

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Extracting research data from historical documents with eScriptorium and Python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
University Library Mannheim

24.11.2022

- 1. Introduction & Digitalization**
- 2. Layout segmentation and OCR via eScriptorium**
- 3. Extraction and structuring via Python**
- 4. Summary**

1. Introduction & Digitalization

Alle Inhalte dieser Präsentation stehen unter der [Lizenz Creative Commons BY 4.0 International](https://creativecommons.org/licenses/by/4.0/), sofern nicht anders angegeben.



1. Introduction & Digitalization

- Request for extraction of structured research data from "**Die Maschinen-Industrie im Deutschen Reich von 1937**" by Prof. Jochen Streb (Chair of Economic History @ University of Mannheim)
- Cooperation project between BERD@NFDI and OCR-D module project "work-specific training" @ UB Mannheim
- Digitization (654 pages) by 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.—
Anteilseigner: G. E. 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 1403 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 Schlangensammel-Trockenapparate System Otto und Getreidetrockner nach dem Zellen-system neben anderen allgemeinen Trocknungsanlagen. 1921 erfolgte neben dem Trocknerbau die Aufnahme der Fabrikation von Holzbearbeitungsmaschinen. 1932 übernahm das Werk zusätzlich die Herstellung von Bewoilt-Apparaturen für die Papierindustrie nach den Patenten Dr. Bruno Wiegler, Berlin.

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

Fernruf: 21 20 95. **Drahtanschrift:** piccoloautomat
Gründung: 1932.
Fabrikationsprogramm: Schokoladen-Verkaufsapparate (Tischautomaten, Kugelstichapparate).
Kapital: RM 20 000.—
Anteilseigner: 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. Hebe-maschinen (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, Fräseerei, 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: 1903.
Fabrikationsprogramm: Entwurf und Bau von Trink- und Gebrauchswasserwerken jeder Größe, Enteisungsanlagen 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; Gewindeschleif-einrichtungen; 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 1200 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.-Grünwald.
Prokuristen: Paul Derpsch, Frieda Thierschmann.

Research Question:

- Identification of changes in legal form of the listed companies
- Extraction of the **company name** and **legal forms** necessary

**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.

Research Question:

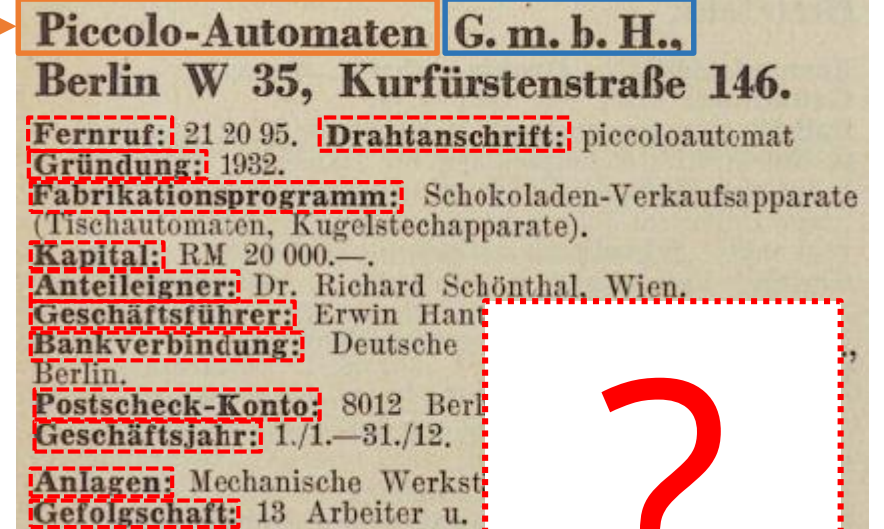
- Identification of changes in legal form of the listed companies
- Extraction of the **company name** and **legal forms** necessary



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.

Research Question:

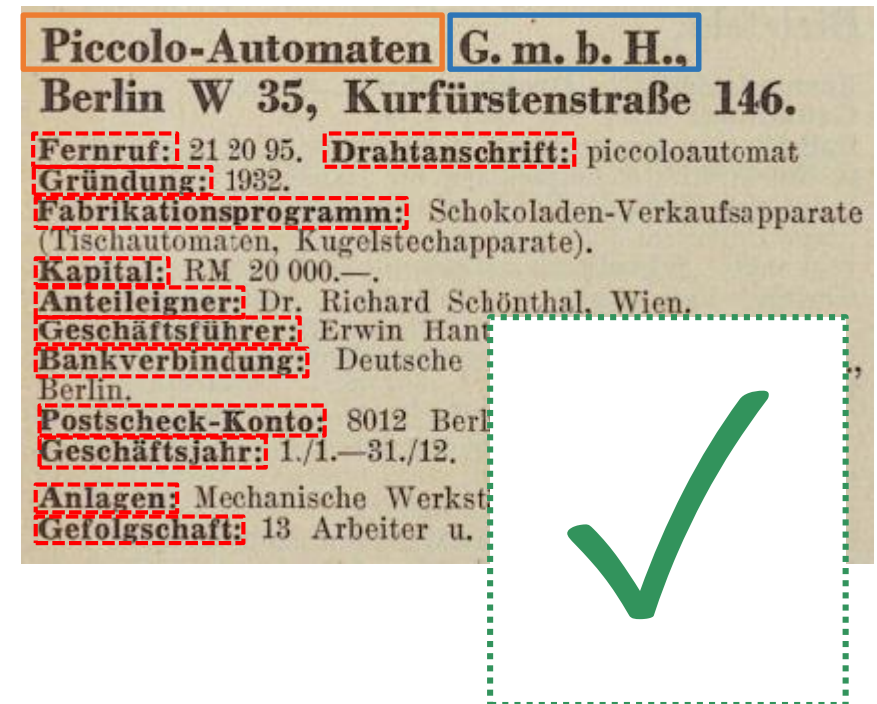
- Identification of changes in legal form of the listed companies
- Extraction of the **company name** and **legal forms** necessary



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 Han
Bankverbindung: Deutsche
Berlin.
Postscheck-Konto: 8012 Berl
Geschäftsjahr: 1./1.—31./12.
Anlagen: Mechanische Werkst
Gefolgschaft: 13 Arbeiter u.

Approach and goal:

- Extract and structure data beyond the research question (realistic additional effort)
- Ensure reusability of the data for other research purposes
- Test workflow to gain experience for comparable projects



Approach and goal:

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

Structured research data

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. Introduction & Digitalization

Challenges (text recognition):

- High quality requirements for research data
- Effort for OCR training difficult to estimate
- Layout (two columns → 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.—.
Anteilhaber: C. R. Nestmann, Meerane (100%).
Geschäftsführer: Obering. A. Wackermann.
Prokurist: J. Frenzel.
Bankverbindungen: Meeraner Bank A.-G., Reichsbank, Meerane.
Postcheck-Konto: 115 485 Leipzig.
Geschäftsjahr: 1./1.—31./12.

Grundbesitz: 3700 qm, davon 1403 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 Schlangensammel-Trockenapparate System Otto und Getreidetrockner nach dem Zellsystem neben anderen allgemeinen Trocknungsanlagen. 1921 erfolgte neben dem Trocknerbau die Aufnahme der Fabrikation von Holzbearbeitungsmaschinen. 1932 übernahm das Werk zusätzlich die Herstellung von Bewoilt-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, Kugelschapparate).
Kapital: RM 20 000.—.
Anteilhaber: Dr. Richard Schönthal, Wien.
Geschäftsführer: Erwin Hantsch, Bln.-Steglitz.
Bankverbindung: Deutsche Bank u. Disconto-Ges., Berlin.
Postcheck-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.
Postcheck-Konto: 4847 Berlin.
Geschäftsjahr: Kalenderjahr.
Grundbesitz: 9500 qm, davon 4500 qm bebaut.
Anlagen: Maschinenbau-Werkstätten (Dreherei, Schleiferei, Eiserei, 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: 1903.
Fabrikationsprogramm: Entwurf und Bau von Trink- und Gebrauchswasserwerken jeder Größe, Enteisungsanlagen 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.
Postcheck-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.
Postcheck-Konto: 2178 Stuttgart.
Geschäftsjahr: Kalenderjahr.
Grundbesitz: 2600 qm, davon 2000 qm bebaut; gepachtet sind 1200 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.-Grünwald.
Prokuristen: Paul Derpsch, Frieda Thierschmann.

462

1. Introduction & Digitalization



Challenges (data structuring):

- No "all-in-one" solution available: Software must be written in-house
- High quality requirements for research data
- Profile separation
- Inconsistent attribute naming throughout the publication:
 - *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.—.
Anteilseigner: G. 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 1403 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 Schlangensammel-Trockenapparate System Otto und Getreidetrockner nach dem Zellen-system neben anderen allgemeinen Trocknungsanlagen. 1921 erfolgte neben dem Trocknerbau die Aufnahme der Fabrikation von Holzbearbeitungsmaschinen. 1932 übernahm das Werk zusätzlich die Herstellung von Bewöld-Apparaturen für die Papierindustrie nach den Patenten Dr. Bruno Wiegler, Berlin.

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

Fernruf: 21 20 95. **Drahtanschrift:** piccoloautomat
Gründung: 1932.
Fabrikationsprogramm: Schokoladen-Verkaufsapparate (Tischautomaten, Kugelschapparate).
Kapital: RM 20 000.—.
Anteilseigner: 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. Hebe-maschinen (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, Eiserei, 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: 1903.
Fabrikationsprogramm: Entwurf und Bau von Trink- und Gebrauchswasserwerken jeder Größe, Enteisungsanlagen 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-, Absaug- 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; Gewindeschleif-einrichtungen; 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 1200 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.-Grünwald.
Prokuristen: Paul Derpsch, Frieda Thierschmann.

462

1. Introduction & Digitalization



Challenges (Infrastructure, staff):

- Multi-part workflow with specific infrastructural and staff requirements
- Digitization (1 project coordinator, 1 research assistant)
- OCR (1 project coordinator, 1 developer)
- Data structuring (1 developer)
- 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.—.
Anteilhaber: G. 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 1403 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 Schlangensystem-Trockenapparate System Otto und Getreidetrockner nach dem Zellen-system neben anderen allgemeinen Trocknungsanlagen. 1921 erfolgte neben dem Trocknerbau die Aufnahme der Fabrikation von Holzbearbeitungsmaschinen. 1932 übernahm das Werk zusätzlich die Herstellung von Bewöld-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, Kugelschapparate).
Kapital: RM 20 000.—.
Anteilhaber: 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. Hebe-maschinen (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, Feiserei, 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, Enteisungsanlagen 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 1200 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.-Grünwald.
Prokuristen: Paul Derpsch, Frieda Thierschmann.

462

2. Layout segmentation and OCR via eScriptorium

2. Layout segmentation and OCR via eScriptorium

eScriptorium

Transcription software

Open source platform for manual or automated segmentation and text recognition of historical manuscripts and prints.



Developer:	PSL (Paris)
Year of release:	2018
Current version:	0.10.5 (2022)
Operating system:	platform independent
Programming Lang.:	Python, JavaScript, Hypertext Markup Language
OCR engine:	Kraken
License:	MIT license
Models:	https://ub-backup.bib.uni-mannheim.de/~stweil/eScriptorium/ (Fraktur)
Additional info:	https://ocr-bw.bib.uni-mannheim.de/tag/escriptorium/

2. Layout segmentation and OCR via eScriptorium

Kraken

OCR engine for layout segmentation and text recognition

Kraken is an all-in-one open source OCR software solution optimized for historical documents and non-Latin writing systems, with fully trainable layout analysis and text recognition.

Developer: Benjamin Kiessling

Year of release: 2015

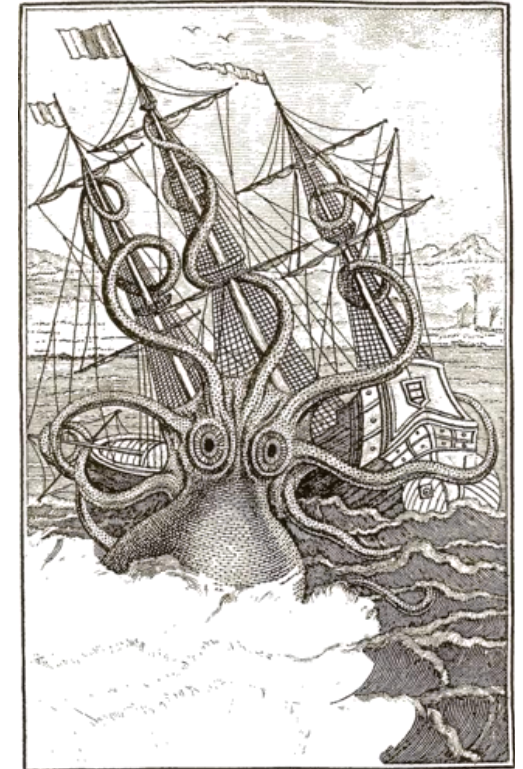
OCR technology: bounding box and baseline-based (eScriptorium)

Export formats: ALTO, PageXML, abbyXML and hOCR

Programming lang.: Python

Models: <https://ub-backup.bib.uni-mannheim.de/~stweil/eScriptorium/> (Fraktur)

License : MIT License



Workflow eScriptorium

Layout segmentation and text recognition

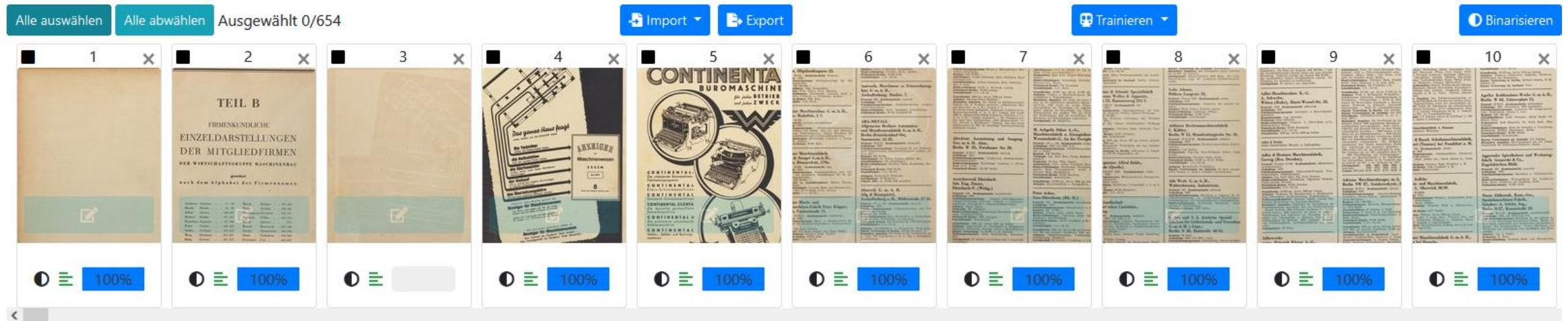
1. Upload of the digitized files
2. Selection of a suitable base model for layout segmentation and text recognition
3. Identification of weak points of the models
4. Creation of ground truth
5. Work-specific training of the models
6. Layout segmentation and text recognition

If no base model can be found, another system should be used or a "from-scratch" model has to be trained (much higher training effort)

2. Layout segmentation and OCR via eScriptorium

Upload of the digitized files

Digitized files can be uploaded via the GUI or the API individually or in a batch



Overview of all 654 images in the newly created project

Layout segmentation

Selection of a suitable base model

Segmentation consists of two steps:

1. Segmentation of text regions
2. Segmentation of text lines (baseline and polygons)



Suitable base model:
cbad_1800_compensated_50

2. Layout segmentation and OCR via eScriptorium



Weakness: profile segmentation

Abwärme — Ackermann

Fabrikationsprogramm: Beton- u. Mörtelmischer; Bauwinden, Basenofen.
Kapital: RM 20 000.—
Anteilhaber: Arthur Schumann, Rich. Findeisen, Ernst Schumann sen.
Geschäftsführer: Arthur Schumann, Rich. Findeisen, Ing. O. Selz.
Bankverbindungen: Stadt- u. Girobank, Leipzig.
Postcheck-Konto: 70 550 Leipzig.
Geschäftsjahr: 1./1.—31./12.

Grundbesitz: 6000 qm, davon 3000 qm bebaut.
Anlagen: Maschinenwerkstätten.
Besondere Angaben: Die Fabrik ist Fabrikationsnachfolgerin der in Konkurs gerathenen „Allgemeinen Baumaschinen-Ges. m. b. H.“, Leipzig o. L.
Die Firma hat sich anfangl. „Schumann, Findeisen & Co., Baumaschinenfabrik G. m. b. H.“, Leipzig, und wurde 1904 in „Baumaschinenfabrik Schumann, Findeisen & Co. G. m. b. H.“, Leipzig, umfirmirt. — Die Fabrikate werden unter dem Schutzzeichen „ABO-Baumaschinen“ sowie „Neoroll- u. Rib-Mischer“ verkauft. Aus dem ersten Schutzzeichen ist 1907 die endgültige Firmenbezeichnung entwickelt worden.

Abwärme Ausnutzung und Saugzug
Ges. m. b. H. Abas,
Berlin W 35, Potsdamer Str. 28.
Fernruf: 22 63 17. **Drachenschrift:** abwasch.
Gründung: 1921.
Fabrikationsprogramm: Ventilatoren, Staubabscheider, Gasanhalterbühler.
Bankverbindungen: Reichsbank, Commerz- u. Privatbank, A.-G., Berlin.
Postcheck-Konto: 118 120 Berlin.

Acetylenwerk Ebersbach
Inh. Eug. Zinser,
Ebersbach-F. (Witbg.)
Fernruf: 216. **Drachenschrift:** acetylenwerk.
Gründung: 1898.
Fabrikationsprogramm: Antogen-Schweißapparate; Antogen-Werkzeuge (Schweiß-, Schneid- u. Lötlöffel); Sauerstoff-, Wasserstoff- u. Dünnsäureventile).
Inhaber: Eugen Zinser.
Präkurist: Th. Friedrich.
Bankverbindungen: Gewerbank, Ebersbach-F.; Reichsbank, Göttingen-Witbg.
Postcheck-Konto: 3228 Stuttgart.
Geschäftsjahr: Kalenderjahr.
Grundbesitz: 1500 qm, davon 1500 qm bebaut.
Anlagen: Apparatebau u. Schweißerei u. Klempnerei, Schneid-, Schlosserei, Montage, Eisen- u. Metalldreherei.
Eigene Vertretung: in Berlin u. im Ausland.
Tochtergesellschaft: 8844 „Spindelwerke, Ebersbach-F.“
Gefolgschaft: 30 Arbeiter (4 Lehrling) und 7 Angestellte (1 Lehrling).

Achenbach Söhne G. m. b. H.,
Buschhütten (Kr. Siegen).
Fernruf: Siegen 5011. **Drachenschrift:** achenbachsöhne.
Gründung: 1846.
Fabrikationsprogramm: Walzwerkzeug- und -walzen-gießerei.
Kapital: RM 1 254 000.—

Anteilhaber: Frau Dr. Barten, Dr. Ing. Ernst Barten, Ernst Gieseler (Geschäftsführer).
Präkuristen: Karl Roth, Eduard Reinschmidt, Heinrich Bester.
Bankverbindungen: Reichsbank, Deutsche Bank und Disconto-Ges., Siegen.
Postcheck-Konto: 873 Dortmund.
Geschäftsjahr: 1./1.—31./6.

Grundbesitz: 16 600 qm, davon 16 000 qm bebaut.
Anlagen: Gießerei u. Modelldreherei; 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.
Die Firma hat sich anfangl. „Schumann, Findeisen & Co., Baumaschinenfabrik G. m. b. H.“, Leipzig, und wurde 1904 in „Baumaschinenfabrik Schumann, Findeisen & Co. G. m. b. H.“, Leipzig, umfirmirt. — Die Fabrikate werden unter dem Schutzzeichen „ABO-Baumaschinen“ sowie „Neoroll- u. Rib-Mischer“ verkauft. Aus dem ersten Schutzzeichen ist 1907 die endgültige Firmenbezeichnung entwickelt worden.

M. Achgelis Söhne A.-G.,
Maschinenfabrik u. Eisengießerei,
Wesermünde-G, An der Zweighahn 1.
Fernruf: 101 u. 146. **Drachenschrift:** achgeliswerke.
Gründung: 1883; seit 1918 A.-G.
Fabrikationsprogramm: Schälhilfsmaschinen in jeder Art u. Größe.
Kapital: RM 225 000.—
Vorstand: Ing. Karl Boos, Georg Brinkmann, Werner Sander.
Präkurist: Abw. Ing. Wilh. Barth.
Aufsichtsrat: Vors. Arthur Friedrichs, Bremerhaven.
Bankverbindungen: Reichsbank, Wesermünde-Gesellschaftsbank, Wesermünde-Ges., Nordd.-Kreditbank, Bremer Bank, Bremerhaven.
Postcheck-Konto: 19 028 Hamburg.
Geschäftsjahr: Kalenderjahr.
Grundbesitz: 18 500 qm, davon 7200 qm bebaut.
Anlagen: Maschinenfabrik u. Eisengießerei.

Peter Acker,
Gau-Odernheim (Rh. H.)
Fernruf: 226. **Drachenschrift:** maschinenacker.
Gründung: 1878.
Fabrikationsprogramm: Höhenförderer u. Pflüge.
Kapital: RM 70 000.—
Inhaber: Johann Acker, Jakob Acker.
Bankverbindungen: Volksbank, Alzey; Spars- u. Darlehnskasse, Gau-Odernheim.
Postcheck-Konto: 23 577 Frankfurt a/M.; 8617 Ludwigshafen.
Geschäftsjahr: 1./1.—31./12.
Grundbesitz: 1200 qm, davon 800 qm bebaut; gepachtet sind 390 qm; gesamte Nutzfläche 1500 qm.
Anlagen: Fabrikationsanlage, Verfahrungs- u. Anstellstraßen, Tischlerei.
Gefolgschaft: 29 Arbeiter, 5 Lehrlinge u. 1 Angestellter (1 Lehrling).

Friedrich Ackermann,
Werkzeug- und Maschinenfabrik,
Wuppertal-Barmen, Oberdenkmalstr. 89.
Fernruf: 54 282.
Gründung: 1912.
Fabrikationsprogramm: Maschinenschraubstöcke, Zahnrad-, Gewindestift-, Frästaten-, Vorrichtungen- und Drehstühle.

Segmentation with cbad_1800

Optimal segmentation

Abwärme — Ackermann

Fabrikationsprogramm: Beton- u. Mörtelmischer; Bauwinden, Basenofen.
Kapital: RM 20 000.—
Anteilhaber: Arthur Schumann, Rich. Findeisen, Ernst Schumann sen.
Geschäftsführer: Arthur Schumann, Rich. Findeisen, Ing. O. Selz.
Bankverbindungen: Stadt- u. Girobank, Leipzig.
Postcheck-Konto: 70 550 Leipzig.
Geschäftsjahr: 1./1.—31./12.

Grundbesitz: 6000 qm, davon 3000 qm bebaut.
Anlagen: Maschinenwerkstätten.
Besondere Angaben: Die Fabrik ist Fabrikationsnachfolgerin der in Konkurs gerathenen „Allgemeinen Baumaschinen-Ges. m. b. H.“, Leipzig o. L.
Die Firma hat sich anfangl. „Schumann, Findeisen & Co., Baumaschinenfabrik G. m. b. H.“, Leipzig, und wurde 1904 in „Baumaschinenfabrik Schumann, Findeisen & Co. G. m. b. H.“, Leipzig, umfirmirt. — Die Fabrikate werden unter dem Schutzzeichen „ABO-Baumaschinen“ sowie „Neoroll- u. Rib-Mischer“ verkauft. Aus dem ersten Schutzzeichen ist 1907 die endgültige Firmenbezeichnung entwickelt worden.

Abwärme Ausnutzung und Saugzug
Ges. m. b. H. Abas,
Berlin W 35, Potsdamer Str. 28.
Fernruf: 22 63 17. **Drachenschrift:** abwasch.
Gründung: 1921.
Fabrikationsprogramm: Ventilatoren, Staubabscheider, Gasanhalterbühler.
Bankverbindungen: Reichsbank, Commerz- u. Privatbank, A.-G., Berlin.
Postcheck-Konto: 118 120 Berlin.

Acetylenwerk Ebersbach
Inh. Eug. Zinser,
Ebersbach-F. (Witbg.)
Fernruf: 216. **Drachenschrift:** acetylenwerk.
Gründung: 1898.
Fabrikationsprogramm: Antogen-Schweißapparate; Antogen-Werkzeuge (Schweiß-, Schneid- u. Lötlöffel); Sauerstoff-, Wasserstoff- u. Dünnsäureventile).
Inhaber: Eugen Zinser.
Präkurist: Th. Friedrich.
Bankverbindungen: Gewerbank, Ebersbach-F.; Reichsbank, Göttingen-Witbg.
Postcheck-Konto: 3228 Stuttgart.
Geschäftsjahr: Kalenderjahr.
Grundbesitz: 1500 qm, davon 1500 qm bebaut.
Anlagen: Apparatebau u. Schweißerei u. Klempnerei, Schneid-, Schlosserei, Montage, Eisen- u. Metalldreherei.
Eigene Vertretung: in Berlin u. im Ausland.
Tochtergesellschaft: 8844 „Spindelwerke, Ebersbach-F.“
Gefolgschaft: 30 Arbeiter (4 Lehrling) und 7 Angestellte (1 Lehrling).

Achenbach Söhne G. m. b. H.,
Buschhütten (Kr. Siegen).
Fernruf: Siegen 5011. **Drachenschrift:** achenbachsöhne.
Gründung: 1846.
Fabrikationsprogramm: Walzwerkzeug- und -walzen-gießerei.
Kapital: RM 1 254 000.—

Anteilhaber: Frau Dr. Barten, Dr. Ing. Ernst Barten, Ernst Gieseler (Geschäftsführer).
Präkuristen: Karl Roth, Eduard Reinschmidt, Heinrich Bester.
Bankverbindungen: Reichsbank, Deutsche Bank und Disconto-Ges., Siegen.
Postcheck-Konto: 873 Dortmund.
Geschäftsjahr: 1./1.—31./6.

Grundbesitz: 16 600 qm, davon 16 000 qm bebaut.
Anlagen: Gießerei u. Modelldreherei; 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.
Die Firma hat sich anfangl. „Schumann, Findeisen & Co., Baumaschinenfabrik G. m. b. H.“, Leipzig, und wurde 1904 in „Baumaschinenfabrik Schumann, Findeisen & Co. G. m. b. H.“, Leipzig, umfirmirt. — Die Fabrikate werden unter dem Schutzzeichen „ABO-Baumaschinen“ sowie „Neoroll- u. Rib-Mischer“ verkauft. Aus dem ersten Schutzzeichen ist 1907 die endgültige Firmenbezeichnung entwickelt worden.

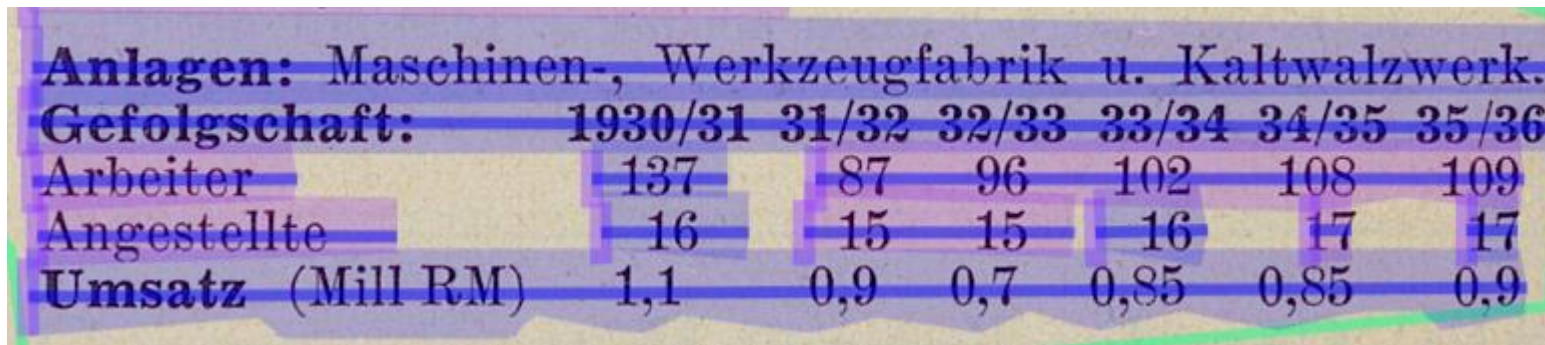
M. Achgelis Söhne A.-G.,
Maschinenfabrik u. Eisengießerei,
Wesermünde-G, An der Zweighahn 1.
Fernruf: 101 u. 146. **Drachenschrift:** achgeliswerke.
Gründung: 1883; seit 1918 A.-G.
Fabrikationsprogramm: Schälhilfsmaschinen in jeder Art u. Größe.
Kapital: RM 225 000.—
Vorstand: Ing. Karl Boos, Georg Brinkmann, Werner Sander.
Präkurist: Abw. Ing. Wilh. Barth.
Aufsichtsrat: Vors. Arthur Friedrichs, Bremerhaven.
Bankverbindungen: Reichsbank, Wesermünde-Gesellschaftsbank, Wesermünde-Ges., Nordd.-Kreditbank, Bremer Bank, Bremerhaven.
Postcheck-Konto: 19 028 Hamburg.
Geschäftsjahr: Kalenderjahr.
Grundbesitz: 18 500 qm, davon 7200 qm bebaut.
Anlagen: Maschinenfabrik u. Eisengießerei.

Peter Acker,
Gau-Odernheim (Rh. H.)
Fernruf: 226. **Drachenschrift:** maschinenacker.
Gründung: 1878.
Fabrikationsprogramm: Höhenförderer u. Pflüge.
Kapital: RM 70 000.—
Inhaber: Johann Acker, Jakob Acker.
Bankverbindungen: Volksbank, Alzey; Spars- u. Darlehnskasse, Gau-Odernheim.
Postcheck-Konto: 23 577 Frankfurt a/M.; 8617 Ludwigshafen.
Geschäftsjahr: 1./1.—31./12.
Grundbesitz: 1200 qm, davon 800 qm bebaut; gepachtet sind 390 qm; gesamte Nutzfläche 1500 qm.
Anlagen: Fabrikationsanlage, Verfahrungs- u. Anstellstraßen, Tischlerei.
Gefolgschaft: 29 Arbeiter, 5 Lehrlinge u. 1 Angestellter (1 Lehrling).

Friedrich Ackermann,
Werkzeug- und Maschinenfabrik,
Wuppertal-Barmen, Oberdenkmalstr. 89.
Fernruf: 54 282.
Gründung: 1912.
Fabrikationsprogramm: Maschinenschraubstöcke, Zahnrad-, Gewindestift-, Frästaten-, Vorrichtungen- und Drehstühle.

Weakness: text lines

When recognizing text lines, the baseline is not always correct and sometimes lines are divided into small parts



The image shows a table with text line segmentation. The table has five rows and seven columns. The first row is a header for the table, and the second row is a header for the data columns. The data rows are 'Arbeiter', 'Angestellte', and 'Umsatz (Mill RM)'. The segmentation is shown by blue and green lines that do not perfectly align with the text, illustrating the weakness mentioned in the text.

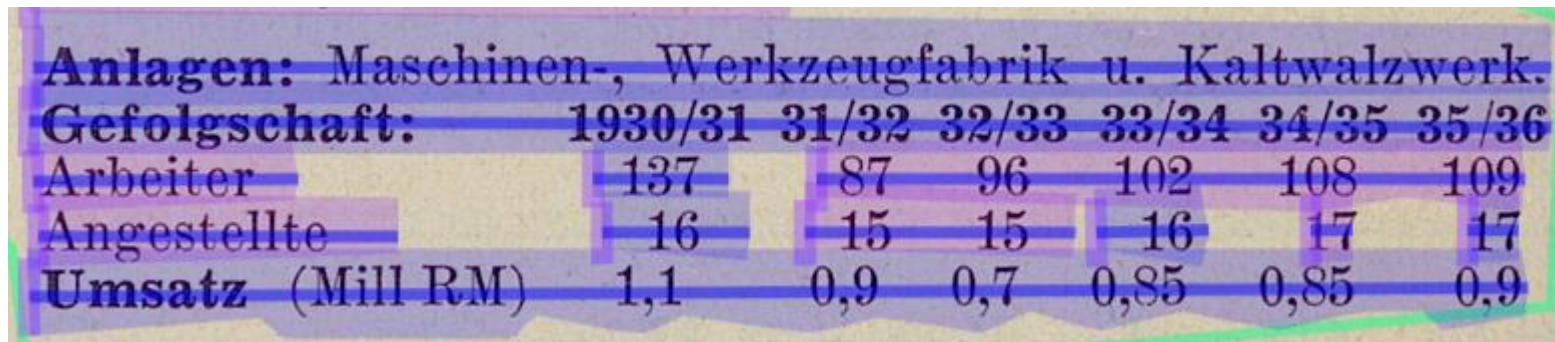
Anlagen: Maschinen-, Werkzeugfabrik u. Kaltwalzwerk.						
Gefolgschaft:	1930/31	31/32	32/33	33/34	34/35	35/36
Arbeiter	137	87	96	102	108	109
Angestellte	16	15	15	16	17	17
Umsatz (Mill RM)	1,1	0,9	0,7	0,85	0,85	0,9

Example: Text line segmentation of a table

Work-specific training

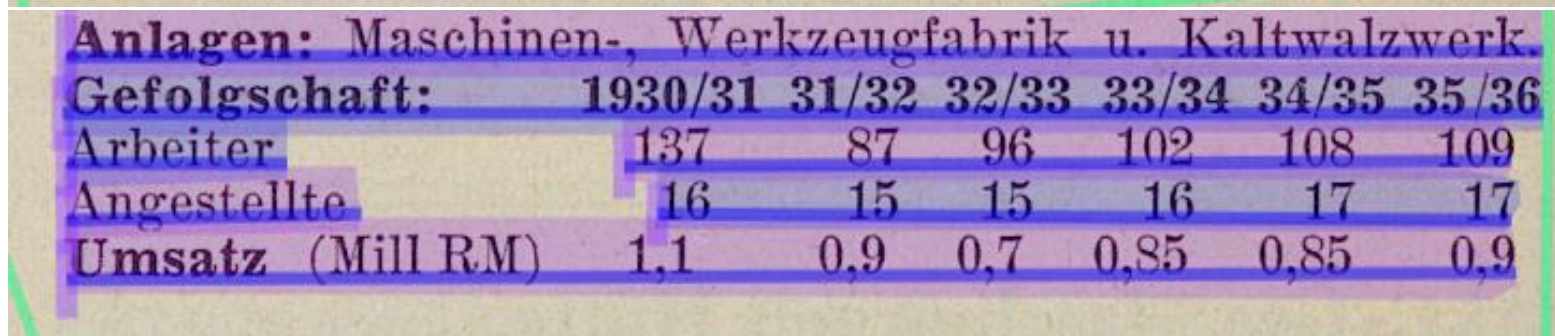
- Ground truth:** 47 pages
- Result:** Recognition of text regions unsatisfactory (solution: usage of a different pre- and post-processing workflow)
- Results (text lines):** Significant improvement

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

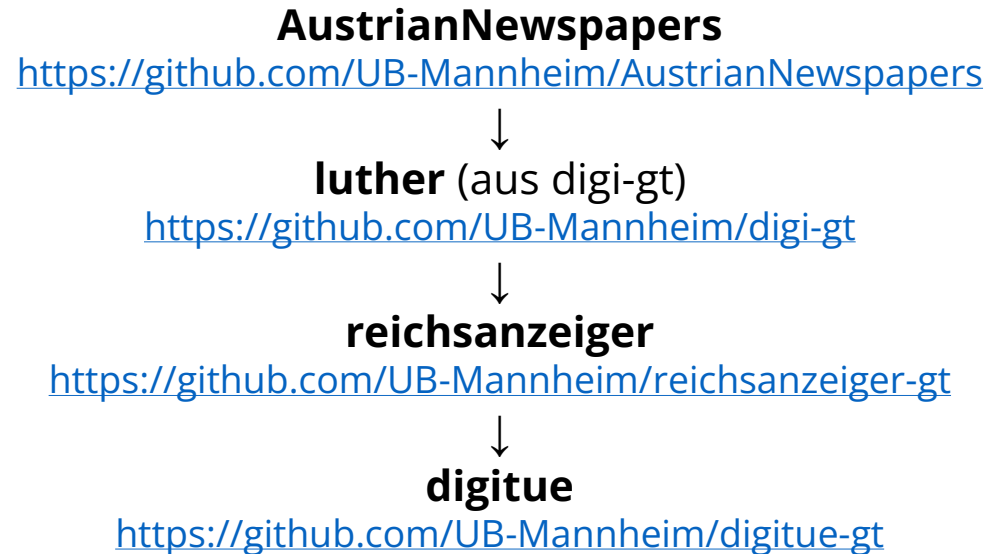
OCR

OCR: Selection of a suitable base model

A suitable base model should:

- 1) cover the character set of the source material (if possible)
- 2) have already good recognition rates on the material in question

Suitable OCR model: digitue

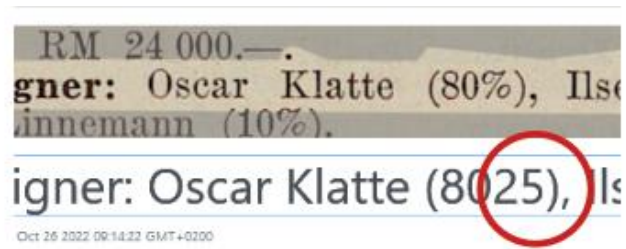


Weaknesses in text recognition

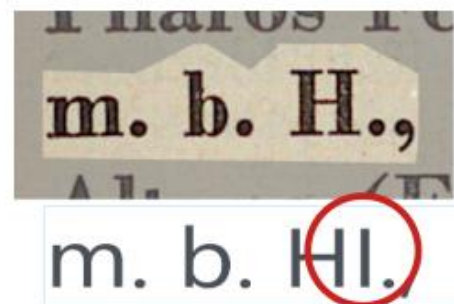
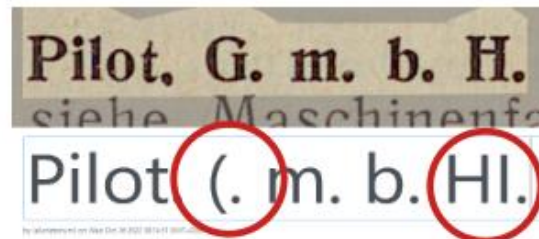
The model already delivers very good results with text recognition rates > 98%

Some specific errors can be identified:

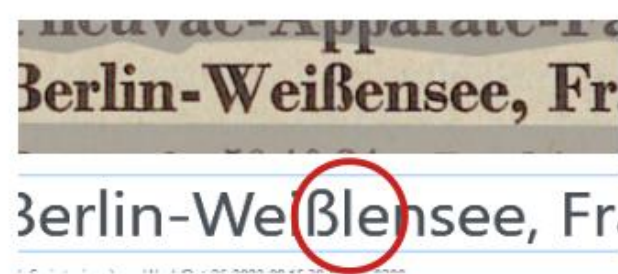
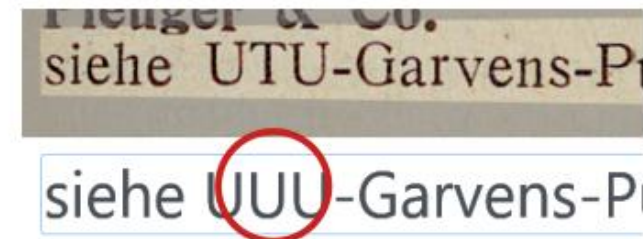
Percent sign



String „G.m.b.H.“



rare characters / character combinations



2. Layout segmentation and OCR via eScriptorium

Work-specific training

Ground truth:

26 pages

Results text recognition:

CER could be improved to > 99.85 %

digitue

RM 24 000.—.
igner: Oscar Klatte (80%), Ilse
innemann (10%).
igner: Oscar Klatte (8025), Ilse

Pilot, G. m. b. H.
siehe Maschinenfa
Pilot. (. m. b. Hl.

Pleuger & Co.
siehe UTU-Garvens-P
siehe UUU-Garvens-P

digitue_trained

M 24 000.—.
er: Oscar Klatte (80%), Ilse
emann (10%).
er: Oscar Klatte (80%), Ilse

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

Pleuger & Co.
siehe UTU-Garvens
siehe UTU-Garvens

3. Extraction and Structuring via Python

3. Extraction and Structuring via Python: Input data



```
1 <?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
2 <PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15" xmlns
3 xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15
4 http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15/pagecontent.xsd">
5 <Metadata>
6 <Creator>escriptorium</Creator>
7 <Created>2022-09-19T12:10:54.541269+00:00</Created>
8 <LastChange>2022-09-19T12:10:54.541290+00:00</LastChange>
9 </Metadata>
10 <Page imageFilename="maschinenindustrie_1937_0452.jpg" imageWidth="2796" imageHeight="3456">
11 <TextRegion id="r_2_1" >
12 <Coords points="127,354 1290,354 1290,685 127,685"/>
13 <TextLine id="eSc_line_835f5cfa" >
14 <Coords points="1269,350 1148,348 1132,348 1109,348 1086,348 1063,348 1040,348 1017,348 994,348 971,348 948,348 925,348 902,348 879,348 856,348 833,348 810,348 787,348 764,348 741,348 718,348 695,348 672,348 649,348 626,348 603,348 580,348 557,348 534,348 511,348 488,348 465,348 442,348 419,348 396,348 373,348 350,348 327,348 304,348 281,348 258,348 235,348 212,348 189,348 166,348 143,348 120,348 97,348 74,348 51,348 28,348 5,348 127,354 1290,354 1290,685 127,685"/>
15 <Baseline points="126,382 1292,387"/>
16 <TextEquiv>
17 <Unicode>Bankverbindungen: Reichsbank, Commerz u. Privat-</Unicode>
18 </TextEquiv>
19 </TextLine>
20 <TextLine id="eSc_line_eff01337" >
21 <Coords points="128,426 126,449 877,449 880,430 877,396 128,389 128,426"/>
22 <Baseline points="128,426 880,430"/>
23 <TextEquiv>
24 <Unicode>Bank A.-G., Dresdner Bank, Berlin.</Unicode>
25 </TextEquiv>
26 </TextLine>
```

Is PAGE-XML enough to answer the research question?

JPG

PAGE-XML

3. Extraction and Structuring via Python: blatt

Blatt functions:

1. reads PAGE-XML
2. removes hyphens
3. converts PAGE-XML to TXT and TSV formats
4. provides module for data structuring

Blatt

pypi package 0.1.6

NLP-helper for OCR-ed pages in [PAGE XML](#) format.

Installation

```
pip install blatt
```

Open Source + CLI



<https://pypi.org/project/blatt>

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

OCR segmentation of the text regions with problems

blatt



1. sorts text lines
2. makes geometric segmentation based on coordinates
3. data cleaning
4. merges segments of subsequent columns and pages

3. Extraction and Structuring via Python: Segmentation

JPG



Transformation
via blatt

Approach: use colon as
separator

Problem: inconsistent attribute
naming, printing mistakes, OCR
errors

e.g.: OCR problems:

*{'Postscheck-Konto', 'Postseheck-Konto', 'Postscheckkonto', 'Postscheck-Konnto', 'Fostscheck-Konto', 'Postcheck-Konto', 'Postscheek-Konto',
'Potscheck-Konto', 'Postsbeck-Konto', 'Postscheckkonto', 'Postschek-Konto', 'Postscheck-Konten', 'Ponstscheck-Konto', 'Postscheck-Konto'}*

Structured
research data

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

...

Sorting and grouping of properties

Postscheck-Konto

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

Geschäftsjahr

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

Fabrikationsprogramm

```
{'Fabfikationsprogramm', 'Fabrikationprogramm', 'Fabrikationsprorammm', 'Fabrikationsprogramm:',  
'Fabrikstionsprogramm', 'Fabrkkationsprogramm', 'Fabrikationsprogramm'}
```


Extract *Drahtanschrift* from *Fernruf*:

Fernruf: 212095. *Drahtanschrift*: piccoloautomat



Fernruf	212095
Drahtanschrift	piccoloautomat

Extraction of the legal form:

{'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. Extraction and Structuring via Python: CSV and XLSX

Structured research data

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

...

CSV and XLSX

	Company	RAW_TEXT	W_TEXT_1	LE_SEGMEN	FABRIKATIONSPROGRAMM	POSTSCHECK-KONTO	FERNRUF	DRAHTANSCHRIFT	BANKVERBINDUNGEN	ANLAGEN
0	Aachener Kratz	Aachener K Cassalette f Aachen, Oli Fernruf: 34 Gründung: Fabrikation Verwendun Kapital: RM Geschäftsfü Prokuristen Bankverbin Disconto-G Postscheck: Geschäftsj	Aachener K	/MI1937/r	Kratzenbeschläge für alle Ver	2952 Köln	34 041	kratzena	Reichsbank, Deutsche Bank u. Disconto-Ges	
1	Aachener Mas	Aachener M Aachen, Ru Fernruf: 25 Drahtansch Gründung: Fabrikation zur Herstell Drahtartike die Kratzen Bankverbin bank A.-G.,	Aachener N	/MI1937/r	Drahtbearbeitungsmaschinen	16 987 Köln	25 205	aachener maschinenbau	Reichsbank, Commerz- u. Privatbank A.-G.,	
2		Aachener N Rothe & Ste Aachen, Bis Fernruf: 25 Gründung: Fabrikation Lieferung v samte Kratz Kapital: RM Anteileigne Peter Rump								

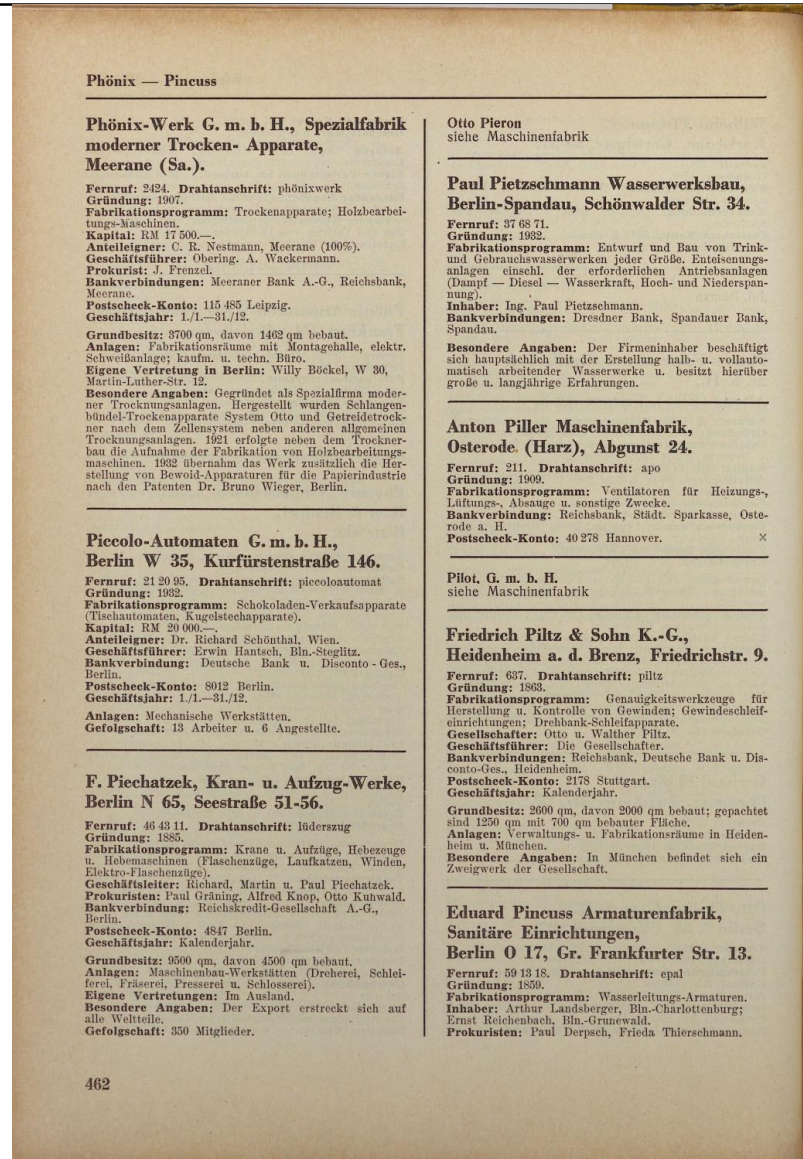
4. Summary

4. Summary



Results:

- Very good text recognition accuracy (> 99.85%) through work-specific training
- eScriptorium very well suited as a platform for ground truth production and training:
 - fast, user-friendly, platform-independent production of transcriptions
 - user-friendly training process (GUI)
- Data structuring:
 - no all-in-one solution → we developed the open source tool *blatt* → it can be reused in similar projects
 - we structured all available data (not only legal forms) → the data can be reused to answer other research questions
- The services of UB Mannheim can be used flexibly and cooperatively in projects with other stakeholders of the University



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.—
Anteilhaber: C. E. Nestmann, Meerane (100%).
Geschäftsführer: Oering, A. Wackernann.
Prokurist: J. Frenzel.
Bankverbindungen: Meeraner Bank A.-G., Reichsbank, Meerane.
Postcheck-Konto: 115 485 Leipzig.
Geschäftsjahr: 1./1.—31./12.

Grundbesitz: 3700 qm, davon 1403 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 Schlangensystem-Trockenapparate System Otto und Getreidetrockner nach dem Zellen-system neben anderen allgemeinen Trocknungsanlagen. 1921 erfolgte neben dem Trocknerbau die Aufnahme der Fabrikation von Holzbearbeitungsmaschinen. 1932 übernahm das Werk zusätzlich die Herstellung von Bewöld-Apparaturen für die Papierindustrie nach den Patenten Dr. Bruno Wiegler, Berlin.

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

Fernruf: 21 20 95. **Drahtanschrift:** piccoloautomat
Gründung: 1932.
Fabrikationsprogramm: Schokoladen-Verkaufsapparate (Tischautomaten, Kugelschapparate).
Kapital: RM 20 000.—
Anteilhaber: Dr. Richard Schöenthal, Wien.
Geschäftsführer: Erwin Hantsch, Bln.-Steglitz.
Bankverbindung: Deutsche Bank u. Disconto-Ges., Berlin.
Postcheck-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. Hebe-maschinen (Flaschenzüge, Laufkatzen, Winden, Elektro-Flaschenzüge).
Geschäftsleiter: Richard, Martin u. Paul Piechatzek.
Prokuristen: Paul Gräning, Alfred Knop, Otto Kuhwald.
Bankverbindung: Reichskredit-Gesellschaft A.-G., Berlin.
Postcheck-Konto: 4847 Berlin.
Geschäftsjahr: Kalenderjahr.
Grundbesitz: 9500 qm, davon 4500 qm bebaut.
Anlagen: Maschinenbau-Werkstätten (Dreherei, Schleiferei, Eiserei, 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, Enteisungsanlagen 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.
Postcheck-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: Genauigkeitwerkzeuge für Herstellung u. Kontrolle von Gewinden; Gewindeschleif-einrichtungen; Drehbank-Schleifapparate.
Gesellschafter: Otto u. Walther Piltz.
Geschäftsführer: Die Gesellschafter.
Bankverbindungen: Reichsbank, Deutsche Bank u. Disconto-Ges., Heidenheim.
Postcheck-Konto: 2178 Stuttgart.
Geschäftsjahr: Kalenderjahr.
Grundbesitz: 2600 qm, davon 2000 qm bebaut; gepachtet sind 1200 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.-Grünwald.
Prokuristen: Paul Derpsch, Frieda Thierschmann.

Time investment:

- **Total:** approx. 2 work weeks (spread over 2 months)
- **Digitization:** 1 work day
- **OCR:** 1 work week:
 - Evaluation of existing OCR models on the material
 - Design of an OCR workflow (layout segmentation + OCR)
 - Ground truth production and subsequent training (layout segmentation + OCR)
 - Final OCR for all pages
- **Data structuring:** 1 work week:
 - Segmentation
 - Post-processing
 - Extension of *blatt*

Team:

- **6 Project participants**
 - Prof. Jochen Streb (**Chair of Economic History**)
 - 1 research assistant for ground truth production and QA (**Chair of Economic History**)
 - 1 project manager for digitization (**UB Mannheim**)
 - 1 project coordinator + OCR (**UB Mannheim**)
 - 1 developer OCR (**UB Mannheim**)
 - 1 developer data structuring (**BERD@NFDI, UB Mannheim**)

Feedback, Questions?

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

Thank You!

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

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