigwerk der Gesellschaft.

**Eduard Pincuss Armaturenfabrik, Sanitäre Einrichtungen, Berlin O 17, Gr. Frankfurter Str. 13.**

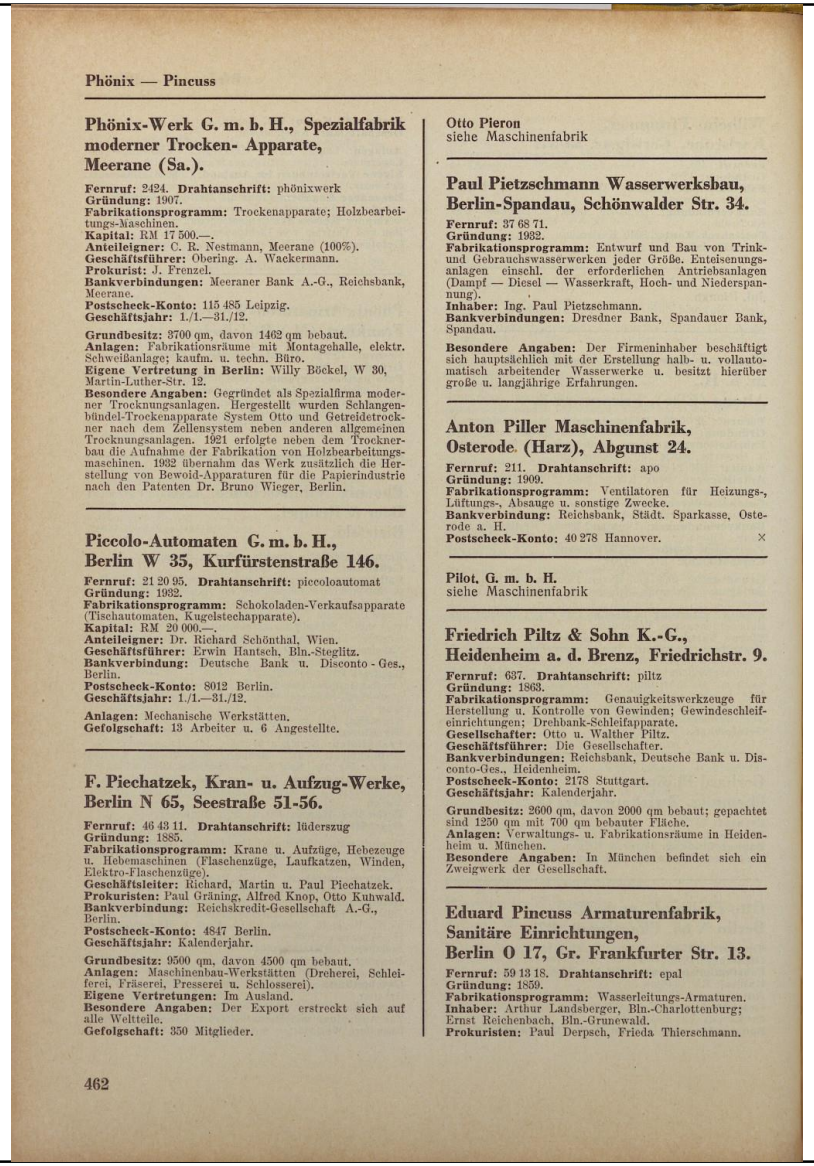
**Fernruf:** 59 13 18. **Drahtanschrift:** epal
**Gründung:** 1859.
**Fabrikationsprogramm:** Wasserleitungs-Armaturen.
**Inhaber:** Arthur Landsberger, Bln.-Charlottenburg; Ernst Reichenbach, Bln.-Grunewald.
**Prokuristen:** Paul Derpsch, Frieda Thierschmann.

462

# 1. Einführung & Digitalisierung

**Herausforderungen (Infrastruktur, Personal):**

- Mehrteiliger Workflow mit spezifischen infrastrukturellen und personellen Anforderungen
- Digitalisierung (1 Projektkoordinator, 1 Hiwi)
- OCR (1 Projektkoordinator, 1 Entwickler)
- Datenstrukturierung (1 Entwickler)
- OCR-Server (eScriptorium)



Phönix — Pincuss

**Phönix-Werk G. m. b. H., Spezialfabrik moderner Trocken- Apparate, Meerane (Sa.).**

**Fernruf:** 2424. **Drahtanschrift:** phönixwerk
**Gründung:** 1907.
**Fabrikationsprogramm:** Trockenapparate; Holzbearbeitungs-Maschinen.
**Kapital:** RM 17 500.—.
**Anteileigner:** C. R. Nestmann, Meerane (100%).
**Geschäftsführer:** Obering. A. Wackermann.
**Prokurist:** J. Frenzel.
**Bankverbindungen:** Meeraner Bank A.-G., Reichsbank, Meerane.
**Postscheck-Konto:** 115 485 Leipzig.
**Geschäftsjahr:** 1./1.—31./12.

**Grundbesitz:** 3700 qm, davon 1462 qm bebaut.
**Anlagen:** Fabrikationsräume mit Montagehalle, elektr. Schweißanlage; kaufm. u. techn. Büro.
**Eigene Vertretung in Berlin:** Willy Böckel, W 30, Martin-Luther-Str. 12.
**Besondere Angaben:** Gegründet als Spezialfirma moderner Trocknungsanlagen. Hergestellt wurden Schlangenbündel-Trockenapparate System Otto und Getreidetrockner nach dem Zellensystem neben anderen allgemeinen Trocknungsanlagen. 1921 erfolgte neben dem Trocknerbau die Aufnahme der Fabrikation von Holzbearbeitungsmaschinen. 1932 übernahm das Werk zusätzlich die Herstellung von Bewoid-Apparaturen für die Papierindustrie nach den Patenten Dr. Bruno Wieger, Berlin.

**Piccolo-Automaten G. m. b. H., Berlin W 35, Kurfürstenstraße 146.**

**Fernruf:** 21 20 95. **Drahtanschrift:** piccoloautomat
**Gründung:** 1932.
**Fabrikationsprogramm:** Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstechapparate).
**Kapital:** RM 20 000.—.
**Anteileigner:** Dr. Richard Schönthal, Wien.
**Geschäftsführer:** Erwin Hantsch, Bln.-Steglitz.
**Bankverbindung:** Deutsche Bank u. Disconto - Ges., Berlin.
**Postscheck-Konto:** 8012 Berlin.
**Geschäftsjahr:** 1./1.—31./12.

**Anlagen:** Mechanische Werkstätten.
**Gefolgschaft:** 13 Arbeiter u. 6 Angestellte.

**F. Piechatzek, Kran- u. Aufzug-Werke, Berlin N 65, Seestraße 51-56.**

**Fernruf:** 46 43 11. **Drahtanschrift:** lüderszug
**Gründung:** 1885.
**Fabrikationsprogramm:** Krane u. Aufzüge, Hebezeuge u. Hebemaschinen (Flaschenzüge, Laufkatzen, Winden, Elektro-Flaschenzüge).
**Geschäftsleiter:** Richard, Martin u. Paul Piechatzek.
**Prokuristen:** Paul Gräning, Alfred Knop, Otto Kuhwald.
**Bankverbindung:** Reichskredit-Gesellschaft A.-G., Berlin.
**Postscheck-Konto:** 4847 Berlin.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 9500 qm, davon 4500 qm bebaut.
**Anlagen:** Maschinenbau-Werkstätten (Dreherei, Schleiferei, Fräserei, Presserei u. Schlosserei).
**Eigene Vertretungen:** Im Ausland.
**Besondere Angaben:** Der Export erstreckt sich auf alle Weltteile.
**Gefolgschaft:** 350 Mitglieder.

Otto Pieron
siehe Maschinenfabrik

**Paul Pietzschmann Wasserwerksbau, Berlin-Spandau, Schönwalder Str. 34.**

**Fernruf:** 37 68 71.
**Gründung:** 1932.
**Fabrikationsprogramm:** Entwurf und Bau von Trink- und Gebrauchswasserwerken jeder Größe. Enteisenungsanlagen einschl. der erforderlichen Antriebsanlagen (Dampf — Diesel — Wasserkraft, Hoch- und Niederspannung).
**Inhaber:** Ing. Paul Pietzschmann.
**Bankverbindungen:** Dresdner Bank, Spandauer Bank, Spandau.

**Besondere Angaben:** Der Firmeninhaber beschäftigt sich hauptsächlich mit der Erstellung halb- u. vollautomatisch arbeitender Wasserwerke u. besitzt hierüber große u. langjährige Erfahrungen.

**Anton Piller Maschinenfabrik, Osterode (Harz), Abgunst 24.**

**Fernruf:** 211. **Drahtanschrift:** apo
**Gründung:** 1909.
**Fabrikationsprogramm:** Ventilatoren für Heizungs-, Lüftungs-, Absauge u. sonstige Zwecke.
**Bankverbindung:** Reichsbank, Städt. Sparkasse, Osterode a. H.
**Postscheck-Konto:** 40 278 Hannover.

Pilot, G. m. b. H.
siehe Maschinenfabrik

**Friedrich Piltz & Sohn K.-G., Heidenheim a. d. Brenz, Friedrichstr. 9.**

**Fernruf:** 637. **Drahtanschrift:** piltz
**Gründung:** 1863.
**Fabrikationsprogramm:** Genauigkeitswerkzeuge für Herstellung u. Kontrolle von Gewinden; Gewindeschleifeinrichtungen; Drehbank-Schleifapparate.
**Gesellschafter:** Otto u. Walther Piltz.
**Geschäftsführer:** Die Gesellschafter.
**Bankverbindungen:** Reichsbank, Deutsche Bank u. Disconto-Ges., Heidenheim.
**Postscheck-Konto:** 2178 Stuttgart.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 2600 qm, davon 2000 qm bebaut; gepachtet sind 1250 qm mit 700 qm bebauter Fläche.
**Anlagen:** Verwaltungs- u. Fabrikationsräume in Heidenheim u. München.
**Besondere Angaben:** In München befindet sich ein Zweigwerk der Gesellschaft.

**Eduard Pincuss Armaturenfabrik, Sanitäre Einrichtungen, Berlin O 17, Gr. Frankfurter Str. 13.**

**Fernruf:** 59 13 18. **Drahtanschrift:** epal
**Gründung:** 1859.
**Fabrikationsprogramm:** Wasserleitungs-Armaturen.
**Inhaber:** Arthur Landsberger, Bln.-Charlottenburg; Ernst Reichenbach, Bln.-Grunewald.
**Prokuristen:** Paul Derpsch, Frieda Thierschmann.

462

# 2. Layoutsegmentierung und OCR via eScriptorium

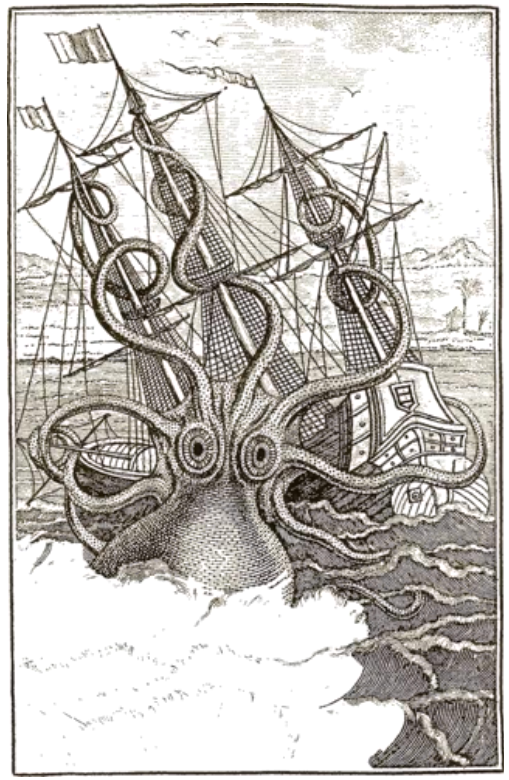
# 2. Layoutsegmentierung und OCR via eScriptorium

**eScriptorium**
Transkriptionssoftware

Open-Source-Plattform zur manuellen oder automatisierten Segmentierung und Texterkennung von historischen Handschriften und Drucken.

| | |
|---|---|
| **Entwickler**: | PSL (Paris) |
| **Erscheinungsjahr**: | 2018 |
| **Aktuelle Version**: | 0.10.5 (2022) |
| **Betriebssystem**: | plattformunabhängig |
| **Programmiersprache**: | Python, JavaScript, Hypertext Markup Language |
| **OCR-Engine**: | Kraken |
| **Lizenz**: | MIT license |
| **Modelle**: | https://ub-backup.bib.uni-mannheim.de/~stweil/eScriptorium/ (Fraktur) |
| **Weitere Infos**: | https://ocr-bw.bib.uni-mannheim.de/tag/escriptorium/ |

# 2. Layoutsegmentierung und OCR via eScriptorium



**Kraken**
OCR-Engine für Layoutssegmentierung und Texterkennung

Kraken ist eine auf historische Dokumente und nicht-lateinische Schriftssysteme optimierte “all-in-one“-OpenSource-OCR-Softwarelösung mit vollständig trainierbarer Layout-Analyse und Texterkennung.

| | |
|---|---|
| **Entwickler:** | Benjamin Kiessling |
| **Erscheinungsjahr:** | 2015 |
| **OCR-Technik:** | bounding box und baseline-based (eScriptorium) |
| **Ausgabeformate:** | ALTO, PageXML, abbyyXML and hOCR |
| **Programmiersprache:** | Python |
| **Modelle:** | https://ub-backup.bib.uni-mannheim.de/~stweil/eScriptorium/ (Fraktur) |
| **Lizenz**: | MIT License |

# 2. Layoutsegmentierung und OCR via eScriptorium

BERD@NFDI

**eScriptorium-Workflow**
Erkennung des Layouts und von Textinhalten

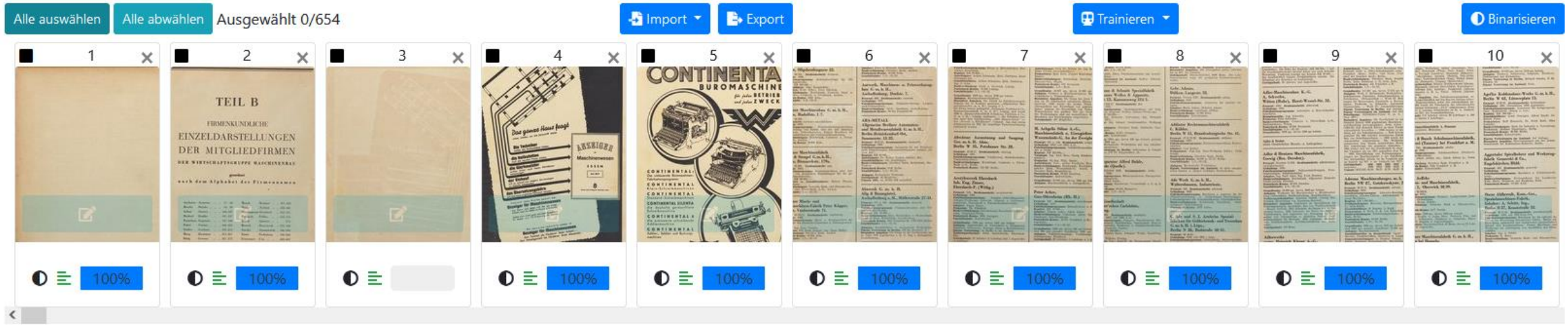
1. Upload der Digitalisate
2. Auswahl eines geeigneten Basismodells für Layoutsegmentierung und Texterkennung
3. Ermittlung von Schwachstellen der Modelle
4. Erstellung von Ground-Truth-Daten
5. Werkspezifisches Nachtraining der Modelle
6. Erkennung von Layout und Textinhalten

Wenn kein Basismodell gefunden werden kann, sollte auf ein anderes System ausgewichen werden oder es  muss ein „from-scratch“ Modell trainiert werden (wesentlich höherer Trainingsaufwand)

# 2. Layoutsegmentierung und OCR via eScriptorium

## Upload der Seiten

Digitalisate können über die GUI oder die API einzeln oder in einem Batch hochgeladen werden




Übersicht über die 654 Digitalisaten in dem neu angelegten Projekt

# Layoutsegmentierung

# 2. Layoutsegmentierung und OCR via eScriptorium

## Auswahl eines geeigneten Basismodells

Die Segmentierung besteht aus 2 Schritten:

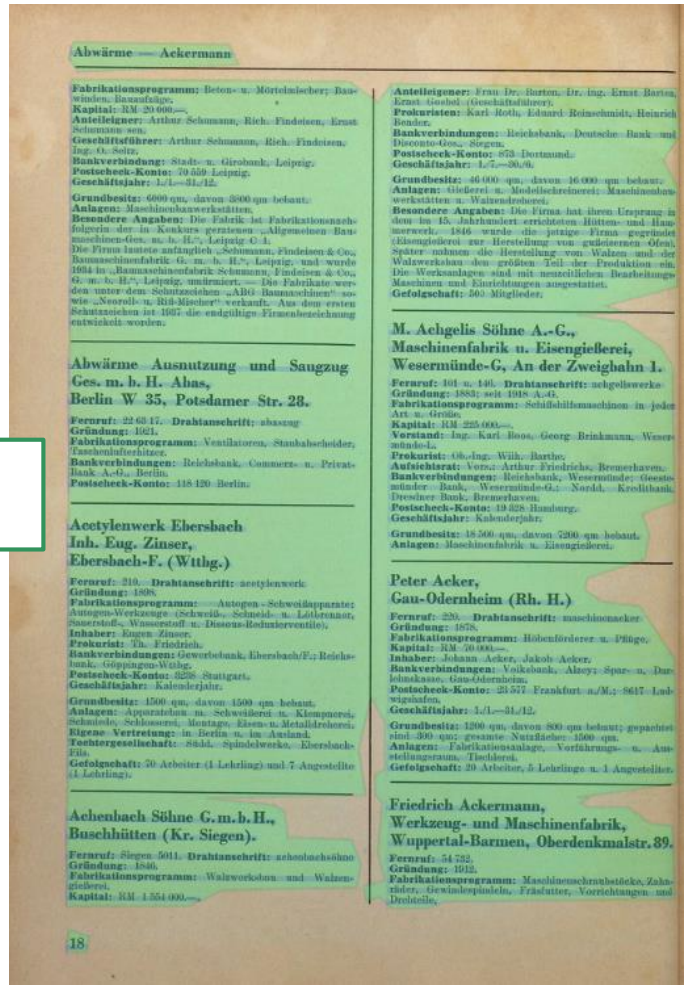
1. Segmentierung der Textregionen
2. Segmentierung der Textzeilen (Baseline und Polygonzug)


**Geeignetes Basismodell:**
**cbad_1800_compensated_50**

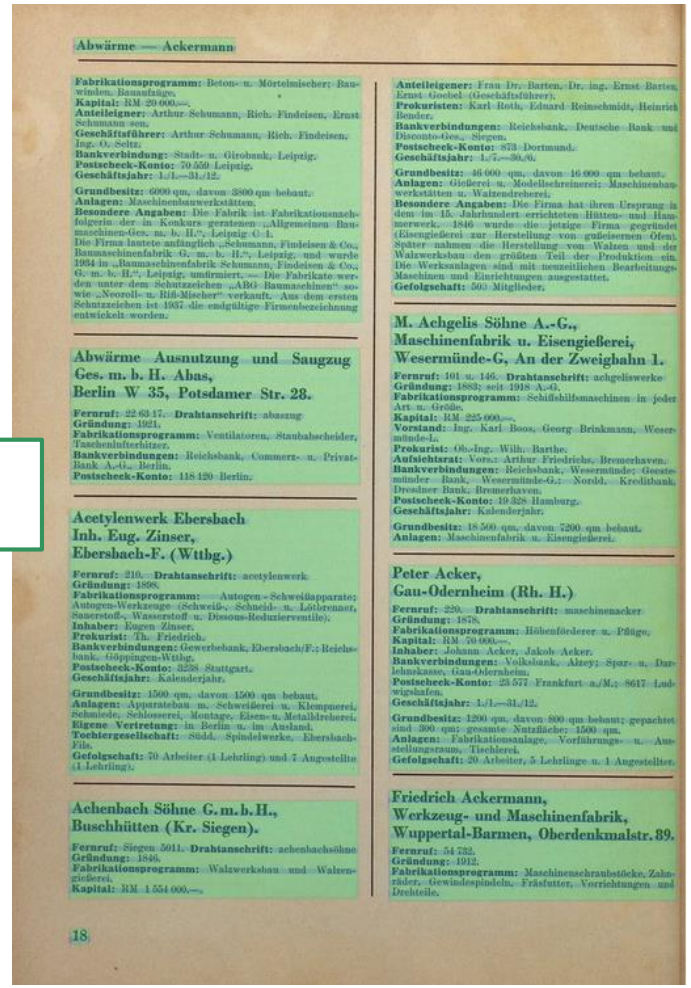
# 2. Layoutsegmentierung und OCR via eScriptorium

## Schwachstelle: Segmentierung der Profile

**Segmentierung mit cbad_1800**

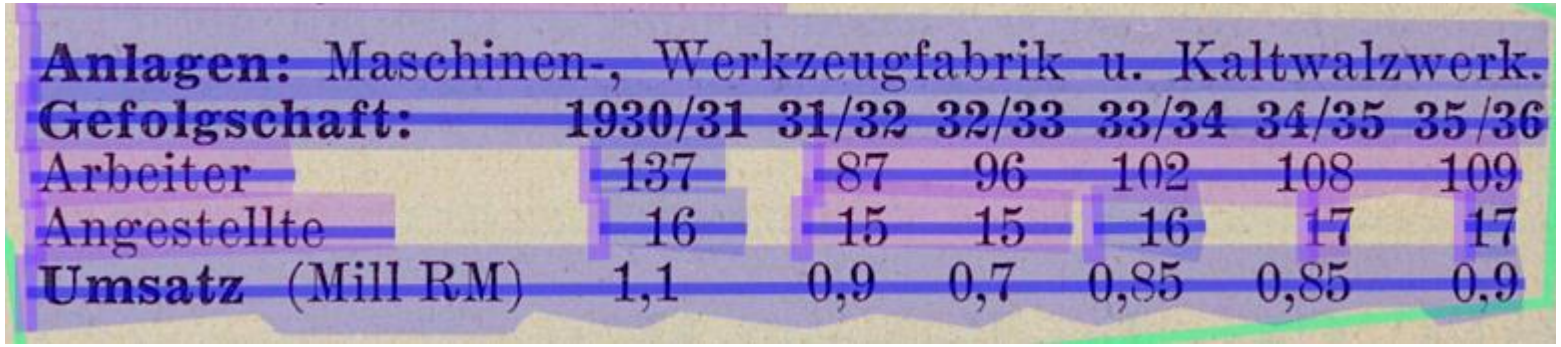



**Optimale Segmentierung**



## Schwachstelle: Textzeilen

Bei der Textzeilenerkennung ist die Baseline nicht immer korrekt und teilweise werden Zeilen in kleine Bereiche unterteilt



| **Anlagen:** Maschinen-, Werkzeugfabrik u. Kaltwalzwerk. | | | | | | |
|---|---|---|---|---|---|---|
| **Gefolgschaft:** | **1930/31** | **31/32** | **32/33** | **33/34** | **34/35** | **35/36** |
| Arbeiter | 137 | 87 | 96 | 102 | 108 | 109 |
| Angestellte | 16 | 15 | 15 | 16 | 17 | 17 |
| **Umsatz** (Mill RM) | 1,1 | 0,9 | 0,7 | 0,85 | 0,85 | 0,9 |

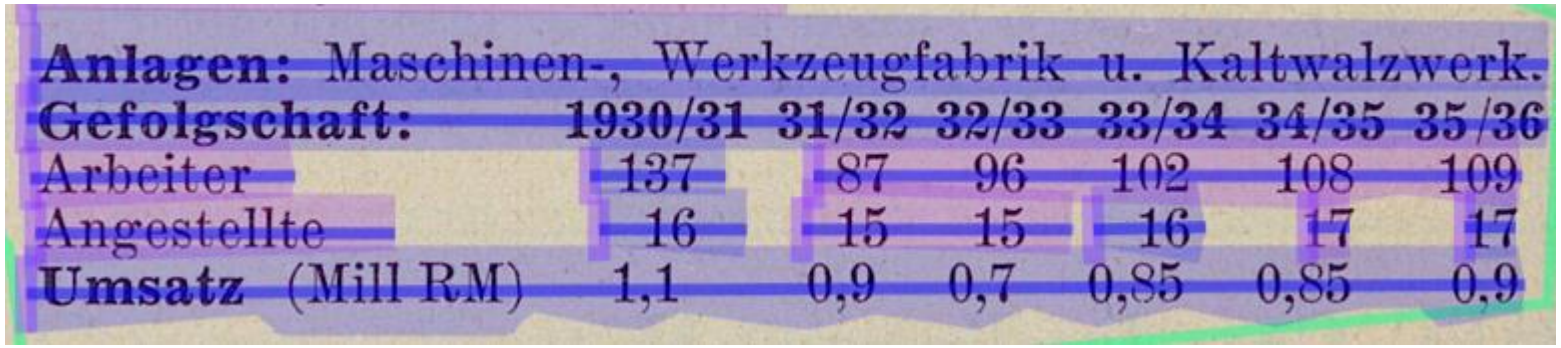
Beispiel: Textzeilensegmentierung einer Tabelle

# 2. Layoutsegmentierung und OCR via eScriptorium

## Werkspezifisches Training

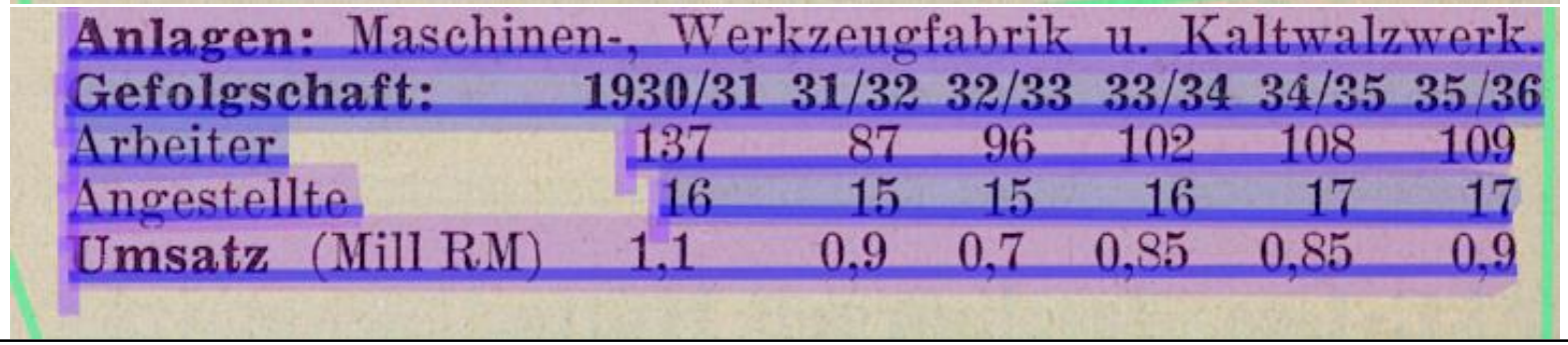
**Ground Truth:** 47 Seiten
**Ergebnisse Textregionen:** Unbefriedigend (Lösung: Verwendung einer anderen Vor- und Nachverarbeitung)
**Ergebnisse Textzeilen:** Deutliche Verbesserung

**cbad_1800**



| Anlagen: Maschinen-, Werkzeugfabrik u. Kaltwalzwerk. | | | | | | |
|---|---|---|---|---|---|---|
| Gefolgschaft: | 1930/31 | 31/32 | 32/33 | 33/34 | 34/35 | 35/36 |
| Arbeiter | 137 | 87 | 96 | 102 | 108 | 109 |
| Angestellte | 16 | 15 | 15 | 16 | 17 | 17 |
| Umsatz (Mill RM) | 1,1 | 0,9 | 0,7 | 0,85 | 0,85 | 0,9 |

**cbad_1800_trained**



| Anlagen: Maschinen-, Werkzeugfabrik u. Kaltwalzwerk. | | | | | | |
|---|---|---|---|---|---|---|
| Gefolgschaft: | 1930/31 | 31/32 | 32/33 | 33/34 | 34/35 | 35/36 |
| Arbeiter | 137 | 87 | 96 | 102 | 108 | 109 |
| Angestellte | 16 | 15 | 15 | 16 | 17 | 17 |
| Umsatz (Mill RM) | 1,1 | 0,9 | 0,7 | 0,85 | 0,85 | 0,9 |

# OCR

# 2. Layoutsegmentierung und OCR via eScriptorium

**Auswahl eines geeigneten Basismodells:**

Ein geeignetes Basismodell sollte

**1)** den Zeichenvorrat des Quellmaterials (möglichst) abbilden

**2)** bereits eine gute Erkennungsrate auf das Quellmaterial garantieren

**Geeignetes OCR-Modell: digitue**

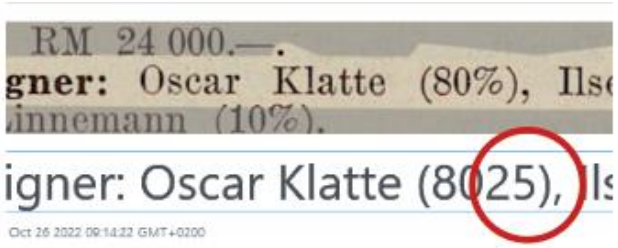
**AustrianNewspapers**
https://github.com/UB-Mannheim/AustrianNewspapers
↓
**luther** (aus digi-gt)
https://github.com/UB-Mannheim/digi-gt
↓
**reichsanzeiger**
https://github.com/UB-Mannheim/reichsanzeiger-gt
↓
**digitue**
https://github.com/UB-Mannheim/digitue-gt

# 2. Layoutsegmentierung und OCR via eScriptorium

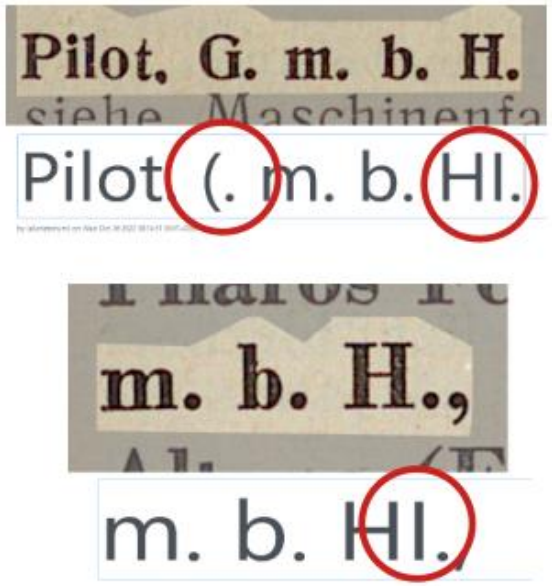
## Schwachstellen in der Texterkennung

Das Modell liefert bereits sehr gute Ergebnisse mit Texterkennungsraten > 98 %
Einige spezifische Fehler lassen sich identifizieren:

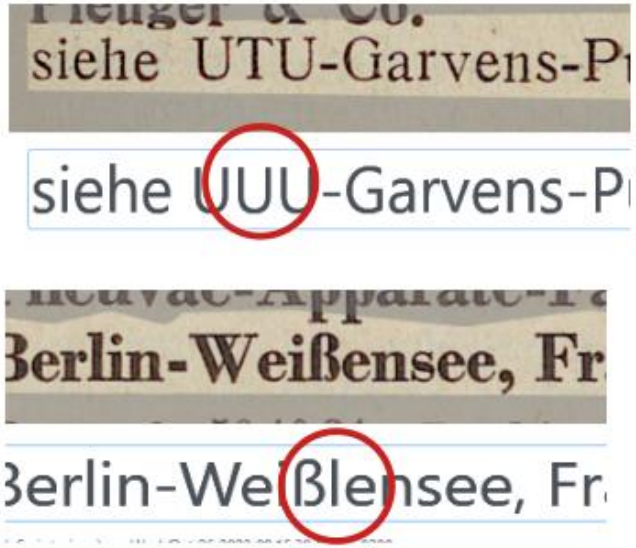
**Prozentzeichen**




**Zeichenkette „G.m.b.H.“**



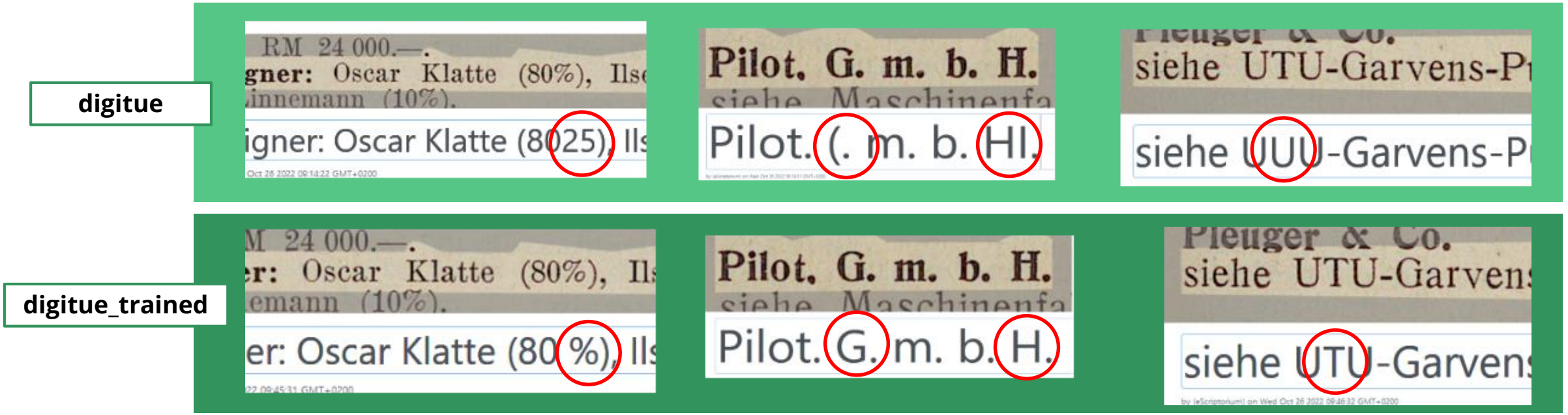

**Seltene Zeichen / Zeichenkombinationen**



# 2. Layoutsegmentierung und OCR via eScriptorium

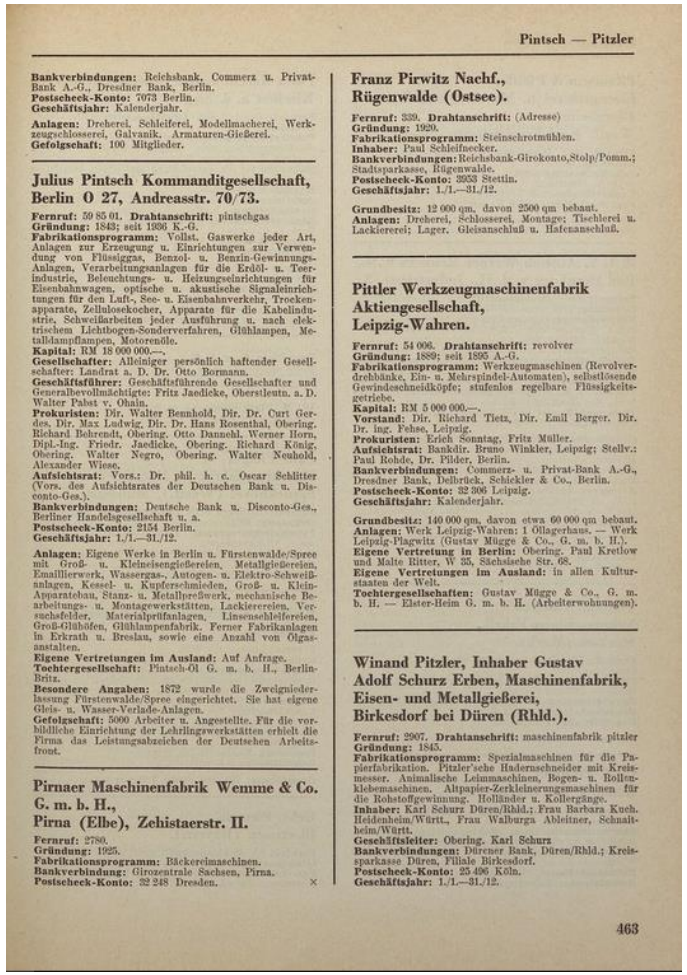
## Werkspezifisches Training

**Ground Truth**: 26 Seiten
**Ergebnisse Texterkennung**: Die Erkennungsrate konnte auf > 99.85 % angehoben werden



# 3. Datenextraktion und –strukturierung via Python

# 3. Datenextraktion und –strukturierung via Python: Dateninput



Pintsch — Pitzler

**Bankverbindungen:** Reichsbank, Commerz u. Privat-Bank A.-G., Dresdner Bank, Berlin.
**Postscheck-Konto:** 7073 Berlin.
**Geschäftsjahr:** Kalenderjahr.
**Anlagen:** Dreherei, Schleiferei, Modellmacherei, Werkzeugschlosserei, Galvanik, Armaturen-Gießerei.
**Gefolgschaft:** 100 Mitglieder.

**Julius Pintsch Kommanditgesellschaft, Berlin O 27, Andreasstr. 70/73.**

**Fernruf:** 59 85 01. **Drahtanschrift:** pintschgas
**Gründung:** 1843; seit 1936 K.-G.
**Fabrikationsprogramm:** Vollst. Gaswerke jeder Art, Anlagen zur Erzeugung u. Einrichtungen zur Verwendung von Flüssiggas, Benzol- u. Benzin-Gewinnungs-Anlagen, Verarbeitungsanlagen für die Erdöl- u. Teerindustrie, Beleuchtungs- u. Heizungseinrichtungen für Eisenbahnwagen, optische u. akustische Signaleinrichtungen für den Luft-, See- u. Eisenbahnverkehr, Trockenapparate, Zellulosekocher, Apparate für die Kabelindustrie, Schweißarbeiten jeder Ausführung u. nach elektrischem Lichtbogen-Sonderverfahren, Glühlampen, Metalldampflampen, Motorenöle.
**Kapital:** RM 18 000 000.—.
**Gesellschafter:** Alleiniger persönlich haftender Gesellschafter: Landrat a. D. Dr. Otto Bormann.
**Geschäftsführer:** Geschäftsführende Gesellschafter und Generalbevollmächtigte: Fritz Jaedicke, Oberstleutn. a. D. Walter Pabst v. Ohain.
**Prokuristen:** Dir. Walter Bennhold, Dir. Dr. Curt Gerdes, Dir. Max Ludwig, Dir. Dr. Hans Rosenthal, Obering. Richard Behrendt, Obering. Otto Dannehl, Werner Horn, Dipl.-Ing. Friedr. Jaedicke, Obering. Richard König, Obering. Walter Negro, Obering. Walter Neuhold, Alexander Wiese.
**Aufsichtsrat:** Vors.: Dr. phil. h. c. Oscar Schlitter (Vors. des Aufsichtsrates der Deutschen Bank u. Disconto-Ges.).
**Bankverbindungen:** Deutsche Bank u. Disconto-Ges., Berliner Handelsgesellschaft u. a.
**Postscheck-Konto:** 2154 Berlin.
**Geschäftsjahr:** 1./1.—31./12.
**Anlagen:** Eigene Werke in Berlin u. Fürstenwalde/Spree mit Groß- u. Kleinesengießereien, Metallgießereien, Emaillierwerk, Wassergas-, Autogen- u. Elektro-Schweißanlagen, Kessel- u. Kupferschmieden, Groß- u. Klein-Apparatebau, Stanz- u. Metallpreßwerk, mechanische Bearbeitungs- u. Montagewerkstätten, Lackierereien, Versuchsfelder, Materialprüfanlagen, Linsenschleifereien, Groß-Glühöfen, Glühlampenfabrik. Ferner Fabrikanlagen in Erkrath u. Breslau, sowie eine Anzahl von Ölgasanstalten.
**Eigene Vertretungen im Ausland:** Auf Anfrage.
**Tochtergesellschaft:** Pintsch-Öl G. m. b. H., Berlin-Britz.
**Besondere Angaben:** 1872 wurde die Zweigniederlassung Fürstenwalde/Spree eingerichtet. Sie hat eigene Gleis- u. Wasser-Verlade-Anlagen.
**Gefolgschaft:** 5000 Arbeiter u. Angestellte. Für die vorbildliche Einrichtung der Lehrlingswerkstätten erhielt die Firma das Leistungsabzeichen der Deutschen Arbeitsfront.

**Pirnaer Maschinenfabrik Wemme & Co. G. m. b. H., Pirna (Elbe), Zehistaerstr. II.**

**Fernruf:** 2780.
**Gründung:** 1925.
**Fabrikationsprogramm:** Bäckereimaschinen.
**Bankverbindung:** Girozentrale Sachsen, Pirna.
**Postscheck-Konto:** 32 248 Dresden.

**Franz Pirwitz Nachf., Rügenwalde (Ostsee).**

**Fernruf:** 339. **Drahtanschrift:** (Adresse)
**Gründung:** 1929.
**Fabrikationsprogramm:** Steinschrotmühlen.
**Inhaber:** Paul Schleifnecker.
**Bankverbindungen:** Reichsbank-Girokonto, Stolp/Pomm.; Stadtsparkasse, Rügenwalde.
**Postscheck-Konto:** 3953 Stettin.
**Geschäftsjahr:** 1./1.—31./12.

**Grundbesitz:** 12 000 qm, davon 2500 qm bebaut.
**Anlagen:** Dreherei, Schlosserei, Montage; Tischlerei u. Lackiererei; Lager. Gleisanschluß u. Hafenanschluß.

**Pittler Werkzeugmaschinenfabrik Aktiengesellschaft, Leipzig-Wahren.**

**Fernruf:** 54 006. **Drahtanschrift:** revolver
**Gründung:** 1889; seit 1895 A.-G.
**Fabrikationsprogramm:** Werkzeugmaschinen (Revolverdrehbänke, Ein- u. Mehrspindel-Automaten), selbstlösende Gewindeschneidköpfe; stufenlos regelbare Flüssigkeitsgetriebe.
**Kapital:** RM 5 000 000.—.
**Vorstand:** Dir. Richard Tietz, Dir. Emil Berger. Dir. Dr. ing. Fehse, Leipzig.
**Prokuristen:** Erich Sonntag, Fritz Müller.
**Aufsichtsrat:** Bankdir. Bruno Winkler, Leipzig; Stellv.: Paul Rohde, Dr. Pilder, Berlin.
**Bankverbindungen:** Commerz- u. Privat-Bank A.-G., Dresdner Bank, Delbrück, Schickler & Co., Berlin.
**Postscheck-Konto:** 32 306 Leipzig.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 140 000 qm, davon etwa 60 000 qm bebaut.
**Anlagen:** Werk Leipzig-Wahren: 1 Ollagerhaus. — Werk Leipzig-Plagwitz (Gustav Mügge & Co., G. m. b. H.).
**Eigene Vertretung in Berlin:** Obering. Paul Krellow und Malte Ritter, W 35, Sächsische Str. 68.
**Eigene Vertretungen im Ausland:** in allen Kulturstaaten der Welt.
**Tochtergesellschaften:** Gustav Mügge & Co., G. m. b. H. — Elster-Heim G. m. b. H. (Arbeiterwohnungen).

**Winand Pitzler, Inhaber Gustav Adolf Schurz Erben, Maschinenfabrik, Eisen- und Metallgießerei, Birkesdorf bei Düren (Rhld.).**

**Fernruf:** 2907. **Drahtanschrift:** maschinenfabrik pitzler
**Gründung:** 1845.
**Fabrikationsprogramm:** Spezialmaschinen für die Papierfabrikation. Pitzler'sche Hadernschneider mit Kreismesser. Animalische Leimmaschinen, Bogen- u. Rollenklebemaschinen. Altpapier-Zerkleinerungsmaschinen für die Rohstoffgewinnung. Holländer u. Kollergänge.
**Inhaber:** Karl Schurz Düren/Rhld.; Frau Barbara Kuch, Heidenheim/Württ., Frau Walburga Ableitner, Schnaitheim/Württ.
**Geschäftsleiter:** Obering. Karl Schurz
**Bankverbindungen:** Dürener Bank, Düren/Rhld.; Kreissparkasse Düren, Filiale Birkesdorf.
**Postscheck-Konto:** 25 496 Köln.
**Geschäftsjahr:** 1./1.—31./12.

463

OCR

```
<?xml version="1.0" encoding="UTF-8"  standalone="yes"?>
<PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15" xmlns
xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15
http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15/pagecontent.xsd">
  <Metadata>
    <Creator>escriptorium</Creator>
    <Created>2022-09-19T12:10:54.541269+00:00</Created>
        <LastChange>2022-09-19T12:10:54.541290+00:00</LastChange>

  </Metadata>
  <Page imageFilename="maschinenindustrie_1937_0452.jpg" imageWi
    <TextRegion id="r_2_1"  >
      <Coords points="127,354 1290,354 1290,685 127,685"/>

      <TextLine id="eSc_line_835f5cfa" >
        <Coords points="1269,350 1148,348 1132,348 1109,348 1086
880,348 857,348 834,348 825,350 818,350 815,350 811,348 811,348
673,348 650,348 632,348 627,348 605,348 582,348 559,348 536,348
307,348 284,348 261,348 238,348 215,348 192,348 169,348 146,348
        <Baseline points="126,382 1292,387"/>
        <TextEquiv>
          <Unicode>Bankverbindungen: Reichsbank, Commerz u. Priv
        </TextEquiv>
      </TextLine>


      <TextLine id="eSc_line_eff01337" >
        <Coords points="128,426 126,449 877,449 880,430 877,396 128,389 128,426"/>
        <Baseline points="128,426 880,430"/>
        <TextEquiv>
          <Unicode>Bank A.-G., Dresdner Bank, Berlin.</Unicode>
        </TextEquiv>
      </TextLine>
```

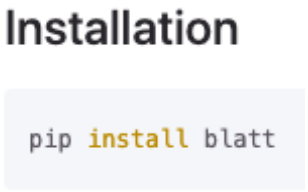
**Reicht PAGE-XML aus, um Forschungsfrage zu beantworten?**

**JPG** **PAGE-XML**

# 3. Datenextraktion und –strukturierung via Python: blatt


**_Blatt_ Funktionalitäten:**

1. liest PAGE-XML
2. entfernt Bindestriche
3. konvertiert PAGE-XML in TXT- und TSV-Formate
4. bietet ein Modul zur Datenstrukturierung




Open Source + CLI


https://pypi.org/project/blatt

https://github.com/UB-Mannheim/blatt

# 3. Datenextraktion und –strukturierung via Python: Segmentierung

OCR-Segmentierung der Textregionen mit Problemen

Abwärme — Ackermann

**Fabrikationsprogramm:** Beton- u. Mörtelmischer; Bauwinden, Bauaufzüge.
**Kapital:** RM 20 000.—.
**Anteileigner:** Arthur Schumann, Rich. Findeisen, Ernst Schumann sen.
**Geschäftsführer:** Arthur Schumann, Rich. Findeisen, Ing. O. Seltz.
**Bankverbindung:** Stadt- u. Girobank, Leipzig.
**Postscheck-Konto:** 70 559 Leipzig.
**Geschäftsjahr:** 1./1.—31./12.

**Grundbesitz:** 6000 qm, davon 3800 qm bebaut.
**Anlagen:** Maschinenbauwerkstätten.
**Besondere Angaben:** Die Fabrik ist Fabrikationsnachfolgerin der in Konkurs geratenen „Allgemeinen Baumaschinen-Ges. m. b. H.", Leipzig C 1.
Die Firma lautete anfänglich „Schumann, Findeisen & Co., Baumaschinenfabrik G. m. b. H.", Leipzig, und wurde 1934 in „Baumaschinenfabrik Schumann, Findeisen & Co., G. m. b. H.", Leipzig, umfirmiert. — Die Fabrikate werden unter dem Schutzzeichen „ABG Baumaschinen" sowie „Neoroll- u. Riffl-Mischer" verkauft. Aus dem ersten Schutzzeichen ist 1937 die endgültige Firmenbezeichnung

**Anteileigener:** Frau Dr. Barten, Dr. ing. Ernst Barten, Ernst Goebel (Geschäftsführer).
**Prokuristen:** Karl Roth, Eduard Reinschmidt, Heinrich Bender.
**Bankverbindungen:** Reichsbank, Deutsche Bank und Disconto-Ges., Siegen.
**Postscheck-Konto:** 873 Dortmund.
**Geschäftsjahr:** 1./7.—30./6.

**Grundbesitz:** 46 000 qm, davon 16 000 qm bebaut.
**Anlagen:** Gießerei u. Modellschreinerei; Maschinenbauwerkstätten u. Walzendreherei.
**Besondere Angaben:** Die Firma hat ihren Ursprung in dem im 15. Jahrhundert errichteten Hütten- und Hammerwerk. 1846 wurde die jetzige Firma gegründet (Eisengießerei zur Herstellung von gußeisernen Öfen). Später nahmen die Herstellung von Walzen und der Walzwerksbau den größten Teil der Produktion ein. Die Werksanlagen sind mit neuzeitlichen Bearbeitungs-Maschinen und Einrichtungen ausgestattet.
**Gefolgschaft:** 500 Mitglieder.

**blatt**

1. sortiert Textzeilen
2. führt eine geometrische Segmentierung auf Basis der Koordinaten durch
3. Datenbereinigung
4. führt Segmente nachfolgender Spalten und Seiten zusammen

# 3. Datenextraktion und –strukturierung via Python: Segmentierung

## JPG

**Piccolo-Automaten G. m. b. H.,**
**Berlin W 35, Kurfürstenstraße 146.**

**Fernruf:** 21 20 95. **Drahtanschrift:** piccoloautomat
**Gründung:** 1932.
**Fabrikationsprogramm:** Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstechapparate).
**Kapital:** RM 20 000.—.
**Anteileigner:** Dr. Richard Schönthal, Wien.
**Geschäftsführer:** Erwin Hantsch, Bln.-Steglitz.
**Bankverbindung:** Deutsche Bank u. Disconto - Ges., Berlin.
**Postscheck-Konto:** 8012 Berlin.
**Geschäftsjahr:** 1./1.—31./12.

**Anlagen:** Mechanische Werkstätten.
**Gefolgschaft:** 13 Arbeiter u. 6 Angestellte.

Transformation via blatt

**Ansatz:** Doppelpunkt als Separator nutzen

**Problem:** inkonsistente Attributbezeichnung, Druckfehler, OCR Fehler

## Strukturierte Forschungsdaten

| Firma | Piccolo-Automaten G.m.b.H., Berlin W 35, Kurfürstenstraße 146 |
|---|---|
| Rechtsform | G.m.b.H. |
| Fernruf | 212095 |
| Drahtanschrift | piccoloautomat |
| Gründung | 1932 |
| Produkt | Schokoloaden-Verkaufsapparate (Tischautomaten, Kugelstechapparate) |

...

*e.g.: OCR problems:*
*{'Postscheck-Konto ', 'Postseheck-Konto','PostscheckKonto', 'Postscheck-Konnto', 'Fostscheck-Konto', Postcheck-Konto', 'Postscheek-Konto', 'Potscheck-Konto','Postscbeck-Konto', 'Postscheckkonto', 'Postschek-Konto','Postscheck-Konten', 'Ponstscheck-Konto', 'Postscheck-Konto'}*

## Sortierung und Gruppierung

***Postscheck-Konto***

{‘Postscheck-Konto', 'Postseheck-Konto','PostscheckKonto', 'Postscheck-Konnto', ‘Fostscheck-Konto', Postcheck-Konto', ‘Postscheek-Konto‘, 'Potscheck-Konto','Postscbeck-Konto', ‘Postscheckkonto’, 'Postschek-Konto‘, 'Postscheck-Konten', 'Ponstscheck-Konto', 'Postscheck-Konto‘}

***Geschäftsjahr***

{'Geschäftjahr', 'Geschäftsjahr', 'Gescbäftsjahr', 'Geschätfsjahr', '.Geschäftsjahr‘}

**Fabrikationsprogramm**

{'Fabfikationsprogramm', 'Fabrikationprogramm','Fabrikationsproramm', 'Fabrikationsprogramm:', 'Fabrikstionsprogramm', 'Fabrkkationsprogramm','Fabrikationsprogramm'}

## Extraktion *Drahtanschrift* aus *Fernruf*:

*Fernruf: 212095. Drahtanschrift: piccoloautomat*

| Fernruf | 212095 |
|---|---|
| Drahtanschrift | piccoloautomat |

## Extraktion der Rechtsformen:

{‘m. b. H.', 'G. m. b. H.', 'Gesellschaft m. b. H.', 'G.m.b. H.', 'Ges. m.b. H.', 'GmbH', 'G. m. b.H.', 'mit beschr. Haftung', 'Ges.m.b.H.', 'G.m.b.H.', 'GmbH.', ‘GmbH‘}

{'A.-G.', 'Aktien-Gesellschaft', 'Actien-Gesellschaft', 'Aktiengesellschaft', 'Akt.-Ges.', 'Aktien-Gesellsch.', 'A. G.‘}

{'K.-G.', 'Kom.-Ges.', 'Komm.-Ges.', 'KG.', 'Kommanditgesellschaft', 'Kom. Ges.', 'Komm.-Ges.'}

# 3. Datenextraktion und –strukturierung via Python: CSV und XLSX

BERD@NFDI

## Strukturierte Forschungsdaten

| Firma | Piccolo-Automaten G.m.b.H., Berlin W 35, Kurfürstenstraße 146 |
|---|---|
| Rechtsform | G.m.b.H. |
| Fernruf | 212095 |
| Drahtanschrift | piccoloautomat |
| Gründung | 1932 |
| Produkt | Schokoloaden-Verkaufsapparate (Tischautomaten, Kugelstechapparate) |

...

## CSV und XLSX

| | Company | RAW_TEXT | W_TEXT_1L | LE_SEGMEN | FABRIKATIONSPROGRAMM | POSTSCHECK-KONTO | FERNRUF | DRAHTANSCHRIFT | BANKVERBINDUNGEN | ANLAGEN |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Aachener Kratz | Aachener K<br>Cassalette &<br>Aachen, Oli<br>Fernruf: 34<br>Gründung:<br>Fabrikation<br>Verwendun<br>Kapital: RM<br>Geschäftsfü<br>Prokuristen<br>Bankverbin<br>Disconto-G<br>Postscheck-<br>Geschäftsja | Aachener K | ./MI1937/n | Kratzenbeschläge für alle Ver | 2952 Köln | 34 041 | kratzena | Reichsbank, Deutsche Bank u. Disconto-Ges | |
| 1 | Aachener Mas | Aachener M<br>Aachen, Ru<br>Fernruf: 25<br>Drahtansch<br>Gründung:<br>Fabrikation<br>zur Herstell<br>Drahtartike<br>die Kratzen<br>Bankverbin<br>bank A.-G.,<br>Postscheck- | Aachener M | ./MI1937/n | Drahtbearbeitungsmaschinen | 16 987 Köln | 25 205 | aachener maschinenbau | Reichsbank, Commerz- u. Privatbank A.-G., | |
| 2 | | Aachener M<br>Rothe & Ste<br>Aachen, Bis<br>Fernruf: 25<br>Gründung:<br>Fabrikation<br>Lieferung v<br>samte Kratz<br>Kapital: RM<br>Anteileigne<br>Peter Rump | | | | | | | | |

# 4. Fazit

# 4. Fazit

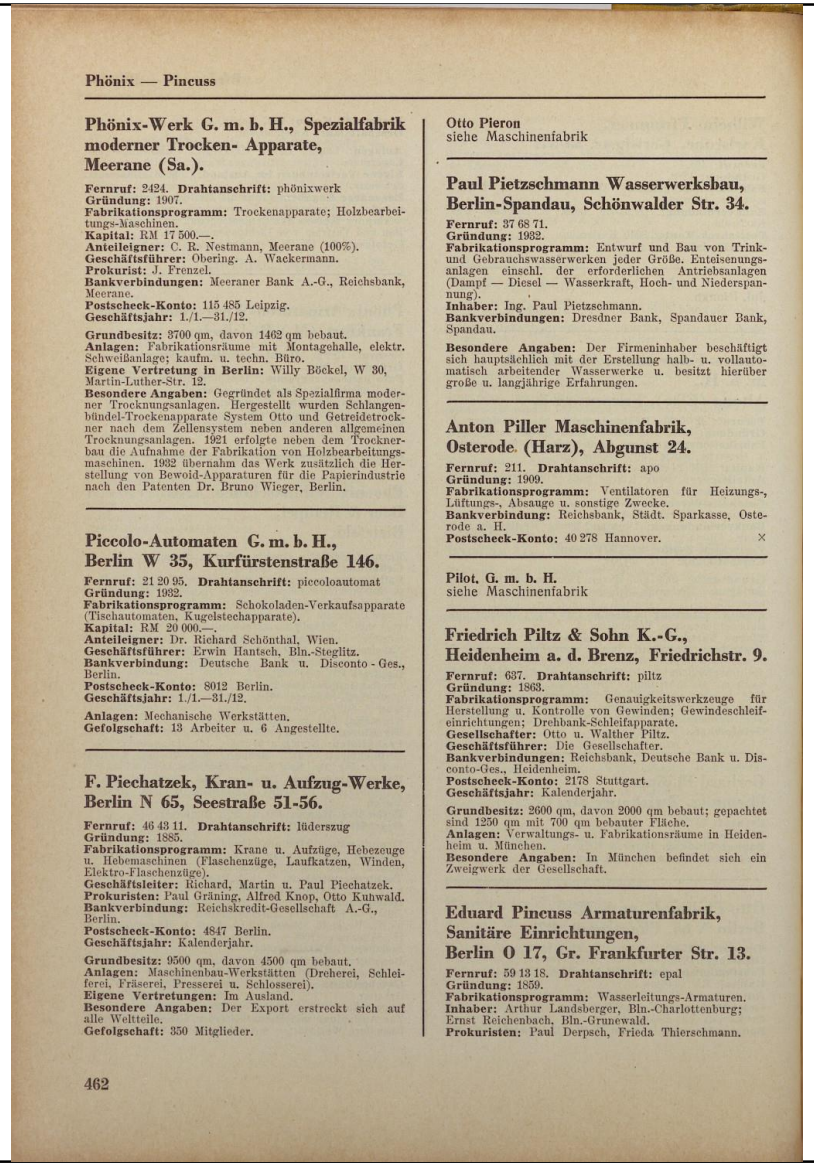
## Ergebnisse:

- Sehr gute Texterkennungsgenauigkeit (> 99,85%) durch werkspezifisches Training

- eScriptorium eignet sich sehr gut als Plattform für Ground-Truth-Produktion und Training:
  - schnelle, benutzerfreundliche, plattformunabhängige Erstellung von Transkriptionen
  - benutzerfreundlicher Trainingsprozess (GUI)

- Datenstrukturierung:
  - keine All-in-One-Lösung → wir haben das Open-Source-Tool *blatt* entwickelt → es kann in ähnlichen Projekten wiederverwendet werden
  - wir haben alle verfügbaren Daten strukturiert (nicht nur Rechtsformen) → die Daten können für andere Forschungsfragen wiederverwendet werden

- Services der UB Mannheim können flexibel und kooperativ in Projekten mit anderen Stakeholdern der Universität genutzt werden



Phönix — Pincuss

**Phönix-Werk G. m. b. H., Spezialfabrik moderner Trocken- Apparate, Meerane (Sa.).**

**Fernruf:** 2424. **Drahtanschrift:** phönixwerk
**Gründung:** 1907.
**Fabrikationsprogramm:** Trockenapparate; Holzbearbeitungs-Maschinen.
**Kapital:** RM 17 500.—.
**Anteileigner:** C. R. Nestmann, Meerane (100%).
**Geschäftsführer:** Obering. A. Wackermann.
**Prokurist:** J. Frenzel.
**Bankverbindungen:** Meeraner Bank A.-G., Reichsbank, Meerane.
**Postscheck-Konto:** 115 485 Leipzig.
**Geschäftsjahr:** 1./1.—31./12.

**Grundbesitz:** 3700 qm, davon 1462 qm bebaut.
**Anlagen:** Fabrikationsräume mit Montagehalle, elektr. Schweißanlage; kaufm. u. techn. Büro.
**Eigene Vertretung in Berlin:** Willy Böckel, W 30, Martin-Luther-Str. 12.
**Besondere Angaben:** Gegründet als Spezialfirma moderner Trocknungsanlagen. Hergestellt wurden Schlangenbündel-Trockenapparate System Otto und Getreidetrockner nach dem Zellensystem neben anderen allgemeinen Trocknungsanlagen. 1921 erfolgte neben dem Trocknerbau die Aufnahme der Fabrikation von Holzbearbeitungsmaschinen. 1932 übernahm das Werk zusätzlich die Herstellung von Bewoid-Apparaturen für die Papierindustrie nach den Patenten Dr. Bruno Wieger, Berlin.

**Piccolo-Automaten G. m. b. H., Berlin W 35, Kurfürstenstraße 146.**

**Fernruf:** 21 20 95. **Drahtanschrift:** piccoloautomat
**Gründung:** 1932.
**Fabrikationsprogramm:** Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstechapparate).
**Kapital:** RM 20 000.—.
**Anteileigner:** Dr. Richard Schönthal, Wien.
**Geschäftsführer:** Erwin Hantsch, Bln.-Steglitz.
**Bankverbindung:** Deutsche Bank u. Disconto - Ges., Berlin.
**Postscheck-Konto:** 8012 Berlin.
**Geschäftsjahr:** 1./1.—31./12.

**Anlagen:** Mechanische Werkstätten.
**Gefolgschaft:** 13 Arbeiter u. 6 Angestellte.

**F. Piechatzek, Kran- u. Aufzug-Werke, Berlin N 65, Seestraße 51-56.**

**Fernruf:** 46 43 11. **Drahtanschrift:** lüderszug
**Gründung:** 1885.
**Fabrikationsprogramm:** Krane u. Aufzüge, Hebezeuge u. Hebemaschinen (Flaschenzüge, Laufkatzen, Winden, Elektro-Flaschenzüge).
**Geschäftsleiter:** Richard, Martin u. Paul Piechatzek.
**Prokuristen:** Paul Gräning, Alfred Knop, Otto Kuhwald.
**Bankverbindung:** Reichskredit-Gesellschaft A.-G., Berlin.
**Postscheck-Konto:** 4847 Berlin.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 9500 qm, davon 4500 qm bebaut.
**Anlagen:** Maschinenbau-Werkstätten (Dreherei, Schleiferei, Fräserei, Presserei u. Schlosserei).
**Eigene Vertretungen:** Im Ausland.
**Besondere Angaben:** Der Export erstreckt sich auf alle Weltteile.
**Gefolgschaft:** 350 Mitglieder.

Otto Pieron
siehe Maschinenfabrik

**Paul Pietzschmann Wasserwerksbau, Berlin-Spandau, Schönwalder Str. 34.**

**Fernruf:** 37 68 71.
**Gründung:** 1932.
**Fabrikationsprogramm:** Entwurf und Bau von Trink- und Gebrauchswasserwerken jeder Größe. Enteisenungsanlagen einschl. der erforderlichen Antriebsanlagen (Dampf — Diesel — Wasserkraft, Hoch- und Niederspannung).
**Inhaber:** Ing. Paul Pietzschmann.
**Bankverbindungen:** Dresdner Bank, Spandauer Bank, Spandau.

**Besondere Angaben:** Der Firmeninhaber beschäftigt sich hauptsächlich mit der Erstellung halb- u. vollautomatisch arbeitender Wasserwerke u. besitzt hierüber große u. langjährige Erfahrungen.

**Anton Piller Maschinenfabrik, Osterode. (Harz), Abgunst 24.**

**Fernruf:** 211. **Drahtanschrift:** apo
**Gründung:** 1909.
**Fabrikationsprogramm:** Ventilatoren für Heizungs-, Lüftungs-, Absauge u. sonstige Zwecke.
**Bankverbindung:** Reichsbank, Städt. Sparkasse, Osterode a. H.
**Postscheck-Konto:** 40 278 Hannover.

Pilot, G. m. b. H.
siehe Maschinenfabrik

**Friedrich Piltz & Sohn K.-G., Heidenheim a. d. Brenz, Friedrichstr. 9.**

**Fernruf:** 637. **Drahtanschrift:** piltz
**Gründung:** 1863.
**Fabrikationsprogramm:** Genauigkeitswerkzeuge für Herstellung u. Kontrolle von Gewinden; Gewindeschleifeinrichtungen; Drehbank-Schleifapparate.
**Gesellschafter:** Otto u. Walther Piltz.
**Geschäftsführer:** Die Gesellschafter.
**Bankverbindungen:** Reichsbank, Deutsche Bank u. Disconto-Ges., Heidenheim.
**Postscheck-Konto:** 2178 Stuttgart.
**Geschäftsjahr:** Kalenderjahr.

**Grundbesitz:** 2600 qm, davon 2000 qm bebaut; gepachtet sind 1250 qm mit 700 qm bebauter Fläche.
**Anlagen:** Verwaltungs- u. Fabrikationsräume in Heidenheim u. München.
**Besondere Angaben:** In München befindet sich ein Zweigwerk der Gesellschaft.

**Eduard Pincuss Armaturenfabrik, Sanitäre Einrichtungen, Berlin O 17, Gr. Frankfurter Str. 13.**

**Fernruf:** 59 13 18. **Drahtanschrift:** epal
**Gründung:** 1859.
**Fabrikationsprogramm:** Wasserleitungs-Armaturen.
**Inhaber:** Arthur Landsberger, Bln.-Charlottenburg; Ernst Reichenbach, Bln.-Grunewald.
**Prokuristen:** Paul Derpsch, Frieda Thierschmann.

462

# 4. Fazit

**Zeitaufwand:**

- **Insgesamt**: ca. 2 Arbeitswochen (verteilt auf 2 Monate)
- **Digitalisierung**: 1 Arbeitstag
- **OCR**: 1 Arbeitswoche:
  - Auswertung vorhandener OCR-Modelle auf dem Material
  - Entwurf eines OCR-Workflows (Layoutsegmentierung + OCR)
  - Ground-Truth-Produktion und anschließendes Training (Layoutsegmentierung + OCR)
  - Endgültige OCR für alle Seiten
- **Strukturierung der Daten**: 1 Arbeitswoche:
  - Segmentierung
  - Nachbearbeitung
  - Erweiterung *blatt*

## Team:

- **6 Projektbeteiligte**
  - Prof. Jochen Streb (**Professur für Wirtschaftsgeschichte @ Uni Mannheim**)
  - 1 Hilfskraft für Ground-Truth-Produktion und QA (**Professur für Wirtschaftsgeschichte**)
  - 1 Projektleitung Digitalisierung (**UB Mannheim**)
  - 1 Projektkoordinator OCR (**UB Mannheim**)
  - 1 Entwickler OCR (**UB Mannheim**)
  - 1 Entwickler Datenstrukturierung (**BERD@NFDI, UB Mannheim**)

Feedback, Fragen?

Jan Kamlah (Development): jan.kamlah@uni-mannheim.de
Renat Shigapov (Development): : renat.shigapov@uni-mannheim.de
Thomas Schmidt (Project coordination): thomas.schmidt@uni-mannheim.de

Danke!

https://github.com/UB-Mannheim/Maschinen-Industrie

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 460037581