Ausbau und Optimierung des DDB-Zeitungsportals – („DDB-Zeitungsportal V. 2.0“)

Beschreibung des Vorhabens – Projektanträge im Bereich „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ (LIS)

LIS-Förderprogramm: e-Research-Technologien (Vorhaben zur *Konsolidierung und Optimierung* bestehender e-Research-Technologien)

Frank Scholze (Frankfurt am Main), Matthias Razum (Eggenstein-Leopoldshafen), Dr. Achim Bonte (Dresden), Barbara Schneider-Kempf (Berlin)

Beschreibung des Vorhabens (Öffentliche Kurzfassung)

1. Ausgangslage und eigene Vorarbeiten

Bedarf für ein nationales Zeitungsportal/DFG-Projekt „Digitalisierung historischer Zeitungen“

In den Jahren 2013 bis 2016 wurde unter Leitung der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB) und unter Beteiligung weiterer Bibliotheken[1] das DFG-geförderte Pilotprojekt „Digitalisierung historischer Zeitungen“ durchgeführt, in dessen Rahmen auch ein neuer ZDB-Katalog mit erweiterten Suchmöglichkeiten für die Zeitungssuche sowie mit spezifischen Features zur Unterstützung von Digitalisierungsprojekten entwickelt[2] und der DFG-Viewer optimiert wurde.[3] Als weiteres Ergebnis wurde ein Masterplan[4] vorgelegt, der auf die Einrichtung einer DFG-Förderlinie zur großflächigen Unterstützung von Digitalisierungsvorhaben zielt. Darin wird auch der aus der Wissenschaft heraus geäußerte dringende Bedarf für einen zentralen Zugang zu digitalisierten Zeitungen deutlich gemacht – etwa in Form eines nationalen Zeitungsportals.

Für ein nationales Zeitungsportal wurden bei einem Workshop mit WissenschaftlerInnen im Herbst 2014 in Bremen folgende zentrale Anforderungen erhoben: (1) eine übergreifende Volltextsuche in den digitalisierten Zeitungsbeständen, (2) browsingbasierte Einstiegspunkte (etwa über Kalender und Zeitungstitel), (3) eine integrierte Anzeigekomponente (Viewer) mit Funktionen unter anderem für Treffer-Highlighting, Seitenrotation und -zoom und dem nahtlosen Kopieren gefundener Volltextstellen (Copy&Paste) sowie (4) eine konsistente Möglichkeit, auf die enthaltenen Zeitungen bzw. Einzelausgaben persistent zu referenzieren und sie somit zitierfähig zu machen.

Deutsche Digitale Bibliothek (DDB)

Im Rahmen des Pilotprojekts stellte sich daraufhin die Frage nach konkreten Realisierungsoptionen für ein nationales Zeitungsportal für Deutschland. Dabei konzentrierten sich die Überlegungen schnell auf die Deutsche Digitale Bibliothek (DDB).[5] Die Deutsche Digitale Bibliothek verfolgt das Ziel, das Kultur- und Wissenserbe aus Deutschland in digitaler Form zusammenzuführen und über ihre Angebote an unterschiedliche Zielgruppen zu vermitteln. Dazu gehören neben dem DDB-Portal vor allem Anwendungen, die auf der API (Programmierschnittstelle) der DDB basieren. Nach einer etwa anderthalbjährigen Betaphase ging die DDB Ende März 2014 mit einer ersten Vollversion in den Regelbetrieb und konnte sich seither als nationales Nachweis- und Zugangsportal für das digitalisierte Kultur- und Wissenserbe in Deutschland etablieren. 2018 wurde die Finanzierung der DDB durch

---

[1] Die Konsortialpartner des Projekts waren die Staatsbibliothek zu Berlin (SBB), die Staats- und Universitätsbibliothek Bremen (SuUB), die Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB), die Deutsche Nationalbibliothek (DNB), die Universitäts- und Landesbibliothek Halle (ULB) sowie die Bayerische Staatsbibliothek (BSB).
[2] Siehe https://zdb-katalog.de/index.xhtml [07.05.2020].
[3] Siehe https://gepris.dfg.de/gepris/projekt/271857751 [07.05.2020].
[4] Empfehlungen zur Digitalisierung historischer Zeitungen in Deutschland (Masterplan Zeitungsdigitalisierung) – Ergebnisse des DFG-Projektes „Digitalisierung historischer Zeitungen“ Pilotphase 2013–2015, 29.01.2016 (Partner: SBB (Berlin), SuUB (Bremen), SLUB (Dresden), DNB (Frankfurt), ULB (Halle), BSB (München)).
https://www.zeitschriftendatenbank.de/fileadmin/user_upload/ZDB/z/Masterplan.pdf [07.05.2020].
[5] Siehe https://www.deutsche-digitale-bibliothek.de/ [07.05.2020].

Bund und Länder verstetigt, womit die langfristige Perspektive gesichert ist. Derzeit weist die DDB fast 33 Mio. Objekte aus über 420 Einrichtungen aller Kultursparten nach.

Für die DDB als technische und organisatorische Basis eines nationalen Zeitungsportals sprachen insbesondere folgende Erwägungen:

- die bestehende technische und organisatorische Infrastruktur mit erprobten und etablierten Technologien und Prozessen im Bereich der Zusammenführung, Verarbeitung und Darstellung verteilter digitaler Bestände und Sammlungen sowie einer vertrauenswürdigen und stabilen Betriebssituation beim technischen Betreiber der DDB (FIZ Karlsruhe),
- die langfristige Perspektive der DDB als gesamtstaatliches Vorhaben, dessen dauerhafte Weiterführung seitens des Bundes und der Länder als Unterhaltsträger gesichert ist,
- die erklärte Bereitschaft der Verantwortlichen innerhalb der DDB, eine um zeitungsspezifische Komponenten erweiterte technische und organisatorische Infrastruktur dauerhaft zu betreiben,
- die Tatsache, dass für durch öffentliche Förderung digitalisierte Sammlungen und Bestände ohnehin Verpflichtungen bzw. starke Empfehlungen bestehen, diese an die DDB zu liefern, und die DDB bereits jetzt zahlreiche digitalisierte historische Zeitungen aus mehreren teilnehmenden Bibliotheken nachweist.

Die Deutsche Digitale Bibliothek wird getragen von einem Kompetenznetzwerk deutscher Kulturerbe-Einrichtungen mit derzeit 14 Mitgliedern. Vier dieser Mitglieder haben sich als Projektpartner für das DDB-Zeitungsportal zusammengefunden: Der DDB-Geschäftsbereich Technik, Entwicklung, Service mit der übergreifenden Projektkoordination und -steuerung sowie der Servicestelle ist an der **Deutschen Nationalbibliothek** angesiedelt. Hier werden die konzeptionelle Weiterentwicklung und die Betreuung der Datenpartner verantwortet. **FIZ Karlsruhe** ist der technische Betreiber der DDB und darüber hinaus verantwortlich für die Softwareentwicklung der Kernkomponenten der DDB. Die Stiftung Preußischer Kulturbesitz, zu der die **Staatsbibliothek zu Berlin (SBB)** gehört, ist gemäß Verwaltungs- und Finanzabkommen als Trägerin der DDB-Geschäftsstelle bestellt und mit dem Geschäftsbereich Finanzen, Recht, Kommunikation und Marketing betraut. Ihr Präsident, Prof. Dr. Hermann Parzinger, ist Vorstandssprecher der DDB. Die **SLUB Dresden** betreibt die DDB-Fachstelle Mediathek-Bild/Ton und war im Rahmen eines Pilotprojekts bereits mit der prototypischen Realisierung einer auf dem DFG-Viewer basierenden Anzeigekomponente für die DDB betraut.[6]

Der Empfehlung des Masterplans („Die Deutsche Digitale Bibliothek (DDB) sollte so bald als möglich ein nationales Zeitungsportal mit dem Zugang zu allen digitalisierten Zeitungen in Deutschland mit den in diesem Masterplan beschriebenen Features entwickeln“)[7] folgend, haben diese vier Einrichtungen 2017 einen Antrag an die DFG gestellt, zur „Errichtung eines nationalen Zeitungsportals auf der Basis der organisatorischen und technischen Infrastruktur der Deutschen Digitalen Bibliothek (DDB)“. Das Vorhaben war von Anfang an auf vier Jahre angelegt, der DFG-Antrag von 2017 zielte auf eine erste, zweijährige Förderphase im Programm „Implementierung von e-Research-Technologien“. Diese wurde 2018 bewilligt und die Projektarbeit startete im Januar 2019.

Der vorliegende Antrag bezieht sich auf die zweite Projektphase im Programm „Konsolidierung und Optimierung bestehender e-Research-Technologien“, für die idealerweise ab Januar 2021 Fördermittel bereitstehen, sodass die Projektgruppe ihre Arbeiten ohne Unterbrechung fortsetzen kann.

Stand der Arbeiten und Anforderungen der Nutzergruppen

Zum Zeitpunkt der Antragstellung (Mai 2020) sind die Arbeiten der ersten Förderphase noch nicht abgeschlossen, aber schon gute Fortschritte erzielt worden. So sind die wesentlichen konzeptionellen Vorarbeiten (Datenmodell, UI-Konzept, Nutzerbefragung, Design, technische Konzepte u.a.) abgeschlossen. Auf ihrer Grundlage wurde ein Prototyp, also eine erste funktionstüchtige Version des Zeitungsportals erstellt, die Echtdaten von drei Datenpartnern enthält, aber noch nicht für die Öffentlichkeit zugänglich ist. Anhand des Prototyps werden im Lauf der ersten Projektphase weitere Tests an Daten und Software durchgeführt und Verbesserungen vorgenommen.

Die Freischaltung des Zeitungsportals für die Öffentlichkeit ist, wie im ersten Antrag beschrieben, für Ende 2020 geplant. Es wird Bestände von mindestens sechs Datenpartnern enthalten und alle vier

[6] Weitere Informationen zu Vorarbeiten und Kompetenzen der vier Antragsteller in Bezug auf Zeitungsdigitalisierung finden sich im Antrag zur ersten Projektphase.
[7] Siehe Masterplan Zeitungsdigitalisierung, S. 60.

oben genannten Anforderungen (Volltextsuche, integrierter Viewer, browsende Zugänge, stabile Referenzierbarkeit) erfüllen.

Da das Zeitungsportal noch nicht online ist, gibt es noch keine Auswertung zur quantitativen und qualitativen Nutzung. Während der ersten Projektphase wurde allerdings deutlich, dass der Aufbau eines nationalen Zeitungsportals einhellig begrüßt wird. Dies äußerte sich konkret im Rahmen einer Online-Umfrage, die Zielgruppen, Nutzungsanlässe und Anforderungen an Funktionalitäten erhob. Statt der erhofften 500 TeilnehmerInnen füllten fast 2.500 Menschen den Fragebogen aus. 96,8% der Befragten stimmten dabei der Aussage zu: „Die Deutsche Digitale Bibliothek plant ein Zeitungsportal, in dem historische Zeitungen gemeinsam und einheitlich zugänglich gemacht werden sollen. – Dieses Portal würde für mich einen Mehrwert bieten."[8]

Auch in der Wissenschaftscommunity, v.a. den Geisteswissenschaften und den Digital Humanities, besteht weiterhin großes Interesse an einem nationalen Zeitungsportal, das geht aus den Reaktionen auf Projektvorstellungen bei Konferenzen wie der DHd2020[9] oder einschlägigen Workshops hervor.

Sowohl aus der Nutzerforschung als auch aus der Zusammenarbeit mit der wissenschaftlichen Begleitgruppe, die das Projekt seit Beginn der ersten Förderphase begleitet, ergaben sich – neben dem generellen Bedarf für ein nationales Zeitungsportal – konkrete Anforderungen, die maßgeblich in die Formulierung dieses zweiten Antrags eingeflossen sind (z.B. Korpusbildung, Nachnutzungsmöglichkeiten, Integration der Artikelebene etc., siehe Abschnitte „Ziele" und „Arbeitsprogramm"). Erhoben wurden die Anforderungen einerseits bei der erwähnten Umfrage und Usability-Tests mit ausgesuchten ProbandInnen,[10] andererseits bei einem zweitägigen Workshop mit der wissenschaftlichen Begleitgruppe, der im Juni 2019 in Berlin stattfand.[11] Diese enge Rückkopplung mit den (zukünftigen) NutzerInnen des Zeitungsportals war für die Projektarbeit und die weitere Planung sehr hilfreich und soll in der zweiten Projektphase wiederholt bzw. weitergeführt werden. Durch die konsequente Ausrichtung der Entwicklungsarbeit an den Wünschen und Anforderungen der zukünftigen NutzerInnen ist eine hohe Akzeptanz des Zeitungsportals zu erwarten.

1.1. Projektbezogene Publikationen

1.1.1. Veröffentlichte Arbeiten aus Publikationsorganen mit wissenschaftlicher Qualitätssicherung, Buchveröffentlichungen sowie bereits zur Veröffentlichung angenommene, aber noch nicht veröffentlichte Arbeiten

Altenhöner, Reinhard: Auf dem Weg zu einem nationalen Zeitungsportal. Eine materialspezifische Kooperation als Treiber eines neuen Dienstes für Wissenschaft und Forschung. In: Kooperative Informationsinfrastrukturen als Chance und Herausforderung. Festschrift für Thomas Bürger zum 65. Geburtstag. Hg. von Achim Bonte/Juliane Rehnolt. Berlin, Boston 2018, S. 144–160. DOI: https://doi.org/10.1515/9783110587524-019.

Hubrich, Jessica: Visualisierung von Titelzusammenhängen. Titelhistorie und Netzwerkgraph im neuen ZDB-Katalog. In: Offen(siv)e Bibliotheken: Neue Zugänge, neue Strukturen, neue Chancen. 32. Österreichischer Bibliothekartag Wien, 15.–18. September 2015. Hg. von Bruno Bauer/Andreas Ferus/Josef Pauser. Graz-Feldkirch 2016, S. 269–282. URN: urn:nbn:de:0290-opus4-21125.

Hubrich, Jessica und Lieder, Hans-Jörg: Zeitungssuche interaktiv. Der neue ZDB-Webkatalog. Verschriftlichung des Bibliothekartagvortrags für den Kongressband. Erschienen im Dezember 2014 in o-bib 2014/1, S. 305–311. DOI: https://doi.org/10.5282/o-bib/2014H1S305-311.

Neudecker, Clemens: Who cares about yesterday's news? Use cases and requirements for newspaper digitization. Proceedings of the IFLA 2016 News Media Conference, 20–22 April 2016, Hamburg, Germany. Online: https://blogs.sub.uni-hamburg.de/ifla-newsmedia/wp-content/uploads/2016/04/Neudecker-Who-cares-about-yesterday%e2%80%99s-news-Use-cases-and-requirements-for-newspaper-digitization.pdf [05.07.2020].

1.1.2. Andere Veröffentlichungen

Dinger, Patrick und Landes, Lisa: Geschichte aus erster Hand. Der Aufbau eines nationalen Zeitungsportals unter Berücksichtigung der Bedürfnisse verschiedener Nutzergruppen. In: Christof Schöch (2020): DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts, S. 141–144. DOI: https://doi.org/10.5281/zenodo.3666690.

[8] Siehe Ergebnisse der Nutzerumfrage in Anlage 2, S. 115.
[9] Zum Programm der DHd 2020 in Paderborn siehe https://dhd2020.de/ [08.05.2020].
[10] Für die Ergebnisse der Umfrage und der Usability-Tests siehe Anlage 2 und 3.
[11] Für das Protokoll des Workshops siehe Anlage 4.

Hubrich, Jessica und Lieder, Hans-Jörg: Die Zeitschriftendatenbank und die Digitalisierung historischer Zeitungen in Deutschland. In: Dialog mit Bibliotheken 28 (2016) 2, S. 22–28.

Lieder, Hans-Jörg: Coordinating Newspaper Digitisation: Some Facts and Figures. Paper for the IFLA International News Media Conference, Hamburg 2016. Online: https://blogs.sub.uni-hamburg.de/ifla-newsmedia/wp-content/uploads/2016/04/Lieder-Coordinating-Newspaper-Digitisation-%E2%80%93-Some-Facts-and-Figures.pdf [08.05.2020].

1.1.3. Patente

Entfällt.

2. Ziele und Arbeitsprogramm

2.1. Voraussichtliche Gesamtdauer des Projekts

Für das Gesamtprojekt ist eine Laufzeit von 48 Monaten vorgesehen. Die erste Projektphase (24 Monate) läuft seit Januar 2019 und ist voraussichtlich Ende 2020 abgeschlossen. Gewünschter Beginn der Förderung für die zweite Projektphase (weitere 24 Monate) ist der 1. Januar 2021.

2.2. Ziele

Das Hauptziel des Projekts „DDB-Zeitungsportal V. 2.0“ besteht darin, die bisherigen Funktionen des Zeitungsportals zu optimieren und neue Funktionen für das Portal zu entwickeln, um Akzeptanz und Nutzung des Zeitungsportals zu steigern. Durch Funktionen wie die individuelle Korpusbildung, die Darstellung der Artikelebene sowie die Erweiterung um zusätzliche Formate und Bestände sollen die Möglichkeiten der (Nach-)Nutzung erweitert und das Zeitungsportal als *der* zentrale und nutzerfreundliche Zugang zu den digitalisierten historischen Zeitungen aus Deutschland etabliert werden. Die Weiterentwicklung des DDB-Zeitungsportals soll folgende Bereiche umfassen:

1. Optimierung und Erweiterung bestehender Funktionen

Folgende, in der ersten Projektphase entwickelte Funktionen werden optimiert und erweitert:

- Die Verzahnung mit der Zeitschriftendatenbank (ZDB) wird ausgebaut. Während in der ersten Projektphase die Integration der ZDB-Daten in das Zeitungsportal im Mittelpunkt stand, wird jetzt der Rückfluss der Daten, also die Einbindung der Zeitungsportal-Daten in die ZDB, umgesetzt (s. AP 3).
- Die übergreifende Suche in den Volltextbeständen wird optimiert (s. AP 4).
- Die Anzahl der (Volltext)-Bestände wird erhöht (durch Akquise neuer Datenpartner und die Erweiterung der akzeptierten Eingangsformate, s. AP 4 und AP 5).

2. Entwicklung neuer Funktionen

Folgende Funktionen kommen neu dazu:

- Eine Funktion zur Bildung, Beschreibung und zum Export individueller Zeitungskorpora wird eingeführt (s. AP 7).
- Um die Weitergabe und damit auch die Nachnutzbarkeit der Daten im Zeitungsportal zu verbessern, werden die bestehenden Zugriffsmöglichkeiten um neue Schnittstellen für Suche und Harvesting erweitert (s. AP 8).
- Das Zeitungsportal wird um die Hierarchieebene der Artikel erweitert. Zukünftig wird es möglich sein, Bestände, die auf Artikelebene erschlossenen sind, im Viewer entsprechend anzuzeigen und die zugehörigen Artikel herunterzuladen und persistent zu referenzieren (s. AP 6).

3. Nutzererwartungen und Anschlussfähigkeit

Das Zeitungsportal muss sich an den Nutzererwartungen orientieren und für aktuelle und zukünftige Entwicklungen offen und anschlussfähig sein. Wie in der ersten soll darum auch in der zweiten

Projektphase die Projektarbeit durch eine enge Koppelung mit der wissenschaftlichen Begleitgruppe und der Nutzerforschung geprägt sein (s. AP 2).

Die Anschlussfähigkeit des Zeitungsportals soll einerseits über die Integration neuer Konzepte und Technologien ermöglicht werden, andererseits über die Intensivierung bestehender Kooperation wie jene mit der Zeitschriftendatenbank (ZDB) und Europeana, insbesondere der Sammlung „Europeana Newspapers". Über die bestehenden Kooperationen hinaus soll der Kontakt zu anderen Infrastrukturprojekten der Digital Humanities bzw. den geisteswissenschaftlichen NFDI-Konsortien intensiviert werden, um Synergieeffekte zu erzeugen (s. AP 3).

Stärker in die Zukunft gerichtet sind auch die Arbeiten in AP 9 („Konzeption von Datenanreicherungs-Diensten für OCR & NER"). Hier werden verfügbare Technologien, die dem aktuellen Stand der Technik entsprechen, für das Zeitungsportal angepasst und erprobt. Ziele sind die Bewertung entsprechender Datendienste vor allem für kleinere Einrichtungen und von technischen Optionen zur Integration der Dienste in die DDB.

2.3. Arbeitsprogramm und Umsetzung

2.3.1. AP 1 – Projektmanagement und Öffentlichkeitsarbeit

Die Gesamtleitung des Projekts liegt bei der DNB. Diese wird, wie in der ersten Projektphase, die Steuerung des Ressourceneinsatzes sowie die Erstellung und Überwachung der Terminplanung für die beteiligten Projektpartner übernehmen. Zu ihren Aufgaben zählen das Risikomanagement, mit dem auf unvorhergesehene Entwicklungen reagiert wird, sowie das Konfliktmanagement, mit dem etwaige Meinungsverschiedenheiten zwischen den Partnern angegangen und gelöst werden.

Außerdem sichert die Projektleitung in ihren Rollen als technische Koordinatorin der DDB und Betreiberin der DDB-Servicestelle gemeinsam mit FIZ Karlsruhe, dem technischen Betreiber der DDB und verantwortlichen Softwareentwickler, die Koordination mit den sonstigen Entwicklungen und operativen Belangen der DDB zu.

Zur Koordination des Projektteams und der unterschiedlichen Arbeitspakete wird am vierzehntägigen Turnus der Telefonkonferenzen, den halbjährlichen Projekttreffen sowie an den etablierten Methoden der agilen Steuerung festgehalten. Konflikte, die auf Arbeitsebene nicht gelöst werden können, werden durch ein bei Bedarf zusammentretendes Steuerungsgremium entschieden, in dem alle Projektpartner auf Leitungsebene vertreten sind.

Zu den in der ersten Projektphase etablierten Aufgaben und Methoden der Koordinierung kommt in der zweiten Projektphase die Öffentlichkeitsarbeit als ständige Aufgabe der Projektleitung hinzu. Es handelt sich dabei v.a. um Maßnahmen, die das Zeitungsportal in der Wissenschaftscommunity noch bekannter machen sollen (Vorträge auf einschlägigen Konferenzen, Blog- und Zeitschriftenartikel etc.). Andererseits richten sie sich auch an die interessierte Öffentlichkeit. Hierbei werden neben den Kommunikationskanälen der DDB auch die Kanäle der vier Projektpartner einbezogen.

Arbeitsschritte und Aktivitäten[12]

- Gesamtplanung und regelmäßiger Review (M1–M24)
- Projektleitung und Projektsteuerung inkl. Risiko- und Konfliktmanagement (M1–M24)
- Koordination mit den anderen DDB-Entwicklungen (M1–M24)
- Evaluation des Projektfortschritts, Berichtswesen (M1–M24)
- Interne Kommunikation: Wiki, JIRA, Mailing-Listen (M1–M24)
- (Fach)-Öffentlichkeitsarbeit (M1–M24)

AP-Beteiligte und Aufwände

	DNB (AP-Leitung)	FIZ	SLUB	SBB
beantragte Personalaufwände	6 PM	2 PM	2 PM	2 PM
Eigenmittel	4 PM	1 PM	1 PM	1 PM

[12] Alle Zeitangaben beziehen sich auf Projektmonate (M1 = 1. Projektmonat, M24 = 24. (letzter) Projektmonat).

2.3.2. AP 2 – Wissenschaftliche Begleitgruppe und Nutzerforschung

Aufbauend auf der engen Zusammenarbeit mit der wissenschaftlichen Begleitgruppe, die das Projekt im Hinblick auf wissenschaftliche Anforderungen und Nutzungsszenarien sowie mit dem Review der Konzeptpapiere und des Prototyps konstruktiv beraten hat, wird die wissenschaftliche Begleitung des Projekts in der zweiten Förderphase fortgesetzt. Hierbei werden insbesondere die neu hinzukommenden Themen Korpusbildung, Datenanreicherung und fortgeschrittene Suchmöglichkeiten mit der wissenschaftlichen Begleitgruppe im Rahmen eines gemeinsamen Workshops adressiert.

Neben der wissenschaftlichen Nutzung wird in Zusammenarbeit mit einer externen Agentur das Nutzungsinteresse und -verhalten der interessierten Öffentlichkeit weiter erhoben. So wird im zweiten Jahr der zweiten Projektphase der dann erreichte Stand des Zeitungsportals ausgewählten NutzerInnen vorgelegt und evaluiert. Daraus werden Rückschlüsse für nötige oder mögliche Weiterentwicklungen des Zeitungsportals auch nach Ende der Projektlaufzeit gewonnen. Die Kosten für die externe Beratung und Nutzerforschung werden aus DDB-Eigenmitteln bestritten.

Arbeitsschritte und Aktivitäten

- Workshop mit Projektgruppe und wiss. Begleitgruppe zum Review der funktionalen Erweiterungen des Zeitungsportals (Vorbereitung und Durchführung) (M12)
- Erstellen eines Aufgabenkatalogs und einer Ausschreibung für die Nutzerforschung (M12–M14)
- Auswahl einer Agentur aufgrund der eingegangenen Angebote (M15)
- Durchführung und Auswertung der Nutzerforschung (M16–M18)

AP-Beteiligte und Aufwände

	DNB	FIZ	SLUB	SBB (AP-Leitung)
beantragte Personalaufwände	2 PM	0 PM	0 PM	6 PM
Eigenmittel	2 PM	1 PM	1 PM	2 PM

2.3.3. AP 3: Verzahnung mit Schwestervorhaben

Das DDB-Zeitungsportal steht in Kontakt mit einer Vielzahl von Akteuren im Bereich Zeitungsdigitalisierung. In diesem Arbeitspaket sind Kooperations- und Austausch-Arbeiten gebündelt.

Eine besonders enge Kooperation besteht mit der Zeitschriftendatenbank (ZDB). In der ersten Projektphase wurde die Nachnutzung von ZDB-Daten im Zeitungsportal umgesetzt. So spielt die Einbindung von ZDB-Titeldaten und eindeutigen ZDB-Identifiern bereits jetzt eine tragende Rolle in der Datenarchitektur des Zeitungsportals. In der zweiten Projektphase wird der hier zugrundeliegende Datenfluss von der ZDB zur DDB weiter optimiert. Es wird geprüft, inwiefern es sinnvoll ist, statt eines Dumps eine Schnittstelle zur Übermittlung der ZDB-Daten zu nutzen. Sollte es bei der Dump-Nutzung bleiben, soll der Bereitstellungsrhythmus für den Dump erhöht werden. Ggf. werden für Anzeige und Suche im DDB-Zeitungsportal auch zusätzliche Informationen aus dem ZDB-Datendump ausgewertet. Außerdem werden die Anzeige von ZDB-Daten im Zeitungsportal und der Verlinkungen aus dem Zeitungsportal zum ZDB-Katalog an das Nutzerfeedback zum Prototypen angepasst.

Ein weiterer Arbeitsschwerpunkt im Rahmen der Kooperation mit der ZDB sind Konkretisierung und Ausgestaltung des Datenflusses in die andere Richtung, also vom Zeitungsportal in die ZDB. Durch den Rückfluss Zeitungsportal-spezifischer Daten in die ZDB wird gewährleistet, dass titelspezifische Verlinkungen vom ZDB-Katalog zum Zeitungsportal angeboten werden können, und dass die Informationen über die im Zeitungsportal aggregierten Titel über die gängigen ZDB-Schnittstellen zur Verfügung stehen und damit vielfältige Nachnutzungsszenarien ermöglicht werden.

Von der wissenschaftlichen Begleitgruppe wurde in ersten Projektphase zudem der Wunsch nach kontextuellen Zusatzinformationen zu allen im DDB-Zeitungsportal nachgewiesenen Titeln geäußert, und zwar in Form kurzer Texte, die strukturierte Informationen anbieten, die über die bislang in der ZDB verfügbaren Formalerschließungsangaben hinausweisen (z.B. politische Ausrichtung,

konfessionelle Bindung, Auflagenstärke, Lesepublikum von Zeitungen). Dieser Wunsch wird in der zweiten Projektphase aufgegriffen. Es wird zunächst geprüft, welche formatspezifischen Möglichkeiten die ZDB bietet. Ist ein Nachweis derartiger Informationen in der ZDB auf sinnvolle Weise möglich, werden exemplarisch entsprechende Informationen in der ZDB hinterlegt und der Datenfluss zwischen den Systemen entsprechend angepasst, damit diese im Zeitungsportal angezeigt werden können. Die Möglichkeit zur Erfassung derartiger Informationen in der ZDB wird bei digitalisierenden Einrichtungen beworben.

Neben der ZDB bestehen enge Beziehungen zur Europeana und deren thematischer Sammlung „Europeana Newspapers“. Aktuell gibt es bei Europeana noch keinen Ingestworkflow für digitalisierte Zeitungen mit Volltexten. Dennoch ist die perspektivische Abstimmung von Standards notwendig, da damit zu rechnen ist, dass sich in Zukunft eine Möglichkeit ergibt, Zeitungen mit Volltexten an Europeana zu liefern. Auch der Austausch über die Nutzergruppen der jeweiligen Zeitungsportale wird in der zweiten Projektphase weitergeführt.

Darüber hinaus sind weitere wichtige Stakeholder in der deutschen und europäischen Wissenschaftslandschaft zu berücksichtigen. Nicht zuletzt um das Zeitungsportal möglichst erfolgreich in der Landschaft der digitalisierten historischen Zeitungen zu positionieren, wird die Projektgruppe auch in der zweiten Projektphase mit den relevanten Akteuren (anderen Zeitungsportalen, einschlägigen Gremien und Arbeitsgruppen sowie relevanten Fachinformationsdiensten (FIDs)) im Gespräch bleiben. Hierbei besonders hervorzuheben sind die in Entstehung befindlichen geisteswissenschaftlichen NFDI-Konsortien, mit denen Synergien und Möglichkeiten der Zusammenarbeit ausgelotet werden.[13] Ziel solcher Gespräche ist u.a. die Sicherstellung formatspezifischer und technischer Kompatibilität der beteiligten Systeme. Die gewonnenen Erkenntnisse über Anforderungen an digitalisierte Zeitungen bzw. Zeitungsportale und von anderen Projekten eingesetzte Techniken und Funktionen fließen auch über die Projektphase hinaus in den Weiterentwicklungsprozess des DDB-Zeitungsportals ein.

Arbeitsschritte und Aktivitäten

- Prüfung der Verankerung zusätzlicher titelbezogener Informationen in der ZDB (M1–M3)
- Optimierung von Anzeige und ZDB-Verlinkungen im DDB-Zeitungsportal aufgrund von Rückmeldungen zum Prototyp (M1–M6)
- Erstellung und Umsetzung des Konzepts zur Lieferung von Daten aus dem DDB-Zeitungsportal an die ZDB (M6–M12)
- Integration von Links zum DDB-Zeitungsportal im ZDB-Katalog (M6–M12)
- Optimierung von Lieferung und Workflows von ZDB an das DDB-Zeitungsportal unter Berücksichtigung von Updateroutinen (M13–M16)
- Austausch mit Europeana bzgl. Datenmodellen, Lieferworkflows und Schnittstellentechnologien sowie den Ergebnissen der Nutzerforschung (M1–M24)
- Kontaktpflege und Austausch mit Stakeholdern (anderen Zeitungsportalen, einschlägigen Gremien und Arbeitsgruppen, relevanten FIDs und NFDI-Konsortien) (M1–M24)

AP-Beteiligte und Aufwände

	DNB	FIZ	SLUB	SBB (AP-Leitung)
beantragte Personalaufwände	1 PM	5 PM	0 PM	5 PM
Eigenmittel	2 PM	1 PM	1 PM	2 PM

2.3.4. AP 4: Evaluierung, Erweiterung und Optimierung der Verarbeitung von Volltexten

Die Volltextsuche ist der zentrale Einstieg ins DDB-Zeitungsportal. Die Möglichkeit, nicht nur die Metadaten, sondern auch die Inhalte der Zeitungsartikel durchsuchen zu können, stellt eine enorme Erweiterung des Suchraums dar. Ziel dieses Arbeitspakets ist es daher einerseits, die Zahl der Volltextbestände weiter auszubauen: In Eigenleistung wird Akquise und Ingest neuer Datenpartner, die bereits Volltexte im Standardformat (ALTO) vorliegen haben, vorangetrieben.

[13] Zum Zeitpunkt der Antragstellung ist das Ergebnis der Begutachtung der NFDI-Anträge noch nicht bekannt. Falls die Förderung bewilligt wird, sind Austausch und Vernetzung angestrebt.

In der ersten Projektphase wurden aber auch relevante Bestände identifiziert, deren Volltexte nicht im ALTO-Format vorliegen. Um auch solche Bestände in das Zeitungsportal einspielen zu können, werden in der zweiten Projektphase von der Fachstelle Bibliothek insbesondere für die Formate PAGE-XML[14] und TEI[15] Transformationsszenarien nach ALTO entworfen und umgesetzt. Dafür werden bestehende Vorarbeiten des DFG-Projekts OCR-D[16] aufgegriffen und speziell für Zeitungen weiterentwickelt. Teilweise handelt es sich dabei um Bestände, bei denen bereits händisch oder automatisiert Entitäten wie Personen, Körperschaften oder Orte ausgezeichnet wurden. Der Viewer wird darum dahingehend angepasst, dass in den Volltexten ausgezeichnete Entitäten erkannt und visualisiert werden können. Zudem soll er um eine Unterstützung für in TEI vorliegende Volltexte erweitert werden, so dass diese auch ohne vorherige Transformation angezeigt werden können.

Der zweite Teil des Arbeitspakets widmet sich der Verbesserung und dem Ausbau der Funktionalität der Volltextsuche. Insbesondere wird die Suche im Hinblick auf historische Schreibweisen, die mögliche Einbindung weiterer Lexika, syntaktische Strukturen und Lemmata im laufenden Betrieb analysiert und weiter verbessert. Parallel hierzu werden die Such- und Ranking-Parameter auf Basis realer Nutzerdaten kontinuierlich getestet und optimiert. Die vorgesehenen Maßnahmen basieren auf den Erfahrungen der ersten Projektphase, den Rückmeldungen aus der wissenschaftlichen Begleitgruppe sowie der Nutzerforschung.

Arbeitsschritte und Aktivitäten

- Anpassung der Workflows für die Datenakquise, Transformation und die Validierung weiterer Volltextformate (M1–M12)
- Viewer-Unterstützung für TEI-Volltexte (M7–M12)
- Viewer-Unterstützung in der Anzeige von Entitäten und weiterführenden Verlinkungen (M13–M18)
- Fortlaufende Verbesserung des Information Retrievals der Volltextsuche (M1–M24)
- Ingest neuer Datenpartner und neuer Bestände mit Volltexten (M1–M24)

AP-Beteiligte und Aufwände

	DNB (AP-Leitung)	FIZ	SLUB	SBB
beantragte Personalaufwände	4 PM	4 PM	4 PM	1 PM
Eigenmittel	6 PM	1 PM	0 PM	0 PM

2.3.5. AP 5: Integration weiterer Zeitungsbestände

Digitalisierte historische Zeitungen sind in der deutschen Wissenschaftslandschaft in vielfältigen Kulturerbe-Einrichtungen zu finden. Während in der ersten Projektphase das Hauptaugenmerk auf bibliothekarischen Beständen lag, bei deren Erschließung die DFG-Praxisregeln und insbesondere das METS/MODS-Anwendungsprofil für Zeitungen berücksichtigt wurden, werden in der zweiten Projektphase die technischen Infrastrukturen weiterentwickelt, um auch anders erschlossene Bestände in das Zeitungsportal einspielen zu können. Dabei geht es hauptsächlich um Bestände und Formate, die im Rahmen des vorbereitenden Arbeitspaketes der ersten Phase (AP 10 „Vorbereitung der Integration weiterer Bestände“) identifiziert wurden.

Der Schwerpunkt des Arbeitspakets liegt auf der Implementierung des IIIF-Standards, der eine große Interoperabilität ermöglicht und in den letzten Jahren verstärkt Verbreitung im Kulturerbe-Sektor erfährt. Für Bestände wie z.B. die Zeitungen von digiPress[17] (BSB), die in Form von IIIF exponiert werden, wird, aufbauend auf den Untersuchungen aus der ersten Projektphase, die Verarbeitung der DDB-Standard-Lieferformate für Zeitungen dahingehend erweitert, dass darin enthaltene IIIF-Links (Links auf Bild-Ressourcen und auf IIIF-Manifeste) verarbeitet und angezeigt werden können. Zu diesem Zweck werden der Validierungs- und Transformations-Workflow bei der Fachstelle Bibliothek

[14] Siehe https://github.com/PRImA-Research-Lab/PAGE-XML [07.05.2020].
[15] Mit TEI werden insbesondere Volltexte im Bereich der Editorik ausgezeichnet. Zunehmend kann TEI als ein Ausgabeformat der OCR-Texterkennung auch im Bereich der Zeitungsdigitalisierung Anwendung finden. Siehe dazu Resch/Kampkaspar: DIGITARIUM 2019.
[16] Siehe https://github.com/OCR-D/ocr-fileformat [07.05.2020].
[17] Siehe https://digipress.digitale-sammlungen.de/ [08.05.2020].

für das Einspielen der entsprechenden Daten in das Zeitungsportal aufgebaut und die Bestände eingespielt. Darüber hinaus werden Best-Practice-Beispiele zur Umsetzung von IIIF-Manifesten veröffentlicht.

Neben der Vorverarbeitung der IIIF-Referenzen in den Lieferformaten sind auch Anpassungen am integrierten DDB-Viewer notwendig, um den IIIF-Standard vollumfänglich zu unterstützen. Zwar implementiert die dem Viewer zugrundeliegende Technologie Kitodo.Presentation bereits die IIIF Presentation API sowie die IIIF Image API, allerdings werden aktuell beispielsweise noch keine Collection-Strukturen unterstützt, wie sie üblicherweise zur Kodierung von Zeitungen zum Einsatz kommen. Durch eine Aktualisierung der verwendeten OpenLayers-Komponente von Version 3 auf Version 6 wird die derzeit Kitodo-spezifische Implementierung der IIIF Image API durch eine generische ersetzt, die in einer sehr viel größeren Community etabliert ist.

Nicht zuletzt ist auch die Verarbeitung von Annotationen und alternativen Image-Derivaten, also Bildern mit anderen Aufnahmetechniken oder Auflösungen, noch verbesserungswürdig, da aktuell nur eine Annotationsebene bzw. ein Image pro Seite angezeigt werden kann – in der Regel der OCR-Volltext oder eine Transkription sowie ein normaler Scan. Es wird eine Funktion implementiert, über die verschiedene Annotationsquellen und Image-Derivate ausgewählt und visualisiert werden können, sofern diese im IIIF-Manifest oder der METS-Datei referenziert wurden.

In der zweiten Projektphase werden außerdem explizit Archive, die in der Regel ihre Metadaten nicht im METS/MODS-Anwendungsprofil vorliegen haben, als bestandshaltende Einrichtungen v.a. regionaler Zeitungssammlungen in den Blick genommen. Hier wird von der DNB untersucht, welche Metadatenformate zur Beschreibung von Zeitungen in Archiven gängig sind und ein Konzept für einen Standardworkflow zur Integration von Archiv-Sammlungen in das Zeitungsportal entwickelt und umgesetzt. Anschließend werden Zeitungsbestände aus Archiven eingespielt.

Arbeitsschritte und Aktivitäten

IIIF

- Erweiterung der DDB-Standard-Lieferformate für Zeitungen für die Integration von Beständen, die IIIF-Links enthalten (M1–M6)
- Formulierung und Bereitstellung relevanter Informationen, inkl. Best-Practice-Beispielen, zur Lieferung von Beständen, die IIIF-Links enthalten, für die Datenpartner (M1–M8)
- Erweiterung der IIIF-Unterstützung im Viewer um „Collections“ und Aktualisierung der OpenLayers-Komponente (M1–M8)
- Aufbau der Prozessierungs- und Anreicherungsworkflows für Bestände, die IIIF-Links enthalten (M7–M12)
- Erweiterung des Mappings um mindestens ein Metadatenfeld für die Anzeige von IIIF-Manifesten (M12–M13)
- Einspielen von Beständen, die IIIF-Links enthalten (M14–M24)

Weitere Datenformate

- Recherche zu gängigen Metadatenformaten für Zeitungsbestände in Archiven (M1–M3)
- Anpassung der Lieferprozesse zur Selektion von Zeitungsdaten (M4–M5)
- Mapping-Anpassungen für das Archiv-Anwendungsprofil Zeitungen inklusive Anpassungen am Data Preparation Tool (M6–M8)
- Prozesserweiterung zur Integration von ZDB-Daten (M7–M8)
- Prozesserweiterung für Download und Indexierung von Volltexten (M7–M9)
- Auswahl von Bildderivaten und Annotationsquellen im Viewer (M9–M12)
- Evaluierung der Umsetzungsergebnisse und Qualitätssicherung (M10–M12)
- Einspielen von Archiv-Beständen (M14–M24)

AP-Beteiligte und Aufwände

	DNB (AP-Leitung)	FIZ	SLUB	SBB
beantragte Personalaufwände	5 PM	0 PM	8 PM	1 PM
Eigenmittel	4 PM	2 PM	2 PM	0 PM

2.3.6. AP 6: Erweiterung des Zeitungsportals für die Darstellung von Zeitungsartikeln

Im Rahmen des DFG-Pilotprojekts „Digitalisierung historischer Zeitungen“[18] wurde die Artikelseparierung erprobt und eine Reihe von Zeitungen auf Artikelebene erschlossen (im Sinne des Masterplans Zeitungsdigitalisierung handelt es sich bei der Erschließung mit Artikelseparierung um den erweiterten Digitalisierungsstandard 1).[19] Um diese und andere artikelseparierte Bestände adäquat anzeigen zu können, wird in der zweiten Projektphase das Zeitungsportal um die Darstellung der Artikelebene erweitert.

Dazu wird das Datenmodell um die Hierarchieebene der Artikel erweitert und ein entsprechendes Mapping definiert. Die daraus resultierenden Erweiterungen des Backends (Index-Struktur) werden zusammen mit den nötigen Anpassungen im Design (Frontend) umgesetzt. Ferner wird die Darstellung der Zeitungen über den DDB-Viewer angepasst: Innerhalb einer Ausgabe können die darin enthaltenen Artikel über ein „Inhaltsverzeichnis“ direkt angesteuert werden. Die Druckfunktion wird um die Artikelebene ergänzt: Neben Ausgaben können auch einzelne Artikel (als Bild und als Volltext) heruntergeladen, gespeichert und gedruckt werden. Die dazu notwendige Generierung der Artikel-Downloads erfolgt dynamisch im Viewer.

Für die einzelnen Zeitungsartikel werden beim Ingest außerdem individuelle IDs vergeben, sodass persistente und zitierbare Links entstehen. Das in der ersten Projektphase für die Ausgaben-Ebene entwickelte Konzept der Zitierfähigkeit wird um die Artikelebene erweitert.

Voraussetzung für alle beschriebenen Funktionen sind artikelseparierte Lieferdaten.

Arbeitsschritte und Aktivitäten

- Mapping-Anpassungen am METS/MODS-Anwendungsprofil Zeitungen für die Artikelebene (M1–M4)
- Erweiterung der Index-Struktur und Befüllung des Index (M1–M4)
- Anpassung des Portaldesigns zur Darstellung der Artikelebene und Implementierung („Inhaltsverzeichnis“) (M5)
- Erweiterung der Viewerfunktionen um dynamische PDF-Erstellung für Download- und Druckfunktion (Bild und Volltext) (M6–M7)
- Erweiterung der zitierfähigen Referenzierung um Artikel (M8–M10)

AP-Beteiligte und Aufwände

	DNB (AP-Leitung)	FIZ	SLUB	SBB
beantragte Personalaufwände	3 PM	4 PM	4 PM	1 PM
Eigenmittel	2 PM	1 PM	1 PM	1 PM

2.3.7. AP 7: Korpusbildung

Grundlage der meisten geisteswissenschaftlichen Arbeiten ist die Korpusbildung. Erst wenn eine Auswahl der relevanten Objekte getroffen wurde, kann mit der Bearbeitung der Forschungsfrage begonnen werden. Dies setzt sich auch für Arbeiten in den digitalen Geisteswissenschaften fort, für die oft ein digitales Korpus, also eine maschinenlesbare Sammlung von Texten und Bildern, gebildet werden muss.

Durch die zunehmend verfügbare Volltexterschließung historischer Zeitungen können verstärkt quantitative Methoden für die Erforschung der textuellen Inhalte angewandt werden. Aber auch die Digitalisate und Metadaten können für die Forschung relevante Informationen enthalten und zu neuen transdisziplinären Erkenntnissen führen. Zukünftig soll es darum möglich sein, im DDB-Zeitungsportal Korpora, die sowohl Volltexte als auch Digitalisate und Metadaten enthalten, über den gesamten im Portal verfügbaren Bestand zu erstellen.

[18] Siehe https://gepris.dfg.de/gepris/projekt/225572475 [07.05.2020].
[19] Siehe Masterplan Zeitungsdigitalisierung, S. 22, 30, 62.

In der DDB können NutzerInnen bereits über den persönlichen Nutzerbereich „Meine DDB“ Suchen- und Favoritenlisten speichern. Dieses Angebot wird in der zweiten Projektphase an die Anforderungen der Korpusbildung im Zeitungsportal angepasst und funktional ausgebaut. Konkret bedeutet dies, dass im jeweiligen persönlichen Nutzerbereich auf Objektebene eine Auswahl für das Korpus getroffen werden kann. Für jedes Korpus wird dabei eine persistente ID vergeben, die ein Datenset definiert und eindeutig referenzierbar macht. Über ein Textfeld wird die intellektuelle Beschreibung des Korpus ermöglicht. Anschließend können die für die Analyse des Korpus benötigten Informationen (Volltexte, Digitalisate, Metadaten) selektiert und heruntergeladen werden.[20]

Arbeitsschritte und Aktivitäten

- Sammlung der Anforderungen unter Einbeziehung der wissenschaftlichen Begleitgruppe (M10–M14)
- Evaluierung der etablierten Metadatenformate für Korpusbeschreibungen (z.B. DCAT)[21] (M15–M16)
- Konzepterstellung für eine Erweiterung der „Meine DDB“-Funktion inkl. Download-Funktion für Korpora (M16–M18)
- Entwicklung und Umsetzung des technischen Konzepts zur Korpusbildung (M19–M24)

AP-Beteiligte und Aufwände

	DNB (AP-Leitung)	FIZ	SLUB	SBB
beantragte Personalaufwände	6 PM	12 PM	0 PM	1 PM
Eigenmittel	2 PM	2 PM	1 PM	0 PM

2.3.8. AP 8: Datenweitergabe

Dieses Arbeitspaket widmet sich der Verbesserung der Datenweitergabe. In der aktuellen Ausbaustufe können sich NutzerInnen einzelne Ausgaben oder einzelne Seiten als JPG- oder PDF-Datei herunterladen sowie über die DDB-API[22] auf den Zeitungsportal-Bestand zugreifen. In der zweiten Projektphase werden diese Möglichkeiten um den Zugriff über standardisierte Schnittstellen für die dezentrale Suche sowie Integration in verteilte Präsentationskontexte erweitert. Dafür werden sowohl die SRU/CQL-Schnittstelle[23] als auch die IIIF Presentation API, die im Umfeld der Kulturinstitutionen etabliert sind, unter dem Gesichtspunkt einer Implementierung in der DDB-Architektur evaluiert. Auf Basis der Evaluation werden die Schnittstellen schließlich technisch konzipiert und dokumentiert und dann von FIZ Karlsruhe entwickelt und integriert. Ziel ist es, die Inhalte des Zeitungsportals für gezielte Suchen und anschließende Übernahme in (virtuelle) Forschungsumgebung bereitzustellen, damit die Daten im Anschluss mit den eigenen Digital-Humanities-Tools analysiert werden können. Mit dieser Funktion wird auch eine Empfehlung der wissenschaftlichen Begleitgruppe umgesetzt, die sich einen besseren Zugang zu den Inhalten gewünscht hatte.

Im Kontext der Datenweitergabe muss auch die Frage der Datenformate behandelt werden. Derzeit akzeptiert die DDB eine Reihe von standardisierten Lieferformaten, die jedoch individuell je Datenpartner im Rahmen des Ingests ggf. aufbereitet und in DDB-interne Formate transformiert werden. Folglich könnten die Schnittstellen zur Datenweitergabe wahlweise die reicheren, aber heterogenen Lieferformate der Datenpartner bereitstellen oder das normalisierte, dafür jedoch reduzierte Internformat. Da im Rahmen der Anreicherungen in AP 9 ohnehin Änderungen an den Lieferdaten erfolgen müssen, können die dafür zu implementierenden Prozesse nachgenutzt werden, um diese Formate auch für die Datenweitergabe aufzubereiten. Die dazu notwendigen Prozessschritte

[20] Da das Zeitungsportal die dauerhafte Sicherung und Zugänglichmachung der Datensets nicht garantieren kann, weil die Verfügbarkeit der Daten immer auch in der Hand der Datenpartner liegt, bleibt es Aufgabe der Forschenden, die einmal heruntergeladenen Daten über geeignete Forschungsdatenrepositorien zu sichern.
[21] DCAT = Data Catalog Vocabulary, siehe https://www.w3.org/TR/vocab-dcat-2/ [07.05.2020].
[22] Siehe https://api.deutsche-digitale-bibliothek.de/doku/display/ADD/API+der+Deutschen+Digitalen+Bibliothek [07.05.2020].
[23] Siehe https://www.loc.gov/standards/sru/ [07.05.2020].

und Werkzeuge werden in AP 8 entwickelt. Dies umfasst mindestens die Transformation der Lieferformate in IIIF-Manifeste zur Bereitstellung über die IIIF Presentation API.

Arbeitsschritte und Aktivitäten

- Evaluation und technische Konzeption der Schnittstellen (SRU/CQL/IIIF Presentation API) (M1–M14)
- Bereitstellung einer Dokumentation und Anleitung zur Nutzung der Schnittstellen (M11–M12)
- Entwicklung von Transformationswerkzeugen und -verfahren zur Erzeugung von IIIF-Manifesten aus den Lieferformaten (M15–M20)
- Entwicklung und Integration der Schnittstellen (M15–M22)

AP-Beteiligte und Aufwände

	DNB	FIZ	SLUB (AP-Leitung)	SBB
beantragte Personalaufwände	4 PM	18 PM	6 PM	1 PM
Eigenmittel	2 PM	2 PM	0 PM	1 PM

2.3.9. AP 9: Konzeption von Datenanreicherungs-Diensten für OCR & NER

Die Kernfunktionalität des Zeitungsportals, die Volltextsuche über alle Bestände, ist abhängig von der Qualität der Texterkennung (OCR) bei den einzelnen Datenpartnern. Sind Volltexte von niedriger Qualität oder gar keine Volltexte vorhanden, wirkt sich dies in hohem Maße auf die Anzahl und Güte der Suchtreffer aus. Des Weiteren wird insbesondere in Zeitungen häufig nach Eigennamen gesucht, also bspw. Namen von Personen, Ortsbezeichnungen oder Körperschaften, sogenannten „Named Entities“. Basierend auf vorhandenen Volltexten können derartige Informationen automatisiert extrahiert (Named Entity Recognition, NER) und bspw. als zusätzliche Suchmöglichkeiten angeboten werden oder in ggf. anschließenden Arbeitsschritten mit den entsprechenden Einträgen in der Gemeinsamen Normdatei (GND) verknüpft werden.[24]

In diesem Arbeitspaket wird untersucht, ob es sinnvoll und möglich ist, seitens des Zeitungsportals den Bearbeitungen beim Datenpartner nachgelagerte Datenanreicherungs-Dienste aufzubauen, die die Möglichkeit bieten, Bestände, die ohne Volltexte geliefert werden, nachträglich mit OCR und ggf. NER anzureichern, um so fehlende Volltexte zu ergänzen und einheitliche Suchmöglichkeiten über sämtliche Zeitungsbestände anbieten zu können. Ein weiterer Anwendungsfall wäre eine Verbesserung bereits bestehender, aber qualitativ minderwertiger Volltexte.

Dazu wird zunächst von der SBB der Stand der Technik im Bereich OCR (insbesondere die Ergebnisse des OCR-D-Projekts[25]) und NER (bspw. die von der SBB im Rahmen des BMBF-Projekts Qurator[26] entwickelte NER für historische Daten[27]) auf der Grundlage eines repräsentativen Samples aus dem Bestand des Zeitungsportals evaluiert. Für die Qualitätsmessung sind dabei manuell transkribierte resp. annotierte und qualitätsgeprüfte Referenzdaten (sog. Ground Truth) zu erstellen.

Aufbauend auf den von der SBB vorgelegten Bericht wird von allen Projektbeteiligten ein Konzept entwickelt, wie Datenanreicherungs-Dienste für OCR/NER aufgebaut, betrieben und in den DDB-Routinebetrieb verankert werden können. Darin werden die aus einem dauerhaften Betrieb entstehenden Anforderungen an die DDB-Infrastruktur beschrieben und beurteilt (bspw. für die Implementierung, Pflege und Betrieb entsprechender Verfahren sowie für die Speicherung und Verfügbarmachung der Ergebnisse der Datenanreicherungen – sowohl innerhalb des Zeitungsportals, als auch für die Datenpartner).

[24] Im Sinne des Masterplans Zeitungsdigitalisierung handelt es sich bei der Erschließung mit Entitäten um den Erweiterten Digitalisierungsstandard 2, Stufe 4.
[25] Siehe https://ocr-d.de [07.05.2020].
[26] Siehe https://qurator.ai/projekt/ [07.05.2020].
[27] Siehe https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_4.pdf [07.05.2020].

Mit den Bemühungen um eine Erhöhung der Datenqualität will die DDB die wissenschaftlichen Auswertungsmöglichkeiten der Zeitungsdaten verbessern und damit auch im übergeordneten Rahmen der im Aufbau befindlichen Nationalen Forschungsdateninfrastruktur einen Beitrag leisten.[28]

Arbeitsschritte und Aktivitäten

- Erstellung eines repräsentativen Testsample aus den Beständen des Zeitungsportals (M1–M3)
- Erstellung eines Testkorpus (sog. Ground Truth) für die Evaluation von OCR und Segmentierung (M4–M6)
- Tests und Evaluation geeigneter OCR-Verfahren (unter besonderer Berücksichtigung der Ergebnisse des OCR-D-Projektes) (M7–M9)
- Erstellung eines Testkorpus für die Evaluation von NER-Verfahren (M10–M12)
- Tests und Evaluation geeigneter NER-Verfahren (M13–M15)
- Bericht zur Auswertung der OCR- & NER-Verfahren für Zeitungen inkl. Empfehlungen für deren Implementierung (M16–M18)
- Formulierung eines Konzepts zum Aufbau von Datenanreicherungs-Diensten inkl. deren Verankerung im DDB-Routinebetrieb (M19–M24)

AP-Beteiligte und Aufwände

	DNB	FIZ	SLUB	SBB (AP-Leitung)
beantragte Personalaufwände	1 PM	3 PM	0 PM	6 PM
Eigenmittel	4 PM	1 PM	1 PM	1 PM

Die folgende Übersicht zeigt den zeitlichen Verlauf der einzelnen Arbeitspakete im Gesamtprojekt.

		Projektmonate																						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Arbeitspakete	AP 1																								
	AP 2																								
	AP 3																								
	AP 4																								
	AP 5																								
	AP 6																								
	AP 7																								
	AP 8																								
	AP 9																								

2.4. Maßnahmen zur Erfüllung der Förderbedingungen und Umgang mit den Projektergebnissen

Der vorliegende Projektantrag beschreibt ein Vorhaben zur Konsolidierung und Optimierung von e-Research-Technologien im Sinne des DFG-Förderprogramms „e-Research-Technologien“. Dem Gesamtvorhaben ging eine Untersuchung im Rahmen des Pilotprojekts „Digitalisierung historischer Zeitungen“ voraus, aus der die Bereitstellung und Zugänglichmachung eines übergreifenden Zeitungsportals für die digitalisierten Zeitungsbestände in Deutschland als wesentliches Desiderat hervorgeht. Aufgrund dieses grundsätzlich formulierten Bedarfs, der im Rahmen der gutachterlichen Evaluierung des Masterplans zur Digitalisierung historischer Zeitungen vonseiten der DFG bestätigt wurde, wird das Zeitungsportal nun entwickelt.

Das Projekt zur Errichtung eines Zeitungsportals für Deutschland ist so angelegt, dass dafür die organisatorische und technische Infrastruktur, die die Deutsche Digitale Bibliothek aufgebaut hat und betreibt, soweit wie möglich nachgenutzt und integriert wird. Dadurch werden nicht nur während der Konzeptions- und Aufbauphase des Zeitungsportals erhebliche Synergien genutzt, mit denen der Entwicklungsaufwand deutlich gesenkt wird. Durch die starke Integration des Zeitungsportals mit den

[28] Mit diesem Arbeitspaket will das Zeitungsportal die Datenpartner jedoch nicht von ihrer Pflicht entbinden, qualitativ hochwertige Bilder, Metadaten und Volltexte zu liefern. Vielmehr soll die Datenqualität schon bestehender Bestände aus früheren Digitalisierungsaktivitäten an die modernen Anforderungen der Suche und Erschließung angepasst werden.

technischen Komponenten der DDB und den erforderlichen Prozessen und Organisationseinheiten wird auch der langfristige Betrieb des Zeitungsportals gesichert. Insbesondere die an der DNB ansässige DDB-Projektkoordination und der technische Betreiber der DDB (FIZ Karlsruhe) werden in der Konzeptions- und Entwicklungsphase des Zeitungsportals dafür Sorge tragen, dass der operative Betriebsaufwand des Zeitungsportals, der zusätzlich zum DDB-Betrieb entsteht, möglichst gering ist. Dies betrifft auch die Wartung und Pflege der Softwarekomponenten.

Die DDB verpflichtet sich, das entwickelte Zeitungsportal nach Fertigstellung langfristig zu betreiben und auch für dessen funktionale Erweiterung und Aktualisierung – dies ggf. mit zusätzlichen Fördermitteln – Sorge zu tragen. Bei einem ähnlichen Vorhaben, dem Archivportal-D, ist dies bereits erfolgreich umgesetzt worden: Nach Abschluss der durch die DFG geförderten Entwicklungsphase ist die Betriebsverantwortung einschließlich der Übernahme entsprechender Aufwände auf die DDB übergegangen. Durch die softwareseitige Integration des DDB-Zeitungsportals in das DDB-Portal werden die spezifischen Betriebsaufwände für das Zeitungsportal vermutlich sogar geringer sein als beim Archivportal-D, für das eine eigene Frontend-Software entwickelt wurde.

Sämtliche im Projekt entwickelte Software wird unter eine Open-Source-Lizenz gestellt und auf GitHub[29] veröffentlicht. Die im Projekt definierten Standards und Best-Practice-Beispiele bzgl. (Meta-)Datenformaten werden über ein offenes, bei FIZ Karlsruhe gehostetes Wiki archiviert und der Öffentlichkeit zugänglich gemacht.[30]

2.5. Erläuterungen zur inhaltlichen und finanziellen Projektbeteiligung von Kooperationspartnerinnen und Kooperationspartnern im Ausland

Entfällt.

3. Literaturverzeichnis

Empfehlungen zur Digitalisierung historischer Zeitungen in Deutschland (Masterplan Zeitungsdigitalisierung) – Ergebnisse des DFG-Projektes „Digitalisierung historischer Zeitungen“ Pilotphase 2013–2015, 29.01.2016 (Partner: SBB (Berlin), SuUB (Bremen), SLUB (Dresden), DNB (Frankfurt), ULB (Halle), BSB (München)), https://www.zeitschriftendatenbank.de/fileadmin/user_upload/ZDB/z/Masterplan.pdf [07.05.2020].

Resch, Claudia und Kampkaspar, Dario: DIGITARIUM: Unlocking the Treasure Trove of 18th-Century Newspapers for Digital Times. In: Digital Eighteenth Century: Central European Perspectives. Hg. von Thomas Wallnig/Marion Romberg/Joelle Weis. Wien, Köln, Weimar 2019, S. 49–64.

[29] Die DDB veröffentlicht alle Softwarekomponenten auf GitHub, siehe https://github.com/Deutsche-Digitale-Bibliothek [08.05.2020].
[30] Siehe https://wiki.deutsche-digitale-bibliothek.de/x/vQQiAQ [08.05.2020].