Data in Brief 54 (2024) 110274

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Reichsanzeiger-GT: An OCR ground truth dataset based on the historical newspaper "Deutscher Reichsanzeiger und Preußischer Staatsanzeiger" (German Imperial Gazette and Prussian Official Gazette) (1819–1945)

Thomas Schmidt*, Jan Kamlah, Stefan Weil

University of Mannheim, University Library, Schloss Schneckenhof, 68161 Mannheim

ARTICLE INFO

Article history:
Received 20 December 2023
Revised 13 February 2024
Accepted 23 February 2024
Available online 7 March 2024

Dataset link: reichsanzeiger-gt (Original data)

Keywords:
OCR
Text recognition
Ground truth
Historical newspapers

ABSTRACT

Reichsanzeiger-GT is a ground truth dataset for OCR training and evaluation based on the historical German newspaper "Deutscher Reichsanzeiger und Preußischer Staatsanzeiger" (German Imperial Gazette and Prussian Official Gazette), which was published from 1819 to 1945 and printed mostly in the typeface Fraktur (Black Letter). The dataset consists of 101 newspaper pages for the years 1820–1939, that cover a wide variety of topics, page layouts (lists, tables, and advertisements) as well as different typefaces. Using the transcription software Transkribus and the open-source OCR engine Tesseract we automatically created and manually corrected layout segmentations and transcriptions for each page, resulting in 65,563 text regions, 412 table regions, 119,429 text lines and 490,679 words. By applying transcription guidelines that preserve the printing conditions, the dataset contains language and printing specific phenomena like the historical use of glyphs like long s (ſ), rotunda r (ꝛ), and

* Corresponding author.
E-mail address: thomas.schmidt@uni-mannheim.de (T. Schmidt).

https://doi.org/10.1016/j.dib.2024.110274

historical currency symbols (ℳ, ℔) among others. The dataset is provided in two variants in PAGE XML format. The first one contains ground truth data with table regions transformed to text regions for easier processing. The second variant preserves all table regions. Researchers can reuse this dataset to train new or finetune existing text recognition or layout segmentation models. The dataset can also be used to evaluate the accuracy of existing OCR models. Using specific, community driven transcription guidelines our dataset is easily interoperable and reusable with other datasets based on the same transcription level.

Specifications Table

Subject	Data Science
Specific subject area	Ground truth data for Optical Character Recognition (OCR) training and evaluation
Data format	Raw, Filtered
Type of data	Text, Image
Data collection	We collected 101 raw images of the corresponding newspaper pages using the digital edition of "Deutscher Reichsanzeiger und Preußischer Staatsanzeiger" [1], shared under the Public Domain Mark 1.0 license. Using the transcription software Transkribus [2], we created an automatic layout segmentation, that was manually corrected afterwards. We then extracted text lines for all pages using the open-source OCR engine Tesseract [3]. The extracted text lines were manually corrected using Transkribus. Each page was corrected twice by different transcribers to assure high data quality.
Data source location	Source of collected images: https://doi.org/10.7801/REICHSANZEIGER
Data accessibility	Repository name: Zenodo Data identification number: https://doi.org/10.5281/zenodo.10144094 Instructions for accessing these data: Images: The Zenodo repository contains a bash script ("download_images.sh") for downloading all images via CLI. The script can be found in the "data" folder of the repository. Transcriptions in PAGE XML format (with table regions): Transcriptions with annotated table regions can be found in the folder "reichsanzeiger-1820–1939_with-TableRegion" of the Zenodo repository. Transcriptions in PAGE XML format (with table regions transformed to text regions): Transcriptions with annotated table regions transformed into text regions can be found in the folder "reichsanzeiger-1820–1939″ of the Zenodo repository.

1. Value of the Data

- The dataset [4] is valuable for training and evaluating text recognition models for historical typefonts (Fraktur and Black Letter). The data captures historical printing conditions, e.g., the use of historical glyphs like long s (ſ), rotunda r (ꝛ), combining Latin small letter E for old German Umlaut (ᵉ), and currency symbols (ℳ, ℔). These historical glyphs can be trained with the dataset to improve their recognition in different historical documents.
- The dataset is valuable for training layout segmentation models for historical domains. Especially the segmentation of table regions proves to be a problem for state-of-the-art OCR models due to the heterogeneity of table structures and their varied layouts [5]. Our dataset provides 412 table regions to further train and evaluate layout segmentation models that focus on recognizing tabular data in historical domains.
- The dataset was created during the third phase of the OCR-D project [6]. During the project, transcription guidelines were developed together with the OCR-D community to ensure the interoperability and the reusability of ground truth datasets using three different transcrip-

tion levels [7]. The transcriptions of our dataset rely on level 2 of the OCR-D Ground Truth Guidelines and can therefore be combined and reused with other datasets that follow the same transcription conventions.
- The dataset provides full texts for historical newspapers in very high quality. For this reason, it can serve as an ideal source for NLP workflows.

2. Background

The heightened interest in historical newspapers from the First (1914–18) and Second World War (1939–45), which offer daily news on economics, science, societal opinions and war events, underscores the importance of robust digitization. Despite the abundance of digitized newspapers, challenges persist in extracting high quality full texts, particularly for early 20th century glyphs like currency symbols and long (ſ), an omnipresent glyph in German books, periodicals and documents printed in Fraktur. Tabular data extraction from newspaper articles faces a similar challenge due to the scarcity of available datasets that can be used to train robust layout segmentation models. These challenges form the background for our dataset, which concentrates on the accurate transcription of historical newspapers and the representation of tabular data.

3. Data Description

Source images: The dataset [4] provides 101 files in PAGE XML format that capture layout segmentation and transcriptions for 101 source images. The source images are scans of newspaper pages of the historical newspaper "Deutscher Reichsanzeiger und Preußischer Staatsanzeiger" (German Imperial Gazette and Prussian Official Gazette), which was published from 1819 to 1945 under six different names and published as a digital edition in 2016 (shared under the Public Domain Mark 1.0 license) [1,8]. The dataset does not include the source images directly but rather enables the download of all images by using the bash script "download_images.sh", located in the "data" folder of the Zenodo repository.

Ground truth in PAGE XML format: The layout segmentation and transcriptions matching the source images are provided as individual PAGE XML [9] files in the "data" folder. For each of the 101 source images, there is a PAGE XML file available in two different variants. These variants are:

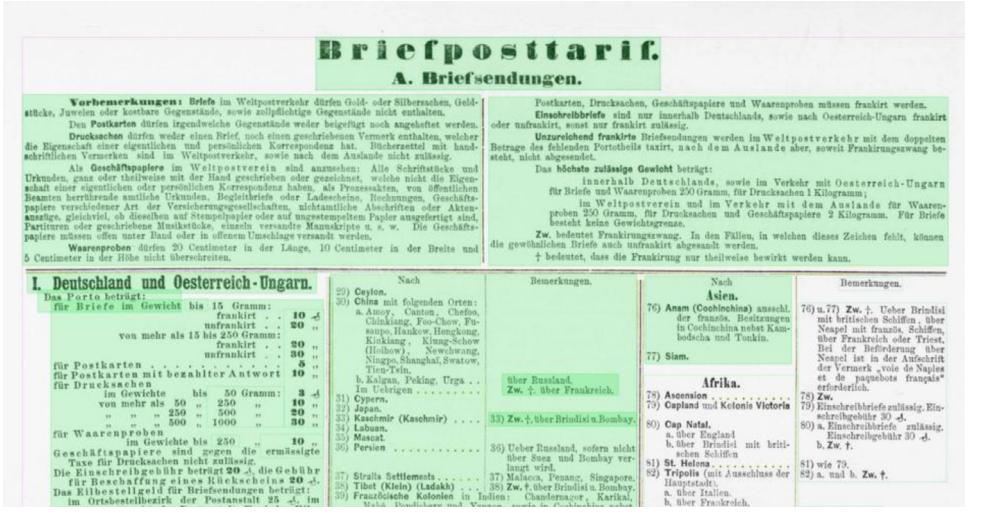
Variant 1: Ground truth in PAGE XML format (with table regions): The folder "reichsanzeiger-1820–1939_with-TableRegion" provides all 101 PAGE XML files with annotated table regions.

Variant 2: Ground truth in PAGE XML format (with table regions transformed to text regions): The folder "reichsanzeiger-1820–1939" provides all 101 PAGE XML files without table regions. All existing table regions were transformed to text regions using Transkribus' built in "Transform tables to region" function. This function converts tables in such a way that each table cell is transformed to an individual text region. Hence the significant difference between the text region counts in variant 1 and 2.

Documentation: Further project documentation can be found at GitHub [10]. Metadata, containing project details, staff, transcription guidelines and sources can be found in the "METADATA.yml" of the Zenodo repository [4]. Overall statistics (e.g., glyph distributions for the whole dataset) can be found at GitHub as well [11].

4. Experimental Design, Materials and Methods

Image selection: Out of the 361,713 available scans at [1] we manually selected 96 double and 5 single newspaper pages for the years 1820–1939. The pages were selected for the following reasons: 1) A representative page layout with common layout components of historical



Briefposttarif.

A. Briefsendungen.

Vorbemerkungen: **Briefe** im Weltpostverkehr dürfen Gold- oder Silbersachen, Geldstücke, Juwelen oder kostbare Gegenstände, sowie zollpflichtige Gegenstände nicht enthalten.

Den **Postkarten** dürfen irgendwelche Gegenstände weder beigefügt noch angeheftet werden.

Drucksachen dürfen weder einen Brief, noch einen geschriebenen Vermerk enthalten, welcher die Eigenschaft einer eigentlichen und persönlichen Korrespondenz hat. Bücherzettel mit handschriftlichen Vermerken sind im Weltpostverkehr, sowie nach dem Auslande nicht zulässig.

Als **Geschäftspapiere** im Weltpostverein sind anzusehen: Alle Schriftstücke und Urkunden, ganz oder theilweise mit der Hand geschrieben oder gezeichnet, welche nicht die Eigenschaft einer eigentlichen oder persönlichen Korrespondenz haben, als Prozessakten, von öffentlichen Beamten herrührende amtliche Urkunden, Begleitbriefe oder Ladescheine, Rechnungen, Geschäftspapiere verschiedener Art der Versicherungsgesellschaften, nichtamtliche Abschriften oder Aktenauszüge, gleichviel, ob dieselben auf Stempelpapier oder auf ungestempeltem Papier angefertigt sind, Partituren oder geschriebene Musikstücke, einzeln versandte Manuskripte u. s. w. Die Geschäftspapiere müssen offen unter Band oder in offenem Umschlage versandt werden.

Waarenproben dürfen 20 Centimeter in der Länge, 10 Centimeter in der Breite und 5 Centimeter in der Höhe nicht überschreiten.

Postkarten, Drucksachen, Geschäftspapiere und Waarenproben müssen frankirt werden.

Einschreibbriefe sind nur innerhalb Deutschlands, sowie nach Oesterreich-Ungarn frankirt oder unfrankirt, sonst nur frankirt zulässig.

Unzureichend frankirte Briefsendungen werden im Weltpostverkehr mit dem doppelten Betrage des fehlenden Portotheils taxirt, nach dem Auslande aber, soweit Frankirungszwang besteht, nicht abgesendet.

Das **höchste zulässige Gewicht** beträgt:

innerhalb Deutschlands, sowie im Verkehr mit Oesterreich-Ungarn für Briefe und Waarenproben 250 Gramm, für Drucksachen 1 Kilogramm;

im Weltpostverein und im Verkehr mit dem Auslande für Waarenproben 250 Gramm, für Drucksachen und Geschäftspapiere 2 Kilogramm. Für Briefe besteht keine Gewichtsgrenze.

Zw. bedeutet Frankirungszwang. In den Fällen, in welchen dieses Zeichen fehlt, können die gewöhnlichen Briefe auch unfrankirt abgesandt werden.

† bedeutet, dass die Frankirung nur theilweise bewirkt werden kann.

I. Deutschland und Oesterreich-Ungarn.

Das Porto beträgt:

für Briefe im Gewicht bis 15 Gramm:		
	frankirt . .	10 ₰
	unfrankirt . .	20 „
von mehr als 15 bis 250 Gramm:		
	frankirt . .	20 „
	unfrankirt . .	30 „
für Postkarten		5 „
für Postkarten mit bezahlter Antwort		10 „
für Drucksachen		
im Gewichte bis 50 Gramm:		3 ₰
von mehr als 50 „ 250 „		10 „
„ „ „ 250 „ 500 „		20 „
„ „ „ 500 „ 1000 „		30 „
für Waarenproben		
im Gewichte bis 250 „		10 „

Geschäftspapiere sind gegen die ermässigte Taxe für Drucksachen nicht zulässig.

Die Einschreibgebühr beträgt 20 ₰, die Gebühr für Beschaffung eines Rückscheins 20 ₰.

Das Eilbestellgeld für Briefsendungen beträgt: im Ortsbestellbezirk der Postanstalt 25 ₰, im

Nach	Bemerkungen.
29) Ceylon.	
30) China mit folgenden Orten: a. Amoy, Canton, Chefoo, Chinkiang, Foo-Chow, Fusuepo, Hankow, Hongkong, Kiukiang, Kiung-Schow (Hoihow), Newchwang, Ningpo, Shanghai, Swatow, Tien-Tsin. b. Kalgan, Peking, Urga . . Im Uebrigen	über Russland. Zw. †. über Frankreich.
31) Cypern.	
32) Japan.	
33) Kaschmir (Kaschmir)	33) Zw. †. über Brindisi u. Bombay.
34) Labuan.	
35) Muscat.	
36) Persien	36) Ueber Russland, sofern nicht über Suez und Bombay verlangt wird.
37) Straits Settlements	37) Malacca, Penang, Singapore.
38) Tibet (Klein) (Ladakh) . . .	38) Zw. †. über Brindisi u. Bombay.
39) Französische Kolonien in Indien:	Chandernagor, Karikal,

Nach	Bemerkungen.
Asien.	
76) Anam (Cochinchina) ausschl. der französ. Besitzungen in Cochinchina nebst Kambodscha und Tonkin.	
77) Siam.	
Afrika.	
78) Ascension	78) Zw.
79) Capland und Kolonis Victoria	79) Einschreibbriefe zulässig. Einschreibgebühr 30 ₰.
80) Cap Natal. a. über England b. über Brindisi mit britischen Schiffen	80) a. Einschreibbriefe zulässig. Einschreibgebühr 30 ₰. b. Zw. †.
81) St. Helena	81) wie 79.
82) Tripolis (mit Ausschluss der Hauptstadt). a. über Italien. b. über Frankreich.	82) a. und b. Zw. †.

Fig. 1. A newspaper page after applying Transkribus' block detection. Some of the automatically created text regions are selected and highlighted in green. Especially the tabular data is recognized poorly compared to Fig. 2, as individual columns, rows and cells are frequently represented by single and/or overlapping text regions.

newspapers, i.e., header, headings, text paragraphs, tables, and lists. 2) A wide range of different knowledge domains, e.g., politics, economics, culture, and official announcements. 3) A representative time period that covers both the changes in fonts and the changes in newspaper layout and contents.

Layout segmentation: Due to layout complexities we applied a granular workflow, using Transkribus [2] and Tesseract [3] as software tools to create the layout segmentation for all 101 pages. 1) Using Transkribus' built in block detection, which is part of the software's layout analysis tools, we automatically generated text regions for all pages. 2) These automatically generated text regions were corrected manually, following guidelines that were based on the OCR-D Ground Truth Guidelines [7] as well as an iteratively adapted and project specific ruleset [12] so that the layout components of the newspaper page (headlines, paragraphs, etc.) could be captured as accurately as possible. Tables in particular were inadequately captured by Transkribus' block detection, which is why all table regions were created manually (see Fig. 1 and 2). 3) We then annotated all regions using a set of 5 structure types: header, heading, paragraph, table, reference [13]. 4) Using Tesseract, we automatically generated bounding boxes and baselines for all existing text lines, which were 5) manually corrected again (see Fig. 3 and 4). 6) We corrected the reading order for all regions.

Transcriptions: 1) After the finished layout segmentation we extracted full texts using Tesseract with frak2021 model trained by Mannheim University Library [15]. 2) The automatically extracted full texts were manually corrected in two runs by a team of four transcribers using Transkribus. Transcriber 1 corrected the transcriptions generated by Tesseract, while transcriber 2 corrected the manual corrections made by transcriber 1. Throughout this process the OCR-D Ground Truth Guidelines level 2 were used as transcription guidelines, since level 2 reproduces "the technical printing conditions […]", while the "interpretation of signs is oriented towards their use in the language and writing system" [7]. Therefore, our dataset does not normalize historical glyphs like long s (ſ) or rotunda r (ꝛ) or the double oblique hyphen (⸗), commonly used to hyphenate words in historical texts typeset in Fraktur, to their modern equivalents like round s, normal r or a standard hyphen (-). In three special cases, we deviated from the OCR-D guidelines in order to capture certain glyphs true to the original. These cases include double oblique hyphen (⸗), em dash (—) instead of en dash (–), and asterisk (*) used for

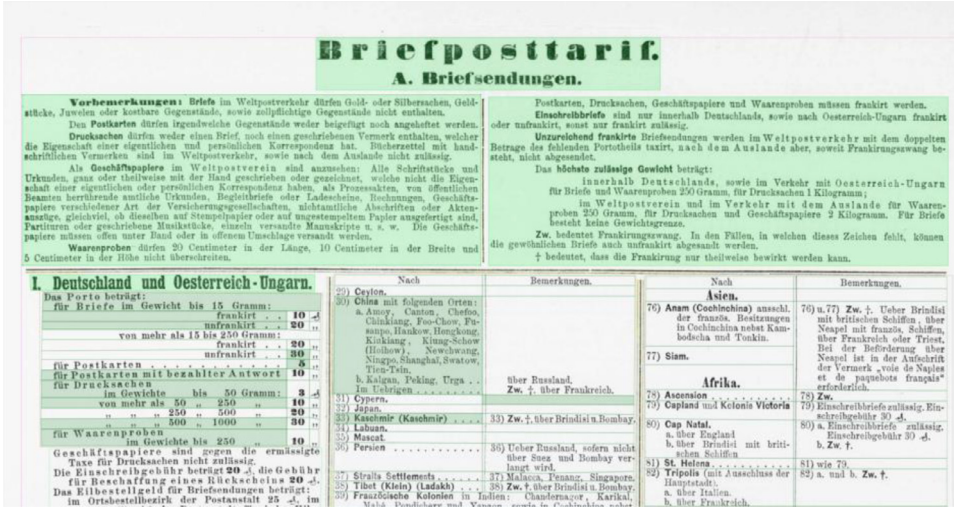


Fig. 2. The same page as in Fig. 1 with manually created table regions. Some of the table cells were selected to highlight the correct representation of the given tabular data.

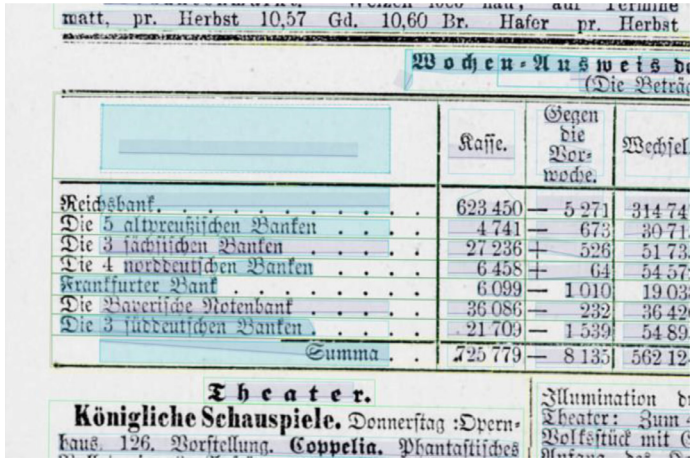


Fig. 3. A newspaper page with incorrectly captured text line bounding boxes highlighted in blue.

both standard asterisk (*) and tear-drop asterisk (⁎) [14]. 3) Finally, we used the finished transcriptions from transcriber 2 to finetune a PyLaia model in Transkribus. Ensuring that the model avoided overfitting to the training material, it was then utilized to identify discrepancies within the finished transcriptions.

Post processing: After finishing the layout segmentation and transcriptions we exported two variants of all 101 pages from Transkribus in PAGE XML format [9]. Variant 1 contains all manually created table regions as table regions (cf. Table 1). Variant 2 contains all table regions transformed to text regions for easier processing (cf. Table 2).


matt. pr. Herbst 10,57 Gd. 10,60 Br. Hafer pr. Herbst

Wochen-Ausweis

	Kasse.	Gegen die Vorwoche.	Wechsel
Reichsbank	623 450	— 5 271	314 74
Die 5 altpreußischen Banken	4 741	— 673	30 71
Die 3 sächsischen Banken	27 236	+ 526	51 73
Die 4 norddeutschen Banken	6 458	+ 64	54 57
Frankfurter Bank	6 099	— 1 010	19 03
Die Bayerische Notenbank	36 086	— 232	36 42
Die 3 süddeutschen Banken	21 709	— 1 539	54 89
Summa	725 779	— 8 135	562 12

Theater.

Königliche Schauspiele. Donnerstag: Opernhaus. 126. Vorstellung. Coppelia. Phantastisches

Illumination

Theater: Zum

Volksstück mit

Fig. 4. The same page as in Fig. 3 with manually corrected text line bounding boxes.

Table 1
Number of regions, lines, words and glyphs for variant 1: Ground truth in PAGE XML format (with table regions).

Text regions	Table regions	Text lines	Words	Glyphs
4491	412	119,430	490,679	2967,330

Table 2
Number of regions, lines, words and glyphs for variant 2: Ground truth in PAGE XML format (without table regions).

Text regions	Table regions	Text lines	Words	Glyphs
65,563	0	119,430	490,679	2967,330

Limitations

The table regions annotated in the dataset largely contain economic data, as tabular data was of particular interest for this project. The first half of the 19th century is underrepresented in comparison to the second half of the century, as only microfiches of poor quality were available for this period. Although the dataset also includes layout regions typical for (historical) newspapers, such as adverts, these are underrepresented in comparison to text regions. Furthermore, the dataset is based on a single newspaper, whose temporal changes in content, layout and typeface are captured by the dataset, but are not representative of the newspaper landscape of the 19th and early 20th century.

The structural annotation of the layout regions is currently rather coarse, as the ground truth was created as part of the OCR-D project, which main goal is to “facilitate research access” to the “Union Catalogue of Books of the 16th–18th century (VD 16, VD 17, VD 18)” [16]. As the OCR-D project is primarily focused on printed books, the ground truth guidelines developed during the project reflect this focus accordingly and do not cover detailed descriptions of structural layout elements found in newspapers. However, as soon as the OCR-D ground truth guidelines are extended for printed documents like newspapers they can be applied to the dataset in order to enhance the current structural annotations.

Ethics Statement

The authors have read the ethical requirements for publication in Data in Brief and hereby affirm that the work did not involve the use of human subjects, animal experiments and/or data collected from social media platforms. The authors did not need permission to use the primary data as it is published under Public Domain Mark 1.0 [17].

Data Availability

reichsanzeiger-gt (Original data) (Zenodo).

CRediT Author Statement

Thomas Schmidt: Project administration, Validation, Data curation, Supervision, Writing – original draft; **Jan Kamlah:** Conceptualization, Methodology, Software, Validation, Data curation, Supervision, Project administration, Writing – review & editing; **Stefan Weil:** Software, Project administration, Writing – review & editing.

Acknowledgements

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grant number 460547474 (https://gepris.dfg.de/gepris/projekt/460547474) and partially by grant number 460037581 (https://gepris.dfg.de/gepris/projekt/460037581).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Weil, "Deutscher Reichsanzeiger und Preußischer Staatsanzeiger" (digital edition), Universitätsbibliothek Mannheim, https://doi.org/10.7801/REICHSANZEIGER.
[2] https://readcoop.eu/transkribus/, last Accessed 26 January 2024.
[3] https://github.com/tesseract-ocr/tesseract, last Accessed 26 January 2024.
[4] T. Schmidt, J. Kamlah, S. Weil, R. Shigapov, UB-Mannheim/reichsanzeiger-gt: 1.0.0, Zenodo (2023), doi:10.5281/zenodo.10144094.
[5] K.A. Hashmi, et al., Current Status and Performance Analysis of Table Recognition in Document Images with Deep Neural Networks, 2021, pp. 3–4, doi:10.48550/arXiv.2104.14272.
[6] https://ocr-d.de/en/about, last Accessed 26 January 2024.
[7] https://ocr-d.de/en/gt-guidelines/trans/level_2_2.html, last Accessed 26 January 2024.
[8] C. Kling, Deutscher Reichsanzeiger und Preußischer Staatsanzeiger: Einleitung zur Veröffentlichung der Digitalausgabe, Report in MADOC (2016) https://ub-madoc.bib.uni-mannheim.de/41378.
[9] S. Pletschacher, A. Antonacopoulos, The PAGE (Page Analysis and Ground-Truth Elements) Format Framework, in: 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 2010, pp. 257–260, doi:10.1109/ICPR.2010.72.
[10] https://github.com/UB-Mannheim/reichsanzeiger-gt, last Accessed 26 January 2024.
[11] https://github.com/UB-Mannheim/reichsanzeiger-gt/wiki/Statistics, last Accessed 26 January 2024.
[12] J. Kamlah, T. Schmidt, "Transkriptionsregeln und Guidelines zur Layoutbearbeitung im DFG-Projekt 'Workflow für werkspezifisches Training auf Basis generischer Modelle mit OCR-D sowie Ground-Truth-Aufwertung'", 2023, https://doi.org/10.5281/zenodo.10203335.
[13] Cf. Kamlah and Schmidt, "Transkriptionsregeln", pp. 32–33.
[14] A full description of all captured special glyphs can be found in the file "METADATA.YML" here: https://github.com/UB-Mannheim/reichsanzeiger-gt/blob/main/METADATA.yml, last Accessed 7 November 2023.
[15] S. Weil, "Tesseract OCR models for historic prints based on Latin script", Vers. 1, 2021, doi:10.5281/zenodo.10125246.
[16] https://ocr-d.de/en/about.html, last Accessed 26 January 2024.
[17] https://digi.bib.uni-mannheim.de/service/, last Accessed 26 January 2024.