Zwischenbericht – Projektanträge im Bereich „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ (LIS)

LIS-Förderprogramm oder Ausschreibung: „e-Research-Technologien“

1 Allgemeine Angaben

1.1 Allgemeine Angaben

DFG-Geschäftszeichen:
KR 3444/14-1
HO 4756/4-1
NI 522/14-1
RA 1965/2-1

Antragsteller:
Prof. Dr. Robert Kretzschmar / Prof. Dr. Gerald Maier, Stuttgart
Dr. Michael Hollmann, Koblenz
Dr. Elisabeth Niggemann, Frankfurt am Main
Matthias Razum, Eggenstein-Leopoldshafen

Thema des Projektes:
Aufbau einer Infrastruktur zur Implementierung sachthematischer Zugänge im Archivportal-D am Beispiel des Themenkomplexes „Weimarer Republik”

Berichtszeitraum: 01.06.2018 bis 31.10.2019

Internetadresse des Vorhabens: www.archivportal-d.de

Mit dem Erstantrag kalkulierte Gesamtdauer des Vorhabens: 24 Monate
Projektbeginn: 01.06.2018
Voraussichtlicher Abschluss der Arbeiten: 31.05.2020
Erstbewilligung der DFG für dieses Projekt vom: 18.07.2012
Derzeit laufende DFG-Bewilligung vom: 15.03.2018

Aktueller Stand des Mittelabrufs aus der laufenden Bewilligung: 247.963,61 Euro

1.2 Bisher abgerufene Mittel

Landesarchiv Baden-Württemberg / Maier, Gerald

Postdoktorandin/Postdoktorand und Vergleichbare:	85.597,12 Euro
Reisen:	2.526,69 Euro
Mittelabruf von weiteren Bewilligungspositionen:	6.617,01 Euro
Noch verfügbare Mittel aus der laufenden Bewilligung:	57.815,18Euro

Bundesarchiv / Hollmann, Michael

Postdoktorandin/Postdoktorand und Vergleichbare:	0 Euro
	(erster Teil wird Anfang November 2019 abgerufen)
Reisen:	792,45 Euro
Mittelabruf von weiteren Bewilligungspositionen:	174,34 Euro
Noch verfügbare Mittel aus der laufenden Bewilligung:	85.933,21 Euro

Deutsche Nationalbibliothek / Niggemann, Elisabeth

Doktorandin/Doktorand und Vergleichbare:	0 Euro
Reisen:	0 Euro
Noch verfügbare Mittel aus der laufenden Bewilligung:	0 Euro

FIZ Karlsruhe / Razum, Matthias:

Doktorandin/Doktorand und Vergleichbare:	123.450 Euro
Reisen:	1.350 Euro
Mittelabruf von weiteren Bewilligungspositionen:	27.456 Euro
Noch verfügbare Mittel aus der laufenden Bewilligung:	90.294 Euro

2 Arbeits- und Ergebnisbericht

Ausgangslage und Zielsetzung des Projekts

Ausgangspunkt des Projekts war der Bedarf der historischen Forschung nach übergreifenden sachthematischen Zugangsmöglichkeiten zu Archivgut. Um diesem Bedarf zu begegnen, werden im Rahmen des vorliegenden DFG-Projekts sachthematische Zugänge als neue Funktionalität innerhalb des deutschlandweit übergreifenden Nachweisportals Archivportal-D entwickelt. Als erste beispielhafte Implementierung soll entsprechend dem aktuellen Forschungsinteresse ein Themenzugang “Weimarer Republik” geschaffen werden. Zwei Digitalisierungsprojekte des Landesarchivs Baden-Württemberg (LABW) und des Bundesarchivs (BARch) zu diesem Thema stellen dabei die ersten Inhalte zur Verfügung, die nach Abschluss der ersten Projektphase durch die Bestände weiterer Archiveinrichtungen ergänzt werden sollen.

Als neues Rechercheinstrument für die Nutzenden steht die Entwicklung einer Themensystematik im Zentrum, mit welcher die verschiedenen Archivobjekte verknüpft sind. Um die umfangreiche Bereitstellung von Archivgut über die Themenzugänge zu ermöglichen, werden verschiedene Strategien verfolgt: Zum einen wird ein Software-Tool erarbeitet, mit welchem die Verknüpfung von Archivgut und sachthematischer Systematik vorgenommen werden kann. Daneben wird zum anderen in einem experimentellen Arbeitspaket ein Algorithmus erarbeitet, der die vorhandenen archivischen Erschließungsdaten analysiert, mit der Systematik und externen Datenquellen abgleicht und automatisiert Empfehlungen für die Einsortierung des Archivguts in die Systematik gibt. Im Ergebnis sollen damit umfangreiche Mengen Archivgut aufbereitet, mit sachthematischen Informationen angereichert und im Themenzugang der Forschung zur Verfügung gestellt werden. Die entwickelten Methoden sollen später auf weitere thematische Zugänge übertragen und nachgenutzt werden können.

AP 1: Entwicklung eines Metadaten-Modells zur sachthematischen Referenzierung und Anpassung der Datenhaltung

Im Rahmen des bisherigen Projektverlaufs wurde eine umfassende Konzeption der neuen datenhaltenden Schicht für die angereicherten Metadaten erarbeitet. Diese Schicht ermöglicht die Einrichtung einer beliebigen Anzahl verschiedener Themenzugänge innerhalb der bestehenden Dateninfrastruktur der Deutschen Digitalen Bibliothek (DDB) beziehungsweise des Archivportals-D und ist für zahlreiche weitere Projekte nachnutzbar und anschlussfähig. Innerhalb dieser Themenzugänge können eine Vielzahl von Systematiken als hierarchische Facetten zur sachthematischen Recherche angelegt werden. Über Indexbegriffe werden die Erschließungsdatensätze mit der Systematik verknüpft. Indexbegriffe werden jeweils mit einer oder mehreren einschlägigen Referenzen zu verschiedenen Normdateien (Voraussetzung ist mindestens eine Referenz zur Gemeinsamen Normdatei, GND) sowie mit der Angabe verschiedener Synonyme gespeichert. Hierdurch wird das Retrieval verbessert, die Einbindung der Erschließungsdatensätze ins Semantic Web ermöglicht und die Bereitstellung eines kontrollierten Vokabulars abgesichert. Die Verknüpfung von Erschließungsdatensätzen mit

Indexbegriffen ist mit einem Verknüpfungsstatus- und Rollenmodell verbunden, welches zum einen die Anbindung des in AP 5 entwickelten Vorschlagsalgorithmus und zum anderen den Aufbau eines Redaktionsworkflows für die nutzerseitige Anreicherung ermöglicht, sodass User Generated Content bereitgestellt, überprüft und freigegeben werden kann.

Daneben wurde eine Reihe von Endpunkten definiert, an denen das Frontend der neuen Rechercheinfrastruktur sowie das zur Produktivreife weiterzuentwickelnde Indexierungstool die Daten abrufen können. Besonders komplex war dabei die Modellierung der Beziehung zwischen den hierarchisch gegliederten archivischen Erschließungsdatensätzen und der Hierarchie von Indexvokabular und Themensystematik. Hier mussten zunächst die verschiedenen Use Cases der Indexierung im künftig umzusetzenden Indexierungstool definiert und abgegrenzt werden.

Aktuell findet die Implementierung der neuen Datenhaltung und -prozessierung statt. Nach dem Einspielen der Themensystematik und des Themenvokabulars in die Vokabularverwaltung xTree[1] und nach dem Import erster indexierter Bestände in die Datenhaltung, wird das Konzept im Testbetrieb überprüft und danach für das Produktivsystem freigegeben. Ein Daten-Rückkanal zur Re-Integration angereicherter Metadaten in die Ursprungssysteme der datenliefernden Einrichtungen ist in Form eines einfachen csv-Exports vorgesehen, da dieser von allen gängigen Archivsoftwares importiert werden kann. Die Erweiterung des API[2]-Umfangs der DDB zur Ausgabe der angereicherten thematischen Metadaten bzw. von thematischen Daten-Subsets soll bis zum Ende der laufenden Projektphase durchgeführt werden.

Eine bedeutende Herausforderung in diesem wie auch den im Folgenden genannten Arbeitspaketen ist die Tatsache, dass es trotz wiederholter Bemühungen leider nicht gelungen ist, die beantragte Projektstelle bei der Deutschen Nationalbibliothek (DNB) zu besetzen. Die erforderlichen Aufgaben der nicht besetzten Projektstelle mussten daher durch die Projektstellen der anderen Projektpartner und die im Projektantrag genannten Eigenleistungsanteile aller Projektpartner aufgefangen werden. Dies führte notwendigerweise zu Verzögerungen in den einzelnen Arbeitspaketen, welche im Rahmen der nun beantragten zweiten Projektphase ausgeglichen werden sollen.

AP 2: Erstellung einer Referenz-Sachklassifikation für den Themenbereich „Weimarer Republik“ und intellektuelle Zuordnung von Beständen

Ziel des Arbeitspakets ist es, eine hierarchisch gegliederte Themensystematik zum Themenkomplex “Weimarer Republik” zu erstellen. Diese Systematik soll einerseits im Rahmen des zu entwickelnden Themenzugangs einen intuitiven, thematisch gegliederten Zugriffspunkt auf unterschiedliches Archivgut aus verschiedenen Archiveinrichtungen bieten und andererseits als Referenz für die Implementierung künftiger Themenzugänge (beispielsweise zu anderen historischen Epochen und Themengebieten, Quellengattungen etc.) dienen.

Hierzu wurden zunächst sowohl verschiedene Klassifikationen aus verwandten Themengebieten, aus der Weimar-spezifischen Fachliteratur und anderen Portalen (z.B. die im Projektantrag erwähnte Klassifikation aus LEO-BW[3]) als auch übergreifende Klassifikationen (Sachgruppen der GND, Dewey Decimal Classification, Library of Congress Subject Headings etc.) analysiert. Ebenso wurden Möglichkeiten geprüft, die bisherige archivische Indexierungspraxis bei der Erstellung der Systematik nachzunutzen. Als Ergebnis wurde eine kombinierte Systematik konzipiert. Sie besteht zum einen aus einer möglichst übergreifenden

[1] xTree ist eine im Rahmen des digiCULT-Verbunds entwickelte Vokabularverwaltungssoftware.
[2] Application Programming Interface, s. https://api.deutsche-digitale-bibliothek.de/.
[3] https://www.leo-bw.de/themenmodul/von-der-monarchie-zur-republik (Abruf 13.11.2019).

und damit nachnutzbaren historischen Themensystematik und zum anderen aus einem aus der Erschließungspraxis abgeleiteten kontrollierten Vokabular an Indexbegriffen, über welches Objekte aus dem Archivportal-D an die Systematik angebunden werden. Die Konzeption der Systematik erfolgte unter Beratung der im Projektantrag genannten Vokabularexpertin Jutta Lindenthal, welche außerdem an der Einspielung der Systematik in die Vokabularverwaltungssoftware xTree und der Definition der dazu erforderlichen Ein- und Ausgangsformate beteiligt ist.

Zentral war von Beginn an der Ansatz, bei den Indexbegriffen jeweils die GND zu referenzieren. Hierdurch wird einerseits eine stärkere Kontrolle bei der Erstellung des Vokabulars ermöglicht und andererseits das Potenzial einer Anbindung der Erschließungsdatensätze an das Semantic Web genutzt. Über die einheitliche nachgelagerte Anreicherung der archivischen Erschließungsdaten mit den Normdaten soll die Auffindbarkeit der archivalischen Quellen für die Forschung deutlich erleichtert werden. Die Teilnahme an einer in der Deutschen Nationalbibliothek eigens abgehaltenen GND-Schulung diente der kohärenten Abgrenzung der zu verwendenden Entitätstypen. Im Rahmen der Schulung wurde daneben zum einen die systematische Überarbeitung von Indexbegriffen abgestimmt, die noch keinen Eintrag in der GND verfügen, zum anderen wurden Möglichkeiten der Neuanlegung dieser Indexbegriffe in der GND geprüft und zum Ende der ersten Projektphase hin in die Wege geleitet.

Im Rahmen eines gemeinsamen Workshops mit Vertreterinnen und Vertretern aus Geschichtswissenschaft, Digital Humanities und Archivwesen wurden Systematik und Indexbegriffe einer kritischen Evaluation durch Forschung und Praxis unterzogen. Gemeinsam mit den Erfahrungen aus der beginnenden Indexierungspraxis wurde das wissenschaftliche Feedback des Workshops zur Grundlage der abschließenden Überarbeitung der Systematik. Herausforderungen bei der endgültigen Erstellung von Systematik und Indexvokabular bildeten vor allem das Festlegen der inhaltlichen Grenzen und der semantischen Kategorisierung des Vokabulars sowie die Erweiterung um einige noch fehlende Indexbegriffe. Bei der beginnenden Indexierung wurden gezielt Beispiele aus möglichst unterschiedlichen archivischen Beständen ausgewählt, um von Anfang an eine möglichste breite inhaltliche Abdeckung des Themenkomplexes zu erreichen und gleichzeitig besonders effektiv nutzbare Trainingsdaten für AP5 zu erzeugen. Die Erfahrungen aus diesem Prozess können später zur Erstellung von Guidelines für weitere Themensystematiken nachgenutzt werden.

Als Ergebnis konnten eine komplexe Themensystematik zum Thema “Weimarer Republik” mit stabilen Kategorien und Indexbegriffen sowie ein breites Spektrum an Trainings- und Validierungsdaten für AP 5 erstellt werden. Die Themensystematik wird im weiteren Verlauf der ersten Projektphase in die Vokabularverwaltungssoftware xTree und auf “xTree public”[4] eingespielt, was eine erste wissenschaftliche Nachnutzbarkeit des Vokabulars ermöglicht.

Der Workshop war insgesamt von einem positiven Feedback zur Systematik und zum Vokabular geprägt. Grundlegende Diskussionspunkte wurden nicht vorgebracht, die formulierten Kritikpunkte und Anregungen ließen sich unproblematisch umsetzen. Nach übereinstimmender Auffassung der Projektmitglieder soll der zweite beantragte Workshop daher zur Evaluierung des geplanten Indexierungstools umgewidmet werden (AP 3). Durch die Überprüfung dieses Tools durch die Fachcommunities und die spätere Freigabe findet hierbei natürlich mittelbar auch eine praktische Evaluation der Systematik statt. Mithilfe der Indexierungsbeispiele weiterer Archivinstitutionen soll zukünftig eine inhaltlich umfassende Abdeckung der einzelnen Themenbereiche des Themenkomplexes “Weimarer Republik” erreicht und die Involvierung der Systematik- und Indexierungspraxis als ergänzende Methode zur archivwissenschaftlichen Erschließung in Form von Themenportalen aufgrund dieser konkreten Veranschaulichungsbeispiele aus verschiedenen Institutionen nachhaltiger diskutiert und dokumentiert werden.

[4] http://xtree-public.digicult-verbund.de/vocnet/ (Abruf 14.11.2019).

Aufgrund der zuvor zu leistenden Grundlagenarbeit, die durch die Stellenbesetzungssituation bei der DNB erschwert wurde, konnte naturgemäß nur ein Teil der Bestände bis zu diesem Zeitpunkt bereits indexiert werden. Der noch ausstehende Teil der zu indexierenden Archivalien soll im weiteren Verlauf der ersten Projektphase sowie im Laufe der nun beantragten zweiten Projektphase bearbeitet wrden. Durch den Abschluss der Arbeiten in AP 3 und 5 wird die Systematik- und Indexierungspraxis erheblich vereinfacht werden. Zugleich ergibt sich so die Möglichkeit, die Arbeit des in AP 5 entwickelten Algorithmus zu überprüfen und weitere Trainings- und Validierungsdaten bereitzustellen.

Die Konzeption der Systematik erfolgte mittels Eigenleistung von DNB, BArch und LABW sowie durch die Projektstellen von BArch und LABW. Die inhaltliche Definition der Kategorien und Indexbegriffe aufgrund ihrer historischen Inhalte sowie die Überarbeitungen erfolgten gemeinsam durch Eigenleistung und Projektstellen von BArch und LABW.

AP 3: Konzeption und Umsetzung von Klassifikationswerkzeugen zur Pflege von hierarchischen Sachklassifikationen sowie zur Zuordnung von Daten(beständen) zu den einzelnen Klassen

Ziel des Arbeitspakets ist zum einen die Entwicklung eines eigenen Vokabularverwaltungstools bzw. die Überprüfung der Nachnutzbarkeit bestehender Tools, namentlich xTree, für die Erstellung und Pflege der Themensystematik und des Indexbegriffsvokabulars. Zum anderen soll im Rahmen des Arbeitspakets ein Werkzeug entwickelt werden, mit dem Archivbestände und einzelne Objekte im Archivportal-D mit Indexbegriffen aus dem kontrollierten Vokabular zur "Weimarer Republik" verknüpft und damit über die Themensystematik verfügbar gemacht werden können.

Als Vokabularverwaltungstool wurde nach Prüfung anderer proprietärer sowie im Open Source-Bereich zur Verfügung stehender Lösungen der Einsatz von xTree beschlossen. Das Tool entspricht im Unterschied zu anderen Lösungen in weiten Teilen dem Thesaurus-Standard ISO 25964[5], ist bereits seit längerem im DDB-Kontext etabliert und ermöglicht über die Plattform "xTree public" eine Veröffentlichung und damit Nachnutzung des Vokabulars. Nach Abschluss letzter inhaltlicher Arbeiten an Vokabular und Themensystematik werden diese aktuell nach xTree überführt und von dort aus bis zum Ende der ersten Projektphase veröffentlicht.

Die Arbeit am Indexierungstool konnte aufgrund der stellenbesetzungsbedingten Verzögerungen im Bereich des Datenmodells und der Systematik erst verspätet aufgenommen werden. Die Konzeption des Tools findet in einem engen Austausch zwischen BArch und LABW als Anwendern, sowie einem Team aus Interaktionsdesignerin und Entwicklern am FIZ Karlsruhe statt. Die konzeptionellen Arbeiten sind abgeschlossen. Aktuell finden lediglich abschließende Arbeiten am Interaktionsdesign der Benutzeroberfläche statt, sodass die Programmierung des Tools in Kürze beginnen kann und bis zum Beginn der geplanten zweiten Projektphase in einen ersten Prototyp münden soll. An diesem Gesamtprozess sind sowohl die verschiedenen Projektstellen als auch das im Rahmen des Eigenanteils in das Projekt eingebrachte Personal beteiligt.

Um große Mengen an Archivalien möglichst intuitiv und mit geringem Aufwand in der Themensystematik verfügbar zu machen, erhält das Indexierungstool eine grafische Benutzungsoberfläche, mit der auf einfache Weise Indexbegriffe per Drag and Drop den verschiedenen archivalischen Objekten zugeordnet werden können. Dabei ist es auch möglich, übergeordnete Hierarchieknoten bis hin zu ganzen Beständen oder der Archivtektonik mit Indexbegriffen zu verknüpfen und diese Verknüpfung an untergeordnete Objekte zu vererben. Dies entspricht der im Projektantrag formulierten Idee einer einfachen, einheitlichen nachgelagerten sachthematischen Anreicherung der archivischen Erschließungsdaten. Anfang

[5] https://www.niso.org/schemas/iso25964 (Abruf 14.11.2019).

2020 ist hierzu ein Workshop geplant, um anhand eines Klickdummies des zu entwickelnden Tools gemeinsam mit künftigen Anwendern aus verschiedenen Fachcommunities die Funktionalität des Tools kritisch zu überprüfen und zu verbessern.

Zum Ende der aktuellen Projektphase wird die Funktionalität des Prototyps mittels der Bestände des Bundesarchivs und des Landesarchivs Baden-Württemberg überprüft. Das Tool soll darüber hinaus in der zweiten Projektphase für andere Archiveinrichtungen freigegeben und damit als Werkzeug für die kollaborative Erschließung und Aufbereitung von Erschließungsdaten für die Nutzenden geöffnet werden. Ebenso soll in dieser zweiten Projektphase der aktuell in AP 5 in Entwicklung befindliche Algorithmus zur automatischen Indexierung als Vorschlagsfunktion in das Tool integriert werden. Nach Abschluss aller Entwicklungsarbeiten am Ende der zweiten Projektphase soll das Tool Open Source zur Nachnutzung bereitgestellt werden. Ebenso soll das Werkzeug bei allen künftigen Themenzugängen im Archivportal-D und in der DDB eingesetzt werden.

AP 4: Konzeption und Realisierung des sachthematischen Recherchezugangs im Frontend des Archivportals-D

Ziel des Arbeitspakets ist es, einen sachthematischen Themenzugang im Frontend des Archivportals-D zu realisieren. Der Themenzugang soll sowohl als thematisch spezifischer, nutzerfreundlicher Rechercheeinstieg mit der thematischen Kontextualisierung von Archivgut auf Treffer- und Merklisten sowie Detailseiten fungieren als auch als visueller und infrastruktureller Prototyp für die Implementierung künftiger Themenzugänge dienen. Des Weiteren soll eine umfassende Überarbeitung des Interaktionsdesigns des Archivportals-D stattfinden, um den Themenzugang als Teil von dessen gestalterischem Gesamtkonzept zu etablieren.

Zunächst wurden die speziell für den Themenzugang zu konzipierenden Seiten (Startseite, Suchergebnisseite, Objektseite) erstellt. Hierzu fand zunächst eine Analyse und anschließende Überarbeitung des aktuellen Archivportal-D-Frontends sowie vorläufiger, zu Beginn der ersten Projektphase erstellten Entwürfe, statt. Danach wurden die für den Themenzugang spezifischen Elemente (Umsetzung des Themen- bzw. Klassifikationsbaums und der Themenfilter, Einstiegsmöglichkeiten in den Themenzugang über die Startseite des Archivportals-D, Design) umgesetzt und in die überarbeiteten Entwürfe eingepflegt. Als Ergebnis steht ein umfassendes Konzept für das Frontend des exemplarischen Themenzugangs „Weimarer Republik“ zur Verfügung, welches die entwickelte Systematik angemessen präsentiert und ihre sachthematischen Inhalte intuitiv recherchierbar macht. Die Einbindung verschiedener themenspezifischer Recherchefacetten (zunächst: sachthematisch und geografisch) wurden hierbei berücksichtigt.

Zentral war von Beginn an der Ansatz, ein für alle Bereiche des Archivportals-D anpassbares und damit nachnutzbares Frontend zu erstellen, welches zudem moderne Nutzererwartungen aufgreift. Durch das Einbinden des Themas Frontend in den Systematikworkshop (AP 2) konnten im Vorfeld wichtige Anregungen und Erfahrungen aus einem diversen Fachpublikum gesammelt werden. Usability und Accessability sollen durch eine übersichtliche Gestaltung der Webseite, interaktive Elemente im Themenzugang und eine eindeutige Visualisierung der verschiedenen Funktionen erreicht werden. So wurde eine für andere Themen adaptierbare Startseite für den Themenzugang mit Auswahl- und Sortierfunktionen der Systematik sowie interaktiven Informationsbereichen und Beispielen konzipiert. Durch verschiedene optische Elemente (Banner, Farbgebung) soll den Userinnen und Usern die inhaltliche Differenzierung zwischen Themenzugang-Bereich und dem übrigen Portalangebot erleichtert sowie allgemein die Funktion des Archivportals-D als zentralem Nachweisportal der Erschließungsdaten deutscher Archive transparent gemacht werden.

Zur Einbindung der hierarchischen Themensystematik in die bisherigen Such- und Filterfunktionen wurden diese (in auch für künftige Themenzugänge adaptierbarer Form) neugestaltet. Zur gleichzeitigen Vereinfachung und Qualitätssteigerung der Recherche wurden ausgehend von einer Analyse der Zugriffshäufigkeiten und der Nutzungsintensität ungenutzte oder ungünstig platzierte Funktionen entfernt (Archivauswahl nach Alphabet) oder neu positioniert (Hilfe) sowie inhaltliche Ergänzungen vorgenommen (z. B. die detaillierte Ansicht der Erschließungshierarchie auf der Objektseite).

Aktuell findet die Implementierung des Frontends im Rahmen eines intensiven Scrum-Prozesses statt. Die Anpassungen an den bisherigen Archivportal-D-Seiten, die zur Umsetzung des Themenzugangs erforderlich sind, sind bereits weitgehend abgeschlossen. Die Implementierung des eigentlichen Themenzugangs erfolgt nach dem Abschluss der Entwicklung der zugehörigen Datenhaltung. Das Release der neuen Rechercheinfrastruktur ist für Mai 2020 geplant. Die Konzeption des Frontends erfolgte in Zusammenarbeit der Projektstellen und der durch Eigenleistung finanzierten Stellen bei BArch und LABW mit den in Eigenleistung finanzierten Stellen bei FIZ Karlsruhe (FIZ 1). Die Umsetzung des Frontends geschieht durch Eigenleistung und Projektstellen bei FIZ Karlsruhe gemeinsam mit den Projektstellen bei BArch und LABW.

AP 5: Definition, Test und Implementierung eines Algorithmus zur Generierung automatisierter Zuordnungsempfehlungen

Ziel des Arbeitspakets ist es, den Umfang der sachthematisch zuordnenbaren Archivbestände signifikant zu erhöhen, indem die manuelle Verknüpfung durch algorithmisch ermittelte Vorschläge (recommendations) unterstützt wird. Hierzu werden die Möglichkeiten einer (teil-) automatisierten Zuordnung zur Referenz-Sachklassifikation konzipiert, implementiert und evaluiert.

Die im Rahmen von AP2 erstellte Systematik wurde mit Hilfe von SPARQL über die vorhandenen GND-IDs und die Schlagwort-Bezeichnern mit externen Datenquellen verknüpft (Wikidata, Wikipedia). Um weitere externe Datenquellen berücksichtigen zu können, wurden Synonyme der Schlagwörter über Wikipedia-Redirect-Seiten und die entsprechenden Felder der marc21-Datei der zugeordneten GND-Ressourcen automatisiert ermittelt. Die Qualität der Verknüpfung mittels Bezeichner/ Synonym wurde mit Hilfe der GND-ID Verknüpfungen evaluiert. Die Evaluierungsergebnisse machten deutlich, dass auf Grund sprachlicher Mehrdeutigkeiten eine manuelle Überprüfung der mittels Bezeichner/Synonym verknüpften Entitäten zusätzlich durchgeführt werden musste. Die Zuordnung der externen Datenquelle wurde AP2 zur abschließenden Überarbeitung der Systematik bereitgestellt.

Für die Generierung der semantischen Feature wurden die textuellen Attribute Titel, Bestandseinleitungen und Enthältvermerk herangezogen, da diese Attribute ebenfalls für die manuelle Klassifikation verwendet werden. Hierbei stellt der Titel das wichtigste Attribut für die Generierung der semantische Feature dar, da unter anderem die Bestandseinleitung nur für 1% der Dokumente und der Enthältvermerk lediglich für 28% der Dokumente vorhanden sind. Allerdings zeigte sich, dass ein signifikanter Anteil der Titel keinen oder nur einen geringen Bezug zum tatsächlichen Inhalt aufwies. So sind 31% der Titel nicht relevant für die Klassifikation, da der entsprechende Titel lediglich die Bandnummer ("Bd. 1”, “Bd. 2”, usw.) enthält.

Auf Grund dieser Eigenschaften der textuellen Attribute muss zusätzlich die Hierarchie der Archivbestände für die Generierung der Feature mit einbezogen werden. Zu diesem Zweck wurde der hierarchische Zusammenhang der Dokumente aus dem Kontext-Attribut der Archivbestände extrahiert und für die Feature-Generierung miteinbezogen.

Zur Ermittlung der semantischen Feature der Archivbestände wurden bislang die wortbasierten Verfahren Word2Vec[6] und fastText,[7] welche mit Hilfe der (deutschen) Wikipedia und Common Crawl trainiert wurden, untersucht und mit TF-IDF verglichen. Beide Verfahren erzeugen eine semantische Repräsentation für jedes Wort, basierend auf der Wort-Kookkurrenz innerhalb des Trainingscorpus. So werden Wörter, die oftmals in einem ähnlichen Kontext auftauchen, zu ähnlichen Vektoren konvertiert. Das Word2Vec-Verfahren wies einen hohen Anteil an unbekannten, sprich nicht trainierten Wörtern zwischen 20% und 30% für alle Textinhalte der Archivbestände auf. FastText hingegen analysiert auch die Morpheme der Wörter, um eine akkurate semantische Repräsentation auch für unbekannte Wörter zu erhalten.
Diese Wort-Vektoren werden anschließend über die Bildung des Mittelwertes normalisiert. Dieser Normalisierungsschritt gewährleistet eine einheitliche Darstellung der textuellen Attribute der Dokumente unabhängig von der Wortanzahl. Neben den beschriebenen textuellen Attributen wurde der hierarchische Aspekt der Dokumente ebenfalls in die Generierung der Feature durch einen gewichteten Mittelwert aller textuellen Feature der übergeordneten Dokumente sowie des Dokumentes selbst miteinbezogen.

Basierend auf den ermittelten semantischen Feature wurde die Klassifikation als sogenanntes “dataless” Verfahren[8] umgesetzt. Hierbei wurden für die Schlagwörter der Systematik gleichermaßen semantische Repräsentation u.a. mit Hilfe der externen Datenquellen erstellt und für jedes Dokument und Schlagwort über die Cosinus-Ähnlichkeit eine Kennzahl für die semantische Ähnlichkeit beider Elemente bestimmt. Ausgehend von diesem Ähnlichkeitswert wurde die Klassifikation mit Hilfe eines Schwellwertes vorgenommen.

Zusammenfassend lässt sich feststellen, dass die bisherigen Modelle insgesamt noch keine zufriedenstellenden Vorschläge, welche zur Realisierung einer (teil-)automatisierten Zuordnung verwendet werden können, liefern und so weitere Forschung hinsichtlich der semantischen Feature nötig ist. Im Rahmen der zweiten Projektphase sollen weitere Arbeiten in diesem Bereich vorgenommen werden. Die vergrößerte Datengrundlage durch die Fortführung der Arbeiten in AP 2 und die Beteiligung weiterer Archiveinrichtungen soll dabei zur Verbesserung der Ergebnisse beitragen.

AP 6: Projektkoordination, Öffentlichkeitsarbeit, Evaluation und Kooperation mit der DDB-Servicestelle
Zur Kommunikation und Koordination innerhalb des Projektteams wurden eine projektinterne Mailingliste und ein internes Projektwiki aufgebaut. Zentrale Informationen, Prozesse und Vorhaben mit Relevanz für das gesamte Projektteam werden über die Mailingliste kommuniziert. Im Wiki werden in verschiedenen Unterbereichen zu den einzelnen Arbeitspaketen Dokumente ausgetauscht sowie Besprechungen sowohl auf Ebene der einzelnen Arbeitspakete als auch auf Ebene des Gesamtteams vor- und nachbereitet und protokolliert.

Bis zum Berichtszeitpunkt fanden mehrere Projekttreffen statt, die seitens der Projektkoordination vor- und nachbereitet wurden. Auf diesen Treffen konnten zum einen übergreifende Fragen und die Zeitplanung besprochen werden, zum anderen wurde auf diesen Treffen auch Zeit für intensive Abstimmungen zu einzelnen Arbeitspaketen bereitgehalten. Absprachen und Planungen innerhalb der Arbeitspakete werden daneben in Telefon- und Videokonferenzen durchgeführt, die in der Regel von der Projektkoordination begleitet werden. Eine intensive Zusammenarbeit ergab sich mit der DDB-Fachstelle Archiv, die der Projektkoordination als zentraler Ansprechpartner in technischen Fragen dient.

Die technische Entwicklung des Frontends und des Verknüpfungstools bei FIZ Karlsruhe findet im Rahmen eines agilen Scrum-Prozesses statt. Die Projektkoordination fungiert hier als

[6] https://arxiv.org/abs/1301.3781 (Abruf 14.11.2019).
[7] https://arxiv.org/abs/1607.04606 (Abruf 14.11.2019).
[8] http://new.aaai.org/Papers/AAAI/2008/AAAI08-132.pdf (Abruf 14.11.2019).

Product Owner, formuliert die Ziele der Arbeit in sogenannten User Stories, aus welchen die Entwicklerinnen und Entwickler konkrete Arbeitsaufträge ableiten, steht für Rückfragen zur Verfügung und nimmt die Ergebnisse ab. Diese Vorgehensweise hat sich im Laufe der bisherigen Arbeiten als sehr produktiv und zielführend erwiesen.

Größere Schwierigkeiten bereitete die Personalsituation aufgrund der nicht möglichen Stellenbesetzung bei der DNB und des überraschenden Todes des technischen Geschäftsführers der DDB. Hinzu kommt der längere krankheitsbedingte Ausfall der Projektleitung beim LABW. Die dadurch entstandenen Zusatzaufwände konnten durch das zur Projektkoordination eingestellte Personal gemeinsam mit den im Projekt als Eigenanteil eingebrachten Personalmitteln aufgefangen werden, hatten jedoch die beschriebenen Verzögerungen in den einzelnen Arbeitspaketen zur Folge.

Öffentlichkeitsarbeit wird im Projekt über verschiedene Kanäle geleistet. Als niedrigschwellige Informationsquelle wird anlassbezogen über den Archivportal-D Twitter-Account sowie die Social Media-Präsenzen der Projektpartner auf das Projekt aufmerksam gemacht. Hinzu kommen Berichte über Projekttreffen und den Systematik-Workshop auf der Archivportal-D Webseite.[9] Ein größerer Beitrag wurde außerdem für den Newsletter der DDB[10] erstellt, ein kurzer Bericht wurde für die Archivnachrichten des Landesarchivs Baden-Württemberg[11] verfasst.

Zur Kommunikation des Projekts in den verschiedenen Fachcommunities wurden bisher zwei Posterpräsentationen (Bibliotheca Baltica 2018 in Rostock, DHd Konferenz 2019 in Frankfurt am Main) sowie ein Vortrag im Rahmen einer Talk-Session auf der GNDCon 2019 in Frankfurt am Main durchgeführt. Daneben wurde auf dem Schleswig-Holsteinischen und dem Sächsischen Archivtag 2019 im Rahmen eines Standes und zweier Vorträge auf das Archivportal-D allgemein sowie auf das Projekt hingewiesen. Auf dem Deutschen Archivtag 2019 in Suhl wurde außerdem im Rahmen einer Fortbildung zu Erschließungs- und Normdaten im Archivportal-D vertieft auf das Projekt eingegangen. Ein fachwissenschaftlicher Beitrag konnte außerdem in der Fachzeitschrift ARCHIVAR platziert werden.[12] Diese Öffentlichkeitsarbeit wurde entsprechend dem Projektantrag sowohl mit Personal aus Eigenmitteln, als auch mit Personal aus Projektmitteln geleistet.

Rückmeldung zur Öffentlichkeitsarbeit ging entsprechend des bisherigen Fokus auf die Fachöffentlichkeit vor allem aus der Archivcommunity ein. Die sachthematische Zugänglichmachung von Archivgut für die Forschung und die allgemeine Öffentlichkeit wurde dabei sehr begrüßt. Verschiedene Archivverwaltungen haben ihr Interesse an einer Beteiligung am Themenzugang “Weimarer Republik” signalisiert.[13] Unterschiedliche Archivsparten haben daneben bereits ihr konkretes Interesse an der Nachnutzung der Themenzugänge im Archivportal-D für eigene sachthematische Zugänge bekundet,[14] was dem formulierten Ziel einer stärkeren Öffnung des Portals innerhalb der Community und einer Verbesserung der Zugangsmöglichkeiten zu Archivgut für die Nutzenden entgegenkommt.

---

[9] https://www.archivportal-d.de/info/aktuelles/Kick-off_sachthematische_Zugaenge/, https://www.archivportal-d.de/info/aktuelles/Bericht_Workshop_Weimar_BArch/ (Abruf 14.11.2019).
[10] https://www.deutsche-digitale-bibliothek.de/content/journal/aktuell/archivportal-d-bald-mit-thematischem-zugang-zur-weimarer-republik?pk_campaign=nl_feb19 (Abruf 14.11.2019).
[11] Nils Meyer: Archivportal-D bald mit thematischem Zugang, in: Archivnachrichten 58 (März 2019), S. 38, https://www.landesarchiv-bw.de/sixcms/media.php/120/64544/Archivnachrichten_58_Webversion.pdf (Abruf 14.11.2019).
[12] Nils Meyer: Sachthematische Zugänge im Archivportal-D. Archive und ihre Bestände zusammenführen und neu entdecken, in: Archivar 72, Heft 1 (Februar 2019), S. 37-38, http://www.archive.nrw.de/archivar/hefte/2019/Ausgabe-1/Archivar-1_2019.pdf (Abruf 14.11.2019).
[13] Siehe die Letters of Intent des Niedersächsischen Landesarchivs und des Landesarchivs Thüringen.
[14] Siehe zum Beispiel den Letter of Intent des Archivs des Deutschen Museums sowie die weiteren an das Projektteam herangetragenen Vorschläge im Antrag zur zweiten Projektphase.

Ein wichtiges Arbeitsgebiet im Bereich Öffentlichkeitsarbeit für das Ende der ersten Projektphase und insbesondere die zweite Projektphase wird die Bewerbung des neuen Themenzugangs in der Öffentlichkeit und in der (geschichts-) wissenschaftlichen Fachöffentlichkeit sein. Sinnvollerweise soll diese Bewerbung in Richtung der Nutzenden erst nach der Fertigstellung des Frontends des neuen Recherchezugangs erfolgen.

Die Dokumentation und Veröffentlichung der Projektergebnisse wird auf einer eigenen Seite im Archivportal-D erfolgen, auf der auch die verschiedenen im Rahmen der zweiten Projektphase zu erstellenden Materialien (z.B. die Guidelines zur Systematikerstellung und die Best Practices zur Indexierung) zum Download angeboten und Links zu den Open-Source-Repositories der entwickelten Tools veröffentlicht werden sollen.

3 Zusammenfassung

Nicht zutreffend, da Zwischenbericht.

4 Weitere Arbeiten und Planungen

AP 1: Entwicklung eines Metadaten-Modells zur sachthematischen Referenzierung und Anpassung der Datenhaltung

- Implementierung des Konzepts zur Datenhaltung und Datenprozessierung
- Import der Systematik aus xTree sowie der ersten bereits Indexierten Bestände
- Prüfung der Funktionalität der neuen Datenbanktabellen für die angereicherten Metadaten sowie der Endpunkte zum Frontend

AP 2: Erstellung einer Referenz-Sachklassifikation für den Themenbereich „Weimarer Republik“ und intellektuelle Zuordnung von Beständen

- Import der Systematik in das Vokabularverwaltungstool xTree und Veröffentlichung über xTree Public
- Indexierung einschlägiger LABW- und BArch-Bestände
- Lieferung von Beispielindexierungen an AP 5

AP 3: Konzeption und Umsetzung von Klassifikationswerkzeugen zur Pflege von hierarchischen Sachklassifikationen sowie zur Zuordnung von Daten(beständen) zu den einzelnen Klassen

- Durchführung eines Workshops zur Toolfunktionalität (02/2020)
- Umsetzung der Entwicklung eines Prototypen des Indexierungstools anhand bestehender Mockups

AP 4: Konzeption und Realisierung des sachthematischen Recherchezugangs im Frontend des Archivportals-D

- Fertigstellung und Tests der Frontendimplementierung für die angereicherte Metadatenschicht (Startseite Themenzugang, neue Filterfunktionen in der Suche, neue Indexbegriffe)

AP 5: Definition, Test und Implementierung eines Algorithmus zur Generierung automatisierter Zuordnungsempfehlungen

- Weitere Verbesserung der Klassifikationsergebnisse durch:
    - Einbeziehung weiterer Attribute
    - Optimierung der Kombination von textuellen und hierarchischen Attributen

    - Einsatz von weiteren Modellen, wie kontextbasierten Modellen, zur Generierung semantischer Feature
- Überprüfung der Referenztexte, welche zur Generierung der Schlagwort-Repräsentationen herangezogen werden, in Zusammenarbeit mit BArch und LABW

AP 6: Projektkoordination, Öffentlichkeitsarbeit, Evaluation und Kooperation mit der DDB-Servicestelle

- Öffentlichkeitsarbeit, insbesondere in der (geschichts-) wissenschaftlichen Fachöffentlichkeit
- Anstoßen der Beteiligung weiterer Archive am Themenzugang “Weimarer Republik”
- Dokumentation der Ergebnisse der ersten Projektphase
- Abschlussevaluation der ersten Projektphase (04/2020 bis 05/2020)

Im Rahmen der aktuell beantragten zweiten Projektphase sollen die Ergebnisse der ersten Projektphase nachhaltig gesichert und erweitert werden, um die Nachnutzung der entwickelten Infrastruktur zu ermöglichen.

5 Veröffentlichung von Daten aus Abschlussberichten

Nicht zutreffend, da Zwischenbericht.

6 Weitere Bemerkungen zum Vorhaben/Anregungen etc.

Keine.

7 Unterschrift(en)

Siehe Anlage.

8 Verzeichnis der Anlagen

Anlage 1: Anlage zu Punkt 7 Unterschrift(en)
Anlage 2: Entwurf Themenzugang „Weimarer Republik“
Anlage 3: Entwurf Indexierungstool