ABSCHLUSSBERICHT

1 Allgemeine Angaben

DFG-Geschäftszeichen: *GE 3075/7-1*

Projektnummer: *422758594*

Titel des Projekts:
Strukturdatenerfassung, ergänzende Digitalisierung und OCR-Erschließung des Deutschen Reichsanzeigers und Preußischen Staatsanzeigers (1871–1945)

Name der Antragstellenden: *Ltd. Bibliotheksdirektorin Dr. Sabine Gehrlein*

Dienstanschrift:
Universitätsbibliothek Mannheim, Schloss Schneckenhof West, 68131 Mannheim

Name(n) der Mitverantwortlichen: –

Name(n) der Kooperationspartnerinnen und -partner: –

Berichtszeitraum (gesamte Förderdauer): *10.2019 – 08.2024 und zusätzlich 09.2024*

2 Zusammenfassung/Summary

Als Ergebnis des Projekts liegt nun eine frei zugängliche digitale Ausgabe des Deutschen Reichsanzeigers und des Preußischen Staatsanzeigers in einer Qualität vor, die deutlich besser ist als die bisherige provisorische digitale Ausgabe auf Mikrofilmbasis.

Diese digitale Ausgabe wird über die Digitalen Sammlungen der UB Mannheim angeboten und ist darüber hinaus in das Deutsche Zeitungsportal integriert. Über die komfortable Weboberfläche des Zeitungsportals kann der Reichsanzeiger auch im Kontext anderer Zeitungen recherchiert und genutzt werden.

Allerdings hat die UB Mannheim den Aufwand unterschätzt, der mit der qualitätsorientierten Digitalisierung einer so umfangreichen Zeitung verbunden ist. Es stellte sich heraus, dass wesentlich mehr „ergänzende Digitalisierung“ notwendig war als geplant, da die Mikrofilme oft unvollständig oder von schlechter Qualität waren. Statt nur die Lücken zu digitalisieren, hat die UB Mannheim daher die Zeitungsbände, die zumeist aus anderen Bibliotheken entliehen waren, komplett neu gescannt. Dies kommt der neuen digitalen Edition sehr zugute und war für viele Bände vielleicht die letzte Möglichkeit, da der Zustand des Zeitungspapiers bereits das Scannen zu einer Herausforderung machte.

Dank zweimaliger Verlängerung der Projektlaufzeit konnten bisher trotz des deutlich erhöhten Aufwands die Jahrgänge ab 1900 und einige Jahrgänge des 19. Jahrhunderts in der angestrebten Qualität bereitgestellt werden. Für die meisten älteren Jahrgänge liegen vorerst die Zeitungsausgaben auf Mikrofilmbasis vor – zwar in Einzelseiten zerlegt, aber mit allen Fehlern der Ausgangsbasis. Die UB Mannheim wird auch nach dem offiziellen Projektende weiterhin das Ziel verfolgen, die digitale Edition auf ein einheitlich gutes Niveau zu bringen.

Summary

As a result of the project, a freely accessible digital edition of the “Deutscher Reichsanzeiger und Preußischer Staatsanzeiger” is now available in a quality that is significantly better than the previous provisional digital edition on a microfilm basis.

This digital edition is offered via the Digital Collections of the Mannheim University Library and is also integrated into the German Newspaper Portal. The Reichsanzeiger can also be researched and used in the context of other newspapers via the convenient web interface of the newspaper portal.

However, Mannheim University Library underestimated the effort involved in the quality-oriented digitisation of such an extensive newspaper. It turned out that much more ‘supplementary digitisation’ was necessary than planned, as the microfilms were often incomplete or of poor quality. Instead of just digitising the gaps, Mannheim University Library has therefore completely rescanned the newspaper volumes, most of which were borrowed from other libraries. This greatly benefits the new digital edition and was perhaps the last option for many volumes, as the condition of the newsprint already made scanning a challenge.

Thanks to two extensions of the project duration, the volumes from 1900 onwards and some volumes from the 19th century have so far been made available in the desired quality despite the significantly increased effort involved. For most of the older volumes, the newspaper editions are available on microfilm for the time being – broken down into individual pages, but with all the errors of the original. Even after the official end of the project, Mannheim University Library will continue to pursue the goal of bringing the digital edition to a uniformly good standard.

3 Arbeits- und Ergebnisbericht

3.1 Ausgangslage und Zielsetzung des Projektes

Im November 2015 stellte die UB Mannheim eine provisorische digitale Edition des Reichsanzeigers[1] erstmals online. Sie zeigte Doppelseiten aus digitalisierten Mikrofilmen, die ab 2016 um experimentelle OCR-Ergebnisse ergänzt wurden. Ein mögliches Leistungsschutzrecht des Mikrofilmherstellers schränkte die Nutzung auf „wissenschaftliche Zwecke“ ein.

Diese provisorische Onlineausgabe mit ihren Beschränkungen sollte durch eine qualitätsgeprüfte und frei zugängliche digitale Edition abgelöst werden. Die UB Mannheim verwendet die Digitalisierungssoftware Kitodo für ihre Digitalisate. Auch der Reichsanzeiger sollte mit Kitodo.Production bearbeitet und mit Kitodo.Presentation präsentiert werden und so gängige Standards erfüllen.

3.2 Arbeitsschritte im Berichtszeitraum

Die Arbeitsschritte werden zunächst gemäß der Arbeitspakete, wie sie im Projektantrag vorgesehen waren, beschrieben und anschließend um zusätzlich angefallene Schritte ergänzt.

1 Stefan Weil, 126 Jahre Zeitung online - Fundgrube für historisch Interessierte und Motor für die Bibliotheks-IT. Konferenzveröffentlichung (2018). https://nbn-resolving.org/urn:nbn:de:0290-opus4-36739

3.2.1 AP 1: Anlegen der Vorgänge in Kitodo

Für dieses Arbeitspaket waren im Antrag 148 Stunden geplant. Die Aufwandsabschätzung war realistisch, aber es waren bibliothekarische Kenntnisse notwendig, so dass dieses Arbeitspaket nicht von Hilfskräften, sondern von Mitarbeiterinnen der Universitätsbibliothek umgesetzt werden musste.

3.2.2 AP 2: Qualitätssicherung der gescannten Mikroverfilmung

Bei der Qualitätssicherung der digitalisierten Mikrofilme kamen zeitweise – während der COVID-19-Pandemie und der damit verbundenen Schließung der Universität Mannheim im Frühjahr 2020 – viele zusätzliche Hilfskräfte, die aus eigenen Mitteln finanziert wurden, zum Einsatz. Natürlich war damit auch ein erheblicher organisatorischer Aufwand verbunden, und nicht alle Arbeitsergebnisse waren verwertbar. In Summe konnten aber alle 22.609 Ausgaben identifiziert und – soweit vorhanden – geprüft werden. Die dank eigener Hilfskräfte eingesparten Projektmittel wurden für Arbeitspaket 3 eingesetzt.

3.2.3 AP 3 : Digitalisierung der fehlenden und fehlerhaften Teile

Die Qualitätssicherung der gescannten Mikroverfilmung zeigte schon früh im Projektverlauf, dass die Abschätzung im Projektantrag mit 57.000 neu zu scannenden Seiten viel zu niedrig angesetzt war. Tatsächlich wurden im Projekt bisher (Stand 24.09.2024) 369.214 Seiten neu gescannt, da es mehr fehlende oder qualitativ schlechte Seiten als erwartet gab und da es weder wünschenswert noch praktikabel war, in einem gebundenen Zeitungsband nur vielleicht jede 10. Seite zu scannen.

Die UB Mannheim besaß nur wenige Zeitungsbände und musste daher schon vor Projektstart und während der Projektlaufzeit Bände per Fernleihe von UB Freiburg, BSB München, UB der LMU München, UB Tübingen und WLB Stuttgart besorgen. Hinzu kamen umfangreiche Schenkungen von Zeitungsbänden aus der Bibliothek des Geheimen Staatsarchivs Preußischer Kulturbesitz Berlin und aus der Bibliothek des ZBW – Leibniz-Informationszentrum Wirtschaft Hamburg.

Die neuen Farbscans in hoher Auflösung (300 und 400 DPI) sind qualitativ wesentlich besser als die schwarz-weißen Mikrofilme aus den siebziger Jahren, obwohl bei vielen Zeitungsbänden das Papier bereits sehr brüchig ist, was das Scannen sehr erschwert. Sie ermöglichen eine bessere Texterkennung, sind angenehmer zu lesen und insbesondere vollständiger, da dabei verdeckter Text im Falz oder durch Knicke im Papier und falsche Belichtung vermieden wurde. Es war klar, dass das Projekt die wahrscheinlich letzte Chance für gute Digitalisate war. Daher wurden statt einzelner Seiten die ganzen Jahrgänge von

Zeitungsausgaben neu gescannt. Dadurch werden auch unschöne Qualitätssprünge zwischen mikroverfilmten und neu gescannten Seiten innerhalb einer Zeitungsausgabe vermieden.

Zum Einsatz kam dabei anfangs ein Zeutschel OS 12000 (Auflösung 300 DPI) und später zwei Microbox book2net Ultra Scanner (Auflösung 400 DPI), die den ursprünglich vorgesehenen Qidenus V-Scanner ersetzten. Von rund 447 Tsd. Seiten stammen 78 Tsd. aus der Mikroverfilmung. 11 Tsd. sind in 300 DPI und 358 Tsd. in 400 DPI neu gescannt.

3.2.4 AP 4: Strukturdatenerfassung des Reichsanzeigers

Anzahl der Strukturelemente	Häufigkeit (Ausgaben)	Anzahl der Strukturelemente	Häufigkeit (Ausgaben)
1	1453	20	296
3	6	21	201
4	6	22	106
5	4	23	83
6	19	24	45
7	145	25	32
8	655	26	18
9	731	27	9
10	1100	28	7
11	1502	29	4
12	1762	30	2
13	1767	31	2
14	1534	32	2
15	1245	37	1
16	951	38	1
17	707	40	1
18	518		
19	353		

Tab. Strukturelemente pro Ausgabe

Die Strukturdatenerfassung kostete sehr viel Zeit, insbesondere weil sie zunächst wie vom Antrag vorgesehen bis auf Artikelebene erfolgte. Schon sehr früh wurde der Nutzen dieser Detaillierung hinterfragt, denn viele Absatzüberschriften wiederholten sich ständig und boten keinen wirklichen Mehrwert. Es war zu erwarten, dass Nutzende eher im Volltext suchen als in Strukturbäumen. Beim ersten Einspielen der Daten in das Deutsche Zeitungsportal zeigte sich auch noch, dass dort so ausführliche Daten gar nicht unterstützt wurden und dass praktisch niemand sonst diesen Aufwand treibt. Daher sind nur die Jahrgänge 1925 bis 1945

mit Strukturdaten bis auf Artikelebene erfasst. Für frühere Jahrgänge beschränkte sich die Erfassung auf Ausgaben und Beilagen. Dadurch sank die Anzahl der Strukturelemente pro Ausgabe von teilweise mehr als 20 (Spitzenwert 40) auf durchschnittlich noch 10. Die meisten Ausgaben vor 1900 enthalten noch keine gültigen Unterstrukturdaten, soweit sie automatisiert auf Basis der Mikrofilme eingespielt wurden.

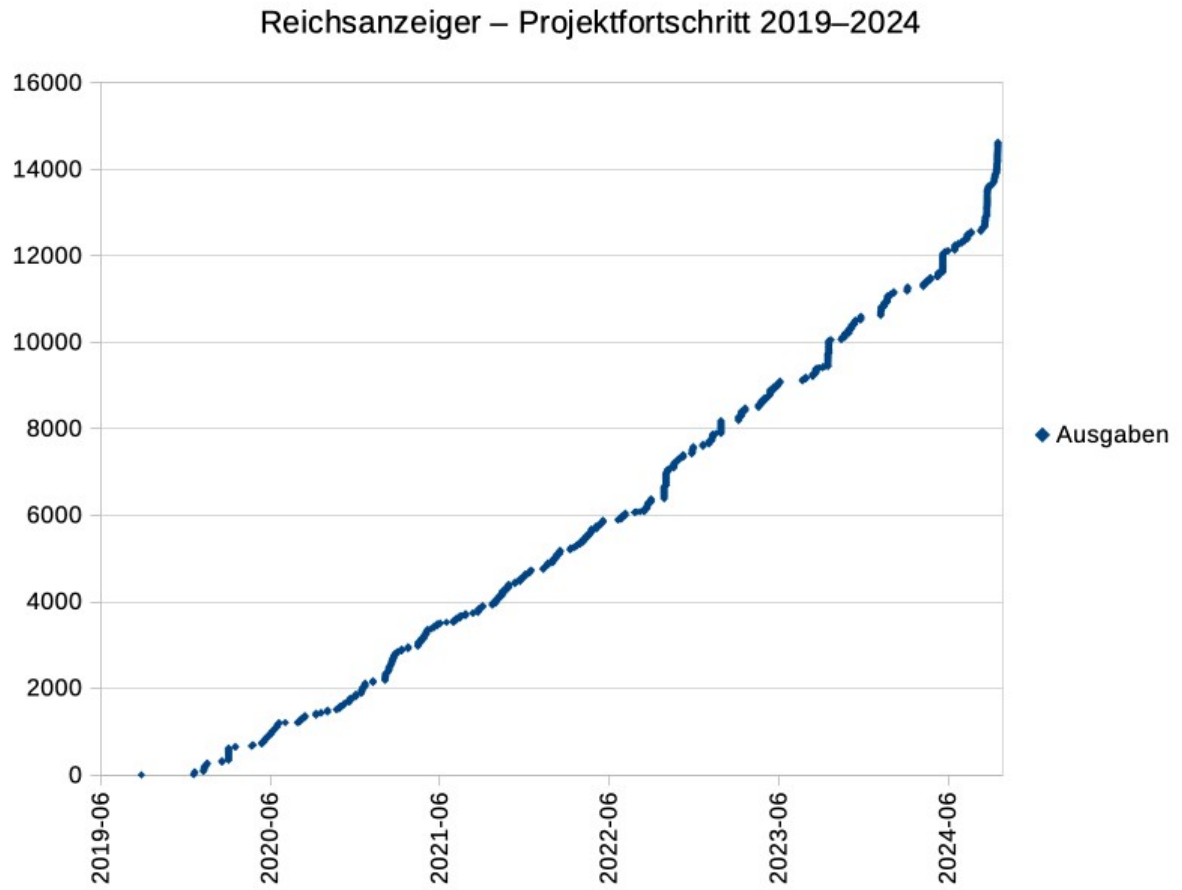

Abb. Publizierte Zeitungsausgaben im Projektzeitraum

Die graphische Darstellung zeigt den fast linearen Projektfortschritt – trotz Pandemie. Im September 2024 wurde begonnen, noch fehlende Jahrgänge des 19. Jahrhunderts durch Mikrofilmdigitalisate ohne neue Scans für fehlerhafte oder fehlende Teile bereitzustellen, was den steilen Anstieg gegen Ende der Kurve erklärt.

3.2.5 AP 5: Generierung und Qualitätskontrolle des Volltextes

Alle Volltexte wurden mit der OCR-Software Tesseract erzeugt. Diese lag bei der Antragstellung in der Version 4.0.0-beta vor. Gegen Ende des offiziellen Projektabschlusses war die Version 5.4.1 (Juni 2024) aktuell. Die UB Mannheim war dabei an der Weiterentwicklung

von Tesseract führend beteiligt: Stefan Weil, der Leiter der Abteilung Digitale Bibliotheksdienste, hat alle Softwareversionen seit 2020 herausgegeben. Sie konnte dank der Förderung durch die DFG im Projekt *Tesseract als Komponente im OCR-D Workflow*[2] die Software modernisieren, die Qualität des Codes auf einen sehr guten Stand bringen und die Geschwindigkeit der Texterkennung nahezu verdoppeln. In einem weiteren DFG-geförderten Projekt *Workflow für werkspezifisches Training auf Basis generischer Modelle mit OCR-D sowie Ground-Truth-Aufwertung*[3] hat sie u. a. das zusätzliche Ausgabeformat PAGE XML ergänzt. Darüber hinaus konnte sie auch den Trainingsprozess ähnlich wie zuvor die Texterkennung wesentlich beschleunigen. Das war sehr hilfreich für das Training neuer Tesseract-Modelle für die Erkennung historischer Schriften.

Die aktuellen Versionen von Tesseract sind ebenso wie die an der UB Mannheim trainierten Tesseract-Modelle ganz zentrale Komponenten von OCR-D und werden auch von anderen Einrichtungen – beispielsweise beim MARCHIVUM Mannheim, an der ULB Halle oder an der TU Darmstadt – für deren Zeitungsdigitalisierung eingesetzt. Erstmals übertraf im November 2019 das Tesseract-Modell *Fraktur5000000* die kommerzielle Software ABBYY Finereader bei der Erkennung von Frakturtexten.[4] Spätere Modelle wie das bisher im Projekt verwendete *frak2021* und das 2023/24 trainierte *german_print* brachten weitere Verbesserungen. Typische Zeichenerkennungsraten mit diesen Modellen liegen über 97 %, bei sehr gut gescannten Textabsätzen werden teilweise 99 % übertroffen. Während 2017 erste OCR-Ergebnisse zwar schon eine Volltextsuche ermöglichten und man zumindest den Inhalt eines Artikels erahnen konnte, sind inzwischen zwar nicht alle, aber doch viele der erzeugten Text fast fehlerfrei und somit auch für die linguistische Forschung nutzbar.

Die Entscheidung für eine Texterkennung nur mit Tesseract fiel schon früh, da bekannt war, dass OCR-D dafür wesentlich mehr Ressourcen (Rechenzeit und Speicherbedarf) benötigt hätte. Ein erneuter Vergleich zwischen Tesseract und drei für OCR-D empfohlenen OCR-Arbeitsabläufen[5] im September 2024 bestätigte die früheren Erfahrungen. Tesseract war dabei rund doppelt so schnell wie der schnellste OCR-D-Arbeitsablauf, der selbst intern auf Tesseract basiert und daher wie zu erwarten fast identische Ergebnisse lieferte. Im Vergleich zum „besten“ OCR-Arbeitsablauf war Tesseract mindestens 300mal schneller und erzielte dennoch ähnliche oder sogar bessere Ergebnisse. Der Speicherverbrauch von OCR-D war je nach Arbeitsablauf 8mal bis mehr als 30mal höher.[6]

2 https://nbn-resolving.org/urn:nbn:de:bsz:180-madoc-522282, https://gepris.dfg.de/gepris/projekt/394264782

3 https://nbn-resolving.org/urn:nbn:de:bsz:180-madoc-671740, https://gepris.dfg.de/gepris/projekt/460547474

4 https://nbn-resolving.org/urn:nbn:de:bsz:180-madoc-537486

5 https://ocr-d.de/en/workflows#recommendations

6 https://github.com/UB-Mannheim/Reichsanzeiger/wiki/Text-recognition

Praktisch bedeuten zusätzliche Messergebnisse mit Tesseract, dass es auf dem OCR-Server der UB Mannheim (AMD EPYC 7413 24-Core CPU) eine Zeitungsseite in durchschnittlich 28 s verarbeiten kann, wobei 24 Seiten parallel verarbeitet werden. Da die CPU mit Hyperthreading auch bis zu 48 Seiten parallel verarbeiten kann, wurde bisher oft diese höhere Zahl verwendet. Laut Testergebnis steigt damit aber die Verarbeitungszeit pro Seite auf 40 s. 48 statt 24 Seiten parallel verdoppelt also nicht wie erhofft den Durchsatz, verkürzt aber immerhin die Dauer für 960 Seiten von knapp 20 min auf rund 15 min. 420.000 Seiten im Projekt lassen sich demnach mit Tesseract in rund 6 Tagen verarbeiten, wenn dabei 24 Seiten parallel laufen. Der „beste“ OCR-D-Arbeitsablauf bräuchte dafür auf dem gleichen Server fast 5 Jahre.

In obigem Vergleich konnten die jüngsten Entwicklungen von OCR-D – insbesondere der METS-Server und die Netzwerkimplementierung – noch nicht einfließen. Die Ergebnisse sind auch sicherlich nicht auf beliebige andere Druckmaterialien übertragbar.

3.2.6 AP 6: Projektmanagement

Im Laufe der langen Projektlaufzeit gab es mehrere Personalveränderungen, die dazu führten, dass das Projekt nacheinander drei Projektleiter hatte. Dies und die Covid-19-Pandemie verursachten zusätzlichen Aufwand beim Projektmanagement.

3.2.7 Weitere Arbeitsschritte

Nicht alle für das Projekt angefallenen Arbeiten waren als Arbeitspakete im Projektantrag aufgeführt.

Gleich zu Projektbeginn konnte – wie im Antrag zugesagt – die Zugriffsbeschränkung für die provisorische Digitalausgabe aufgehoben und durch einen freien Zugang ohne Nutzungsbeschränkungen (Public Domain Mark 1.0) ersetzt werden.

Damit die Zeitung in die Digitalen Sammlungen der Universitätsbibliothek aufgenommen werden konnte, war es notwendig, die dort vorhandene Installation von Kitodo.Presentation (Version 2) an die besonderen Anforderungen für Periodika anzupassen. Der Wechsel auf die neuere Version 3 erwies sich schwieriger als erwartet, so dass zum Projektabschluss nur eine noch nicht produktiv geschaltete Testinstallation existierte. Diese soll aber in Kürze freigeschaltet werden. Bei der Arbeit mit der neuen Version konnte der damit betraute Mitarbeiter mit etlichen Fehlermeldungen (issues) und Codebeiträgen (pull requests) zur Weiterentwicklung von Kitodo.Presentation und dem DFG-Viewer beitragen. Synergien gab

es auch durch den Projektmitarbeiter im parallel laufenden DFG-Projekt *Integration von Kitodo und OCR-D zur produktiven Massendigitalisierung.*[7]

Im Dezember 2021 nahm die Projektleitung Kontakt mit der DDB Fachstelle Bibliothek auf, um die Präsentation des Reichsanzeigers auch im Deutschen Zeitungsportal vorzubereiten. Dessen spezielle Anforderungen an die Metadaten machten umfangreiche Datenbereinigungen notwendig. Beim Versuch, im Dezember 2022 erste Daten ins Zeitungsportal einzuspielen, wurden allerdings noch 7614 von 8541 Datensätzen (ein Datensatz entspricht einer Zeitungsausgabe) als fehlerhaft abgewiesen. Im Juli 2023 waren von 9906 Datensätzen nur noch 1053 fehlerhaft, so dass 8853 Datensätze im August 2023 zunächst ins Testsystem und dann auch ins Produktivsystem des Zeitungsportals eingespielt werden konnten. Im November 2023 war bei 13212 Datensätzen nur einer fehlerhaft. Die Datenlieferungen im 1. Quartal 2024 mit 13933 Datensätzen und im 2. Quartal 2024 mit 14520 Datensätzen waren sogar komplett fehlerfrei. Mit der Datenlieferung im 3. Quartal sind zum Projektabschluss 15497 Zeitungsausgaben der UB Mannheim im Deutschen Zeitungsportal zugänglich, davon 12820 Ausgaben des Reichsanzeigers.[8] Leider enthielt diese letzte Lieferung 180 fehlerhafte Datensätze, die aber gleich ebenso wie die internen Abläufe korrigiert wurden, so dass im kommenden Quartal wieder eine fehlerfreie Datenlieferung mit dann rund 16000 Ausgaben des Reichsanzeigers zu erwarten ist.

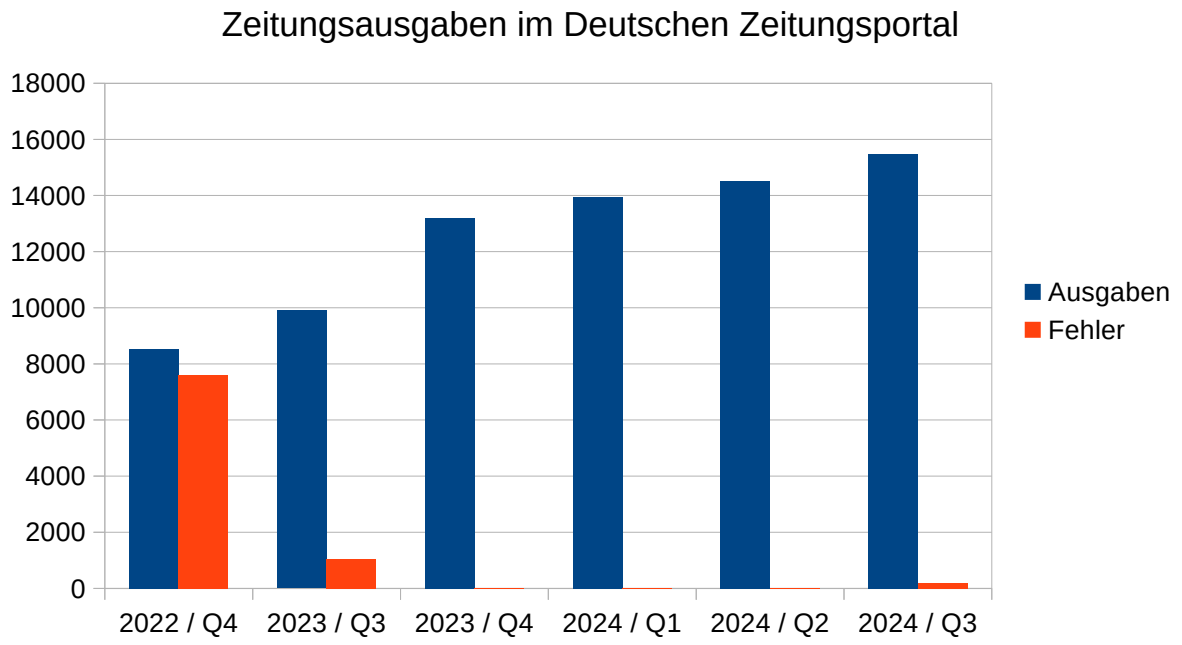

Abb. Zeitungsausgaben im Deutschen Zeitungsportal

7 https://gepris.dfg.de/gepris/projekt/460478737

8 https://www.deutsche-digitale-bibliothek.de/search/newspaper?query=&provider=Universität+Mannheim.+Universitätsbibliothek

Damit Datenfehler schon vor der Lieferung gefunden und korrigiert werden, hat die Universitätsbibliothek Mannheim das Prüfprogramm *validate-mets* für METS-Daten erstellt.[9] Dieses Prüfprogramm prüft u. a. mit den Schematron-Validierungen der Deutschen Digitalen Bibliothek / Fachstelle Bibliothek[10] für Zeitungen und andere digitalisierte Medien, also mit den gleichen Regeln, die auch die DDB selbst anwendet. Auch andere Einrichtungen können das Prüfprogramm frei verwenden, was bisher nur vereinzelt erfolgt, aber auf dem nächsten Kitodo-Anwendertreffen beworben werden soll.

Weitere Skripte für statistische Auswertungen und Routineaufgaben wie beispielsweise das Schneiden der Doppelseiten aus der Mikroverfilmung sind ebenfalls auf GitHub dokumentiert[11].

Die Öffentlichkeitsarbeit für die Digitalausgabe war kein eigenes Arbeitspaket, erfolgte aber in einem Vortrag zum Projekt auf der BiblioCon 2024 und weiteren Vorträgen anderer Projekte der Universitätsbibliothek, die Daten des Reichsanzeigers nutzten. Ein Beispiel dafür ist die linguistische Auswertung der Texte des Reichsanzeigers mit automatischer Erkennung von Entitäten (Reichsanzeiger-NLP/NER).[12] Erfolgreich war die Teilnahme der UB Mannheim als Datengeber beim letzten Hackathon *Coding da Vinci* 2022 in Karlsruhe.[13] Das dort von Studierenden realisierte Projekt *ansights* ermöglicht, Dokumente (beispielsweise Flugblätter aus einem anderen Datenset *Flug- und Extrablätter aus der Revolutionszeit 1918 bis 1920)* inhaltlich automatisiert auszuwerten und diese mit überschneidenden Inhalten des *Deutschen Reichsanzeigers und Preußischen Staatsanzeigers* zu verknüpfen.[14]

9 https://github.com/UB-Mannheim/Reichsanzeiger/tree/master/scripts/validate

10 https://github.com/Deutsche-Digitale-Bibliothek/ddb-metadata-schematron-validation

11 https://github.com/UB-Mannheim/Reichsanzeiger/ und zugehöriges Wiki

12 https://ub-mannheim.github.io/reichsanzeiger-nlp/

13 https://codingdavinci.de/de/daten/deutscher-reichsanzeiger-und-preussischer-staatsanzeiger

14 https://codingdavinci.de/de/projekte/ansights

4 Öffentlich zugängliche Projektergebnisse

4.1 Publikationen mit wissenschaftlicher Qualitätssicherung

Schmidt, Thomas; Kamlah, Jan; Weil, Stefan (2024) *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).* Data in Brief Amsterdam [u. a.] 54 Article 110274 1-7[15]

4.2 Weitere Publikationen und öffentlich gemachte Ergebnisse

Vorträge

Weil, Stefan. *126 Jahre Zeitung online-Fundgrube für historisch Interessierte und Motor für die Bibliotheks-IT.* 107. Deutscher Bibliothekartag 2018 (Berlin)[16]

Weil, Stefan. *Neue Frakturmodelle für Tesseract.* Kitodo Anwendertreffen 2019 (Hamburg)[17]

Weil, Stefan; Gottschling, Tünde. *Fünf Jahre DFG-Projekt Deutscher Reichsanzeiger – was haben wir gelernt?* 112. BiblioCon 2024 (Hamburg)[18]

Quellcode und sonstige öffentlich zugänglichen Projektergebnisse

https://github.com/UB-Mannheim/Reichsanzeiger/
https://github.com/UB-Mannheim/Reichsanzeiger/wiki

15 https://doi.org/10.1016/j.dib.2024.110274
16 https://nbn-resolving.org/urn:nbn:de:0290-opus4-36739
17 https://nbn-resolving.org/urn:nbn:de:bsz:180-madoc-537486
18 https://nbn-resolving.org/urn:nbn:de:0290-opus4-192038