Evaluierungskonzept für SoNAR (IDH)

Sina Menzel & Vivien Petras

Humboldt-Universität zu Berlin

- Version I: September 2019 -

Abstract

Dieses Dokument beschreibt das Konzept zur Qualitätssicherung im Projekt SoNAR (IDH). Dabei werden Ziele und methodisches Vorgehen der Evaluierungen im Projektverlauf beschrieben.

Inhalt

1. Einleitung

Die Begleitforschung im Projekt SoNAR (IDH) dient der Qualitätssicherung der einzelnen Projektschritte und obliegt dem Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin in Arbeitspaket 4 (AP4). Das vorliegende Konzept beinhaltet die Planung zur Umsetzung der Evaluierungsziele gemäß des Projektantrags (S. 14).

Die Evaluierung erfolgt projektbegleitend und iterativ. Maßgeblich dafür sind insgesamt vier neuralgische Punkte, die in Teilschritten untersucht werden. Das vorliegende Dokument differenziert diese Teilschritte aus und definiert Ziele sowie konkretes methodisches Vorgehen. Dabei bauen die Evaluierungsschritte z.T. methodisch aufeinander auf. Die Ergebnisse der Evaluierungen werden mit den betreffenden Projektpartnern ausgetauscht und dadurch Impulse und für einen erfolgreichen Projektverlauf gegeben (vgl. Stiller et al. 2016). Das vorliegende Dokument ist dynamisch angelegt und wird im Projektverlauf angepasst werden. Dadurch wird das flexible Reagieren auf Projektentwicklungen ermöglicht. Es wird sichergestellt, dass allen Projektpartnern stets die aktuellste Version bereitsteht, Versionierungen werden mit Datum gekennzeichnet.

2. AP4-2 Evaluierung I: Datennormalisierung

2.1 Ziel

Die Evaluierung I bezieht sich auf die Ergebnisse des Arbeitsschrittes AP1-1. Sie beantwortet die Frage, ob die Normalisierung der für SoNAR (IDH) vorliegenden Daten in ein einheitliches Datenformat sowie deren Migration in eine neue Zieldatenbank ohne Daten- und Potenzialverluste erreicht wurde. Dies bezieht sich stets auf die Gesamtmenge der für das Projekt bereitgestellten Daten und wird gemessen im Vergleich zu den Rohdaten.

2.2 Vorbereitung

Basis sämtlicher Evaluierungen der Teilergebnisse im Projekt SoNAR (IDH) ist die Ausgangslage der zugrunde liegenden Daten. Da diese aus diversen Repositorien extrahiert sind, gilt es, die Eigenschaften und statistischen Kennzahlen der Daten in der Aufbauphase auszudifferenzieren.

Für diesen Arbeitsschritt wird AP1 statistische Auswertungen der Daten bereitstellen. Es wird eine Ausdifferenzierung der verschiedenen Datentypen angestrebt: Normdaten, Metadaten und Volltexte. Diese werden differenziert nach:

- der Anzahl der Datensätze (MARC 21) bzw. Teile der Findbücher (EAD) gemessen an den Kontrollnummern (*unique identifiers*);
- der Anzahl und Namen der Felder (MARC 21) und Pfade (EAD) mit absoluten und relativen Häufigkeiten;
- dem Umfang der einzelnen Datensätze und Datenpfade gemessen an den Textzeichen in den befüllten Datenfeldern;
- der Anzahl der Ausgaben, Dateien und Identifier (Volltexte).

Die vorliegenden Volltextdaten sind durch OCR oder manuelle Transkription zwar maschinenlesbar, jedoch nicht fehlerfrei erschlossen. Da es sich um historische Texte handelt, die im Original z.T. in Frakturschrift vorliegen, kann mit einer höheren Rate an OCR-Fehlern gerechnet werden (Kann/Hintersonnleitner 2015, S.76). Dementsprechend werden von AP1 nach aktuellem Stand Normalisierungen vorgenommen (z.B. Ausgleich gängiger OCR-Fehler), die in regelmäßigen Abständen kommuniziert und erörtert werden[1].

[1] Eine dezidierte Evaluierung der OCR-Normalisierung ist nur bei ausreichender Kapazität vorgesehen.

Sich ergebende Datenreduktionen und -bereinigungen durch AP1 werden abgestimmt und festgehalten. Dies betrifft auch bereits vorhandene Inkonsistenzen und Fehler in den Rohdaten, die im Vorfeld festgestellt werden. Entsprechende Verringerungen der Datenmenge werden mit allen Projektpartnern abgestimmt und in der Evaluation berücksichtigt.

2.3 Vorgehen

Laut Baierer et al. ist der *mapping approach* entscheidend für die Qualität der resultierenden Normalisierungen (2014, S. 2). Daher findet schon während des Datenmappings ein engmaschiger, bilateraler Austausch mit AP1 statt, sodass die Herangehensweise und die erstellten Algorithmen transparent dokumentiert und kommuniziert werden können.

Basis der Evaluation ist zunächst die formale Überprüfung des einheitlichen Schemas. Dabei wird betrachtet, ob die Gesamtmenge der aufbereiteten Daten in einem einheitlichen, generischen Format vorliegt. Anschließend werden die normalisierten Daten inhaltlich überprüft. Im Zuge des AP1-1 werden in Rücksprache mit AP2 Metadatenfelder festgelegt, die für die HNA maßgebliche Informationen enthalten können (*application profile*). Für die vordefinierten Eigenschaften muss in den normalisierten Daten weiterhin eine Entsprechung vorliegen. Hierüber erfolgt ein formaler Abgleich. Anschließend werden in den normalisierten Daten Auszählungen der eindeutig identifizierbaren Werte vorgenommen (z.B. die Anzahl der Kontrollnummern) und mit den Werten der Rohdaten aus AP4-1 verglichen. Dies lässt Schlüsse über mögliche Datenverluste zu.

Ein wichtiges Kriterium in der Metadatenqualität ist die Vollständigkeit (*completeness*, vgl. Park und Tosaka 2010). Diese bezieht sich auf das Vorhandensein erwünschter Metadateninformationen in den Feldern sowie die Möglichkeit des Zugriffs auf diese Datenfelder in der Nutzung (Bruce und Hillmann 2004, S. 5). Das o.g. *application profile* gilt als Standard für die Evaluierung, anhand dessen der Grad an Vollständigkeit stichprobenartig überprüft werden kann. Dabei wird unterschieden zwischen:

- der Anzahl der vorhandenen und befüllten Felder aus dem Standard in einem Datensatz;
- der Anzahl der befüllten Felder, die persistente Verweise auf Normdaten enthalten;
- dem Umfang eines Datensatzes gemessen an den Textzeichen in den befüllten Datenfeldern.

Zur Ermittlung der Kennzahlen werden Anfragen zu bestimmten Datenfelder und -typen sowohl an die Rohdaten, als auch an die normalisierten Daten gestellt und die Resultate abgeglichen. Die Anfrageergebnisse der Rohdaten bilden dabei den Gold Standard, kommunizierte Datenreduktionen durch AP1 werden berücksichtigt.

3. AP4-3a Evaluierung II-a: Entitätenerkennung

3.1 Ziel

Die Evaluierung II-a zielt auf die Ergebnisse des Arbeitsschrittes AP1-2, in der Entitäten (*named entities*; NE) in den vorliegenden Volltexten automatisiert ausgezeichnet werden sollen. Sie beantwortet die Frage nach der Güte der durch AP1 erarbeiteten NER-Algorithmen (*named entity recognition*). Der Erstellung der dafür notwendigen Gold Standards liegen community-basierte Richtlinien zugrunde, die im Verlauf der Evaluierung mit Blick auf die konkreten Ergebnisse der automatisierten Auszeichnungen durch AP4 iterativ optimiert werden. Die Erstellung der Richtlinien in Abstimmung mit AP1 ist daher ein zweites, separates Teilziel in II-a.

3.2 Vorbereitung

Beachtet werden nach aktuellem Stand und gemäß der vorliegenden Normdaten: Klassen, Personen, Geografika, Körperschaften, Werke, sowie Konferenzen. Ausgezeichnete Entitäten, die keiner dieser

Klassen zugeordnet werden können, werden als Sonstige gekennzeichnet[2]. Es handelt sich in den relevanten Projektdaten um historische Volltexte, das bedeutet, dass unter anderem eine geringe Standardisierung der Orthografie zu erwarten ist (vgl. Labusch et al. 2019). Daher ist die Domänenadaption der NER-Richtlinien durch AP1 von zentraler Bedeutung. Hierüber erfolgt ein kontinuierlicher Austausch.

Für die Evaluierung ist die Erstellung eines Gold Standards notwendig, der manuell annotierte und dadurch maximal präzise Auszeichnungen der Entitäten einer zuvor kuratierten und mit automatisierten NE-Auszeichnungen vorverarbeitete Datenmenge beinhaltet (*automated unsupervised pretagging*). Diese Aufgabe wird durch die im Projekt eingestellte studentische Hilfskraft übernommen (AP4). Die dafür notwendige manuelle Annotation wird mithilfe eines In-House Tools vorgenommen, das durch die SBB bereitgestellt wird. Die Annotation folgt ebenfalls den NER-Richtlinien. Diese werden während des Annotationsprozesses kontinuierlich überprüft und ggf. angepasst (vgl. Oouchida et al. 2009; siehe auch 3.1). Das zugrundeliegende Datenformat für die Annotation der Dokumente ist dem IOB-Format aus dem GermEval Task 2014 angelehnt[3].

3.3 Vorgehen

Die Evaluierung der Güte der NER-Algorithmen erfolgt über die Messung der Fehlerrate in den automatisiert erzeugten Auszeichnungen im Testkorpus. Diese Fehlerrate wird mithilfe der Gold Standards ermittelt und ist ein Indikator für die zu erwartende Fehlerrate auf das gesamte Datensample in SoNAR (IDH)(*expected prediction error rate*, Bengio/Grandvalet 2004, S. 1089). Die Volltexte werden nach erfolgter Auszeichnung in k Einheiten identischer Datenmengen (*folds*) aufgeteilt und gegeneinander evaluiert (*k-fold cross validation*, Arlot/Celisse 2010, S. 53). Jede automatisiert bzw. manuell erzeugte Auszeichnung bildet dabei einen Datenpunkt.

Anhand der sich ergebenden Kennwerte Precision, Recall und F-Score[4] werden anschließend Schlüsse auf die Performanz der automatisiert erzeugten NE gezogen. Die Berechnung erfolgt gemäß der Evaluationsrichtlinien der CoNLL Konferenz 2003, die Precision und Recall exakt an den manuell erstellten Gold Standards misst (*true positives*, Tjong Kim Sang/De Meulder 2003). Zusätzlich werden erweiterte Messungen hinzu gezogen, die auch unvollständig ausgezeichnete Token beachtet (Manning 2006, Batista 2018). Insgesamt werden folgende Fälle unterschieden:

- Korrekte Auszeichnungen (correct: COR), *true positives;*
- Unvollständige Auszeichnungen (partial: PAR), *boundary errors;*
- Verpasste Auszeichnungen (missing: MIS), *false negatives;*
- Falsche Auszeichnungen (spurious: SPU), *false positives*.

[2] Der Umfang der Klassen für die NER ist nach dem aktuellen Projektstand festgelegt. Die Auszeichnung weiterer Klassen (z.B. die Identifizierung von Konzepten im psychologischen, kognitionswissenschaftlichen Sinne) übersteigt nach momentaner Einschätzung die Projektressourcen. Im Falle einer projektweit beschlossenen Erweiterung bleibt das methodische Vorgehen in der Evaluierung hiervon unberührt.

[3] "The IOB format is a simple text chunking format that divides texts into single tokens per line, and, separated by a whitespace, tags to mark named entities. [...] To mark named entities that span multiple tokens, the tags have a prefix of either B- (beginning of named entity) or I- (inside of named entity). O (outside of named entity) tags are used to mark tokens that are not a named entity." (Neudecker 2017, verfügbar mit Beispiel unter: https://github.com/EuropeanaNewspapers/ner-corpora/blob/master/README.md)

[4] Die Kennzahl *Precision* beschreibt die Genauigkeit eines Systems, also hier den Anteil korrekt ausgezeichneter an der Gesamtmenge der ausgezeichneten Entitäten. Der *Recall* hingegen beschreibt die Trefferquote, also den Anteil gefundener korrekt ausgezeichneter Entitäten in Bezug auf alle im entsprechenden Dokument vorhandenen Entitäten. Da sich beide Maße gegenseitig beeinflussen, gibt es als kombinierte Kennzahl den *F-Score*. Dieser ergibt sich aus dem gewichteten harmonischen Mittel von *Precision* und *Recall*, denen jeweils abhängig von den Systemanforderungen gleiche oder unterschiedliche Gewichtung zukommen kann.

Im Fall, dass die Auszeichnung auf zwei Ebenen erfolgt, also eingebettete und sublexische Entitäten[5] ebenfalls erfasst werden (*second level*), wird in der Evaluierung nach erster und zweiter Ebene differenziert und analog zum Evaluationsplan des GermEval Tasks 2014 vorgegangen, der die Messwerte anhand der Ebenen unterscheidet: Ebene 1 separat (nur *first level* NE), Ebene 2 separat (nur *second level* NE), sowie beide kombiniert (Padó 2014, S. 3 f.).

Die ermittelten Performanz-Werte sowie ggf. Optimierungspotenziale werden gemeinsam mit AP1 erörtert. Auf Basis dieser Absprachen wird angestrebt, dass die NER-Algorithmen anhand der Evaluierungsergebnisse angepasst werden und anschließend die automatisierte NE-Auszeichnung der gesamten Volltext-Datenmenge erfolgt.

4. AP4-3b Evaluierung II-b: Entitätenverlinkung

4.1 Ziel

Die Evaluierung zielt ebenfalls auf die Ergebnisse des Arbeitsschrittes AP1-2. Nach erfolgter Auszeichnung der Entitäten kann im zweiten Schritt die weitere Anreicherung der Volltexte durch persistente Verlinkungen der NE auf die entsprechenden Normdatensätze der GND[6] erfolgen, das sogenannte *Named Entity Linking* (NEL). Die Evaluierung II-b beantwortet die Frage nach der Güte der durch AP1 erarbeiteten NEL-Algorithmen.

4.2 Vorbereitung

Voraussetzung für die Errechnung der Güte-Kennzahlen für das NEL ist das Ergänzen der Gold Standards. Gegenstand der Evaluierung ist in diesem Schritt der Anteil korrekter Verlinkungen auf die vorher ausgezeichneten Entitäten. Im Vorfeld erfolgt die automatisierte, unüberwachte Entitätenverlinkung durch AP1. Die durch die automatisiert erzeugten Links ergänzten Gold Standards aus II-a werden daraufhin durch die Hilfskraft in AP4 auf inkorrekte Verlinkungen hin überprüft und bereinigt (Disambiguierung). Dafür wird ebenfalls das o.g. In-House Annotationstool genutzt, dessen Anpassung an den Evaluationsschritt II-b erfogt nach Rücksprache ebenfalls durch die SBB.

4.3 Vorgehen

Auf Basis der ausgebauten Gold Standards kann anschließend analog zum Vorgehen in II-a die Performanz errechnet werden. Kennzahlen sind auch hier Precision, Recall und F-Score. Als Datenpunkte gelten weiterhin die ausgezeichneten Entitäten. Diese werden als Kandidaten für potenzielle Verlinkungen gesehen. Demnach gibt es auch hier mehrere Fälle (Hachey et al. 2012, S. 21): Entitäten, die ausgezeichnet UND mit Links ausgestattet sind (C), sowie Entitäten, die ausgezeichnet, aber NICHT mit Links ausgestattet sind (NIL). Für die Evaluierung werden folgende Fälle betrachtet:

- Korrekt verlinkte korrekt ausgezeichnete Entitäten (correct: COR), *true positives;*

[5] Eingebettete Entitäten sind Entitäten innerhalb von Entitäten, also z.B. *Berlin* (LOC) in *Berliner Mauer* (LOC), LOC zeichnet Geografika aus. Sublexische Entitäten dagegen sind eingebettet in Token, die keine Entität darstellen. Dabei unterscheidet man zwei Typen: Erstens Derivate, z.B. *norddeutsch* (LOC-deriv) und zweitens Komposita, z.B. *Troja* (LOC-part) in *Troja-Ausstellung* (vgl. Padó 2014, S. 1).

[6] Bei einer Erweiterung des Verlinkungs-Schrittes auf weitere Normdatenbanken oder andere Datenquellen wie z.B. Wikidata wird das vorliegende Dokument entsprechend angepasst. Das methodische Vorgehen in der Evaluierung bleibt davon unberührt.

- Falsch verlinkte korrekt ausgezeichnete Entitäten (spurious: SPU), *false positives*[7].

Die errechneten Kennzahlen geben Aufschluss über die Erfolgsrate der angewandten NEL-Algorithmen. Die Ergebnisse und Optimierungspotenziale werden mit AP1 diskutiert.

5. AP4-4 Evaluierung III: Standardisiertes Forschungsdesign

5.1 Ziel

Die Evaluierung zielt auf die Ergebnisse des Arbeitspaketes AP2-1, dem standardisierten Forschungsdesign zur Nutzung der Forschungstechnologie SoNAR (IDH). Die Frage nach dessen Validität kann durch qualitative Erhebungen beantwortet werden. Durch die Erhebung werden darüber hinaus Rückschlüsse auf Bedarf und Umfeld für die Forschungstechnologie erwartet.

5.2 Vorbereitung

Das Forschungsdesign ist standardisiert und damit losgelöst von wissenschaftlichen Einzelthemen. Bei der Überprüfung wird daher zunächst ein Vergleich mit bestehenden Standards geisteswissenschaftlicher Arbeit angestellt (z.B. Stiller et al. 2016, S. 252 f.), um erste Parallelen und Unterschiede herauszustellen. Zusätzlich werden frühzeitig zwei Beobachtungsstudien durchgeführt (Visualisierungsworkshop AP3 und HNR-Workshop AP2), deren Ergebnisse zur Konzipierung der Leitfäden für die spätere Hauptstudie (siehe 5.3) dienen. Zusätzlich findet ein engmaschiger Austausch mit AP2 statt, aus dem ebenfalls geäußerte Anforderungen aus der Fachwissenschaft festgehalten und zur konkreten Studienvorbereitung genutzt werden.

Bei den Workshops sollen ebenfalls erste Kontakte zu FachwissenschaftlerInnen geknüpft werden, die auf dem Gebiet der HNA tätig sind. In allen Beobachtungen werden schriftliche Notizen festgehalten, zusätzlich darf in Absprache mit AP3 das im Visualisierungsworkshop festgehaltene Ton- und Bildmaterial nachgenutzt werden. Das Beobachtungsmaterial wird systematisch ausgewertet (kodiert), die Ergebnisse fließen in die Leitfadenkonzipierung für die Interviewstudie und die Anforderungstests (siehe 6.2) ein.

5.3 Vorgehen

Das standardisierte Forschungsdesign wird durch eine Interviewstudie evaluiert, die gemeinsam mit der Evaluierung der Visualisierung und des Interfacedesigns durch Anwendungstests (AP4-5) am selben ExpertInnen-Sample durchgeführt wird. Zur Skizzierung künftiger Nutzungsszenarien sowie zur Überprüfung von individuellen Forschungsdesigns, die auf der Mikroebene zur Anwendung kommen, wird auf Fallbeispiele zurückgegriffen. Dafür werden FachwissenschafterInnen zum Vorgehen bei eigenen Forschungsarbeiten im Bereich der HNA in Form von teilstrukturierten Leitfadeninterviews befragt (*case studies*, Lazar et al. 2017, S. 153- 185). Ziel der Fallstudien ist das Festhalten exemplarischer Forschungsprozesse und deren Aufbereitung in einzelne Teilschritte, sowie erste Erkenntnisse zu Anforderungen an die Visualisierung. Angestrebt sind ca. 90 min.

[7] Zusätzlich existieren folgende Fälle:

- Verpasste Verlinkungen korrekt ausgezeichneter Entitäten, zu denen eine GND-Referenz existiert (missing: MIS), *false negatives*.
- Fehlende Verlinkungen falsch ausgezeichneter Entitäten, bzw. fehlende GND-Referenzen korrekt ausgezeichneter Entitäten (absent: ABS), *true negatives*.

Als Zusatz wird daher die Überprüfung von fehlenden Verlinkungen in ausgezeichneten Entitäten (false negatives) durch die Studentische Hilfskraft angestrebt. Dies ist aber nur bei ausreichenden Zeitressourcen möglich und wird daher erst im Verlauf des Arbeitsschrittes entschieden.

individuelle Sitzungen mit den ExpertInnen, die in einen Interviewteil (ca. 60 min.) und einen Testteil zur Auswertung des aktuellen Visualisierungsprototyps (ca. 30 min.) aufgeteilt sind (vgl. 6.3).

Als ExpertInnen gelten Personen, die wissenschaftliche Arbeiten im Bereich HNA vorweisen. Die Akquise erfolgt über Kontaktaufnahme während der Beobachtungsstudien (s.o.), über Kontaktvermittlung durch die Projektpartner (SBB, HHU, FHP), über das Netzwerk von AP4 sowie durch die Sichtung geeigneter HNA-Publikationen und die damit verknüpften Publikationsnetzwerke. In Bezug auf den konkreten geschichtswissenschaftlichen Fachbereich der ProbandInnen wird ein möglichst diverses Sample angestrebt. Die Sitzungen werden nach Möglichkeit im Forschungslabor des Instituts für Bibliotheks- und Informationswissenschaft stattfinden. Alternativ ist eine Durchführung in der Arbeitsumgebung der ProbandInnen oder per Videochat möglich. Es werden keine signifikanten Änderungen der Testergebnisse durch die Durchführung in einer Laborumgebung erwartet (vgl. Greifeneder 2012).

Die Befragungen werden mit Einverständnis der ProbandInnen aufgezeichnet und anschließend mithilfe eines qualitativen Analysetools kodiert und ausgewertet. Zusätzlich findet eine Protokollierung der Sitzungen statt. Um die Validität der Leitfäden sicherzustellen, findet ein Pretest für den Interviewteil und den Anwendungstest statt.

In der Analysephase der Evaluierung III wird besonders auf etwaige Abweichungen zum theoretischen Gerüst des Forschungsdesigns von AP2 geachtet (“Gaps or holes in existing theory”, Ridder 2017, S. 287 ff.). Ergebnisse werden AP2 rückgemeldet und ggf. Änderungsbedarf diskutiert.

6. AP4-5 Evaluierung IV: Visualisierung und Interfacedesign

6.1 Ziel

Die Evaluierung zielt auf die Ergebnisse des Arbeitspaketes AP3-3 und beantwortet die Frage nach der Angemessenheit der Visualisierungen und des Interfacedesigns in Bezug auf den Nutzungskontext der HNA. AP3 folgt bei der Konzipierung der Prototypen der nutzerzentrierten Designpraxis, indem zunächst Anforderungen gesammelt werden (u.a. durch AP2 und den Visualisierungsworkshop), auf deren Basis anschließend aufeinander folgende Prototypen erstellt werden. Diese werden iterativ in Reaktion auf die Rückmeldungen der NutzerInnen optimiert[8]. Die Erhebung und fundierte Analyse dieser Rückmeldungen ist Ziel der Evaluierung IV.

6.2 Vorbereitung

Die Evaluierung IV basiert auf dem Konzept der *Grounded Theory*, das auf Basis sozialwissenschaftlicher Methoden induktive Aussagen generiert (vgl. Hunger/Müller 2016, S. 259 f.). Dabei steht die begleitende, gegenstandsverankerte Herangehensweise im Zentrum.

Isenberg et al. (2008) haben dieses Konzept an die Evaluierung von Visualisierungen angepasst. Über Feldforschung wird dabei im Vorfeld der ersten Visualisierungen der Nutzungskontext

[8] Diesem iterativen Vorgehen entspricht auch die Evaluierung IV. Auf Grundlage des verschachtelten Modells nach Munzner (2009) werden vier aufeinander aufbauende Ebenen der Visualisierungskonzeption mit ihren jeweils zu evaluierenden Aspekten (*threats*, S. 922) berücksichtigt:

Ebene 1: Familarisierung mit der Domäne (AP2-1 und AP4-4);

Ebene 2: Verwendung adäquater Daten für die Zielgruppe (AP2-1 und AP2-4);

Ebene 3: Konzeption der Visualisierungen (AP3-2 und AP4-5);

Ebene 4: Erstellung der Algorithmen für den automatisierten Betrieb der Infrastruktur (AP3-4 und AP4-5).

ausdifferenziert. Dies wird in SoNAR (IDH) durch die Vorarbeit aus AP2-1, die beiden Workshops (AP3-1 und AP2-2), sowie der Evaluierung III (AP4-4) gewährleistet. Anschließend werden aufgabenbasierte Anwendungstests mit diversen ProbandInnen durchgeführt: HNA-ExpertInnen, interessierten Laien und Laiinnen sowie *Usability*-ExpertInnen. Dafür wird die Methode des *Think-Aloud-Testings* angewandt. Durch diese Methode, wird die Konstruktion von mentalen Modellen im Umgang mit dem Prototyp erfasst (vgl. Mayr et al. 2016, S. 99 f.). Dabei werden die ProbandInnen gebeten, während oder direkt nach der Erfüllung einer Aufgabe ihre Gedanken dazu laut auszusprechen (vgl. Eccles/Arsal 2017, S. 514). Gegenstand der Tests in Evaluierung IV ist jeweils aktuellste Version des Visualisierungsprototypen zum Testzeitpunkt. Mit Einverständnis der ProbandInnen werden die Tests zusätzlich zum schriftlichen Protokoll durch Audioaufnahmen und Screen-Recording dokumentiert. Den Anwendungstests liegt ein einheitlicher Leitfaden zugrunde. Zusätzlich zur aufgabenbasierten Herangehensweise wird in ausgewählten Tests exploratives Vorgehen (vgl. Dörk et al. 2011; Walsh 2015) angeregt.

6.3 Vorgehen

Im Sinne der *Grounded Evaluation* nach Isenberg et al. 2008 bilden bereits die Interviewergebnisse aus der Evaluierung III sowie das Standardisierte Forschungsdesign aus AP2-1 die Grundlage für Evaluierung IV, da sie den Nutzungskontext abstecken. In den ExpertInneninterviews wird anhand der Fallbeispiele angestrebt, Abfragen zu aktuell genutzten Software-Lösungen zur HNA-Visualisierung einzubeziehen (z.B. Gephi oder NodeXL), sowie deren Vor- und Nachteile aus Sicht der ExpertInnen erfragt.

Darüber hinaus ist bereits in den Interviews eine gezielte Abfrage der individuellen Strategien der ExpertInnen im Umgang mit Unsicherheiten in den Daten (bspw. ungenaue temporäre Angaben) geplant. Der Umgang mit vorhandenen Datenunsicherheiten in der Visualisierung (vgl. Windhager et al. 2019) ist einer der Aspekte der Evaluierung IV und damit ebenfalls Teil der Anwendungstests.

Anders als in der Evaluierung III, in der lediglich ExpertInnen befragt werden, sind für die Evaluierung IV auch Erhebungen nicht-fachwissenschaftlicher Einschätzungen sinnvoll. Daher wird zum einen angestrebt, das Sample für das Prototyp-Testing um interessierte Laien zu erweitern. Auch hier wird für die Akquise auf das Netzwerk von AP4 sowie der Projektpartner zurückgegriffen. Im Sinne der *Grounded Theory* ist die Samplebildung offen und kann z.B. nach dem Schneeballprinzip erweitert werden (vgl. Mertes 2013, S. 157). Optional können Aufrufe über Mailinglisten und Social Media erfolgen. Desweiteren werden zwei Tests mit *Usability*-ExpertInnen stattfinden. Diese Perspektive wird in der Evaluierung ergänzt, um maximal objektive Einschätzungen zur Benutzungsfreundlichkeit (*Usability*) der Infrastruktur zu erheben, losgelöst von individuellen Vorlieben (vgl. Tory/Möller 2005, S. 3).

Für die *Usability*-Tests wird auf die Methode des *Cognitive Walkthroughs* zurückgegriffen. Dabei wird eine fachspezifische Aufgabe aus dem Bereich der HNA durch eine/n *Usability*-ExpertIn am Prototyp “kognitiv durchwandert”. Ziel ist es, die intuitive Bedienbarkeit des Systems mit größtmöglicher Objektivität einzuschätzen (vgl. Blackmon 2004). Eine denkbare Aufgabe wäre hier z.B. die Wiederverwendung eines durch AP2 erstellten Use Cases.

Die Testleitfäden (aufgabenbasiert und explorativ) und das Erhebungsinstrument (Software und Hardware) werden im Vorfeld durch einen Pretest validiert und ggf. angepasst (siehe 5.3). Analog zum Vorgehen in Evaluierung III werden die Aufnahmen und Protokolle der Anwendungstests mithilfe eines qualitativen Analysetools codiert und ausgewertet. Zwischenergebnisse werden AP3 mitgeteilt und Optimierungspotenzial abgestimmt. Die Frequenz dieser Abstimmungen wird im weiteren Projektverlauf geklärt.

7. Zusammenfassung

Die erläuterten Evaluierungen I-IV können in einen quantitativen und einen qualitativen Teil ausdifferenziert werden (siehe Tabelle 1).

Ersterer umfasst die Evaluierungen I-II, ist systemzentriert und konzentriert sich damit auf die Performanz der Algorithmen, die die zugrundeliegende Datenmenge verarbeiten. Damit sind die quantitativen Evaluierungen intrinsisch, denn sie beziehen sich lediglich auf interne Faktoren, in diesem Fall den vorliegenden Datendump im Projekt SoNAR (IDH). Die gewählten Methoden beinhalten skalierte Metriken und liefern daher nach ihrer Anwendung klare Kennzahlen.

Anders ist dies im qualitativen Teil, der nutzerzentriert ist und daher auf externe Faktoren - in diesem Fall die Einschätzungen von ProbandInnen - zurückgreift. Dieser Teil umfasst die Evaluierungen III-IV. Mithilfe der gewählten Methoden aus der qualitativen Sozialforschung ist es möglich, dem innovativen Anspruch des Projektvorhabens durch eine offene und flexible Art der Datenerhebung gerecht zu werden. Die Ergebnisse sind hier lediglich nominal skalierbar und weisen keine statistische Repräsentativität auf. Um dennoch reliable und valide Aussagen treffen zu können, werden verschiedene Maßnahmen der Qualitätssicherung getroffen (Leitfäden, begründetes Sampling, Pretesting, einheitliche Codierung).

Arbeitspaket	Evaluierung	Zu evaluierendes Arbeitspaket	Methode	Erwarteter Zeitraum
Quantitativer Teil				
AP4-2	I	AP1-1	Statistische Analyse	Oktober 2019-Dezember 2019
AP4-3	II-a	AP1-2	Gold Standards Guideline-Entwicklung NER Testing Performanzmessung	November 2019-April 2020
	II-b	AP1-2	Gold Standards NEL Testing Performanzmessung	
Qualitativer Teil				
AP4-4	III	AP2-1	Case Studies Interview	ab April 2020
AP4-5	IV	AP3-2; AP3-3	Case Studies Cognitive Walkthrough Think-Aloud Usability-Testing	

Tabelle 1: Übersicht der Evaluierungsschritte im Arbeitspaket 4.

8. Referenzen

Arlot, Sylvain; Celisse, Alain (2010): A survey of cross-validation procedures for model selection. In: Statist. Surv. 4 (0), S. 40–79. DOI: 10.1214/09-SS054.

Baierer, Konstatin; Dröge, Evelyn; Petras, Vivien; Trkulja, Violeta (2014): Linked Data Mapping Cultures: An Evaluation of Metadata Usage and Distribution in a Linked Data Environment. In: DC-2014-The Austin Proceedings. Online verfügbar unter http://dcpapers.dublincore.org/pubs/article/view/3699.

Batista, David S. (2018): Named-Entity evaluation metrics based on entity-level. Blogeintrag vom 09.05.2018. Online verfügbar unter: http://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/.

Benikova, Darina; Biemann, Chris; Reznicek, Marc (2014): NoSta-D Named Entity Annotation for German: Guidelines and Dataset. LREC. Online verfügbar unter https://www.semanticscholar.org/paper/NoSta-D-Named-Entity-Annotation-for-German%3A-and-Benikova-Biemann/87dced7d3aa2a3ebe270bfeca13db5708d9537ce.

Blackmon, M. H. (2004). Cognitive Walkthrough. In: W. S. Bainbridge (Hrsg.), Encyclopedia of Human-Computer Interaction, 2 volumes (Vol. 1, pp. 104–107). Great Barrington, MA: Berkshire Publishing Group.

Bruce, Thomas R.; Hillmann, Diane I. (2004): The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In: Metadata in Practice (ALA Editions).

Dörk, Marian; Carpendale, Sheelagh; Williamson, Carey (2011): The Information Flaneur: A Fresh Look at Information Seeking. In: Desney Tan, Geraldine Fitzpatrick, Carl Gutwin, Bo Begole und Wendy A. Kellogg (Hrsg.): Conference proceedings and extended abstracts / the 29th Annual CHI Conference on Human Factors in Computing Systems. CHI 2011, Vancouver, BC, May 7 - 12, 2011. the 2011 annual conference. Vancouver, BC, Canada. S. 1215-1224.

Eccles, David W.; Arsal, Güler (2017): The think aloud method: what is it and how do I use it? In: Qualitative Research in Sport, Exercise and Health 9 (4), S. 514–531. DOI: 10.1080/2159676X.2017.1331501.

Greifeneder, Elke (2012): Does it matter where we test? Online user studies in digital libraries in natural environments. Dissertation. Humboldt-Universität zu Berlin, Berlin. Online verfügbar unter https://doi.org/10.18452/16545.

Hachey, Ben; Radford, Will; Nothman, Joel; Honnibal, Matthew; Curran, James R. (2013): Evaluating Entity Linking with Wikipedia. In: Artificial Intelligence 194, S. 130–150. DOI: 10.1016/j.artint.2012.04.005.

Hunger I., Müller J. (2016) Barney G. Glaser/Anselm L. Strauss: The Discovery of Grounded Theory. Strategies for Qualitative Research, Aldine Publishing Company: Chicago 1967, 271 S. (dt. Grounded Theory. Strategien qualitativer Forschung, Bern: Huber 1998, 270 S.). In: Salzborn S. (eds) Klassiker der Sozialwissenschaften. Springer VS, Wiesbaden.

Isenberg, Petra; Zuk, Torre; Collins, Christopher; Carpendale, Sheelagh (2008): Grounded evaluation of information visualizations. In: Proceedings of the 2008 conference on BEyond time and errors novel evaLuation methods for Information Visualization - BELIV '08. Florence, Italy: ACM Press, S. 1. Online verfügbar unter http://portal.acm.org/citation.cfm?doid=1377966.1377974.

Kann, Bettina; Hintersonnleitner, Michael (2015): Volltextsuche in historischen Texten. In: Bibliothek Forschung und Praxis 39 (1). DOI: 10.1515/bfp-2015-0004.

Labusch, Kai; Neudecker, Clemens; Zellhöfer, David (2019): BERT for Named Entity Recognition in Contemporary and Historic German. Preprint.

Lazar, Jonathan; Hochheiser, Harry; Feng, Jinjuan Heidi (2017): Research methods in human-computer interaction. Second edition. Cambridge, MA: Morgan Kaufmann. Online verfügbar unter http://proquest.tech.safaribooksonline.de/9780128093436.

Manning, Christopher (2006): Doing Named Entity Recognition? Don't optimize for F. Blogeintrag vom 25.08.2006. Online verfügbar unter: https://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html

Manning, Christopher D.; Raghavan, Prabhakar; Schutze, Hinrich (2008): Evaluation in information retrieval. In: Christopher D. Manning, Prabhakar Raghavan und Hinrich Schütze (Hrsg.): Introduction to information retrieval. New York: Cambridge University Press, S. 139–161.

Mertes, Nathalie (2013): Fallstudien. In: Konrad Umlauf, Michael S. Seadle, Petra Hauke und Simone Fühles-Ubach (Hrsg.): Handbuch Methoden der Bibliotheks- und Informationswissenschaft. Bibliotheks-, Benutzerforschung, Informationsanalyse. Berlin, Boston: DE GRUYTER SAUR. DOI: 10.1515/9783110255546.

Munzner, Tamara (2009): A Nested Model for Visualization Design and Validation. In: IEEE Transactions on Visualization and Computer Graphics 15 (6), S. 921-928. DOI: 10.1109/TVCG.2009.111.

Neudecker, Clemens (2017): ner-corpora README. Named Entity Recognition data for Europeana Newspapers. Online verfügbar unter: https://github.com/EuropeanaNewspapers/ner-corpora/blob/master/README.md

Oouchida, Kenta; Kim; Takagi, Toshihisa; Tsujii, Jun'ichi (2009): GuideLink: A Corpus Annotation System that Integrates the Management of Annotation Guidelines. In: Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation.

Ridder, Hans-Gerd (2017): The theory contribution of case study research designs. In: Business Research 10 (2), S. 281–305. DOI: 10.1007/s40685-017-0045-z.

Stiller, Juliane; Gnadt, Timo; Romanello, Matteo; Thoden, Klaus (2016): Anforderungen ermitteln, Lösungen evaluieren und Erfolge messen – Begleitforschung in DARIAH-DE. In: Bibliothek Forschung und Praxis 40 (2). DOI: 10.1515/bfp-2016-0025.

Tjong Kim Sang, Erik F.; Meulder, Fien de (2003): Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -, Bd. 4. Edmonton, Canada: Association for Computational Linguistics, S. 142–147. Online verfügbar unter http://portal.acm.org/citation.cfm?doid=1119176.1119195, zuletzt geprüft am 15.07.2019.

Tory, M.; Moller, T. (2004): Human factors in visualization research. In: IEEE Transactions on Visualization and Computer Graphics 10 (1), S. 72–84. DOI: 10.1109/TVCG.2004.1260759.

Walsh, David; Hall, Mark M. (2015): Just Looking Around: Supporting Casual Users Initial Encounters with Digital Cultural Heritage. In: Proceedings of the First International Workshop on Supporting Complex Search Tasks at ECIR 2015. Vienna, AU.

Windhager, Florian; Salisu, Saminu; Mayr, Eva (2019): Exhibiting Uncertainty: Visualizing Data Quality Indicators for Cultural Collections. In: Informatics 6 (3), 29 ff. DOI: 10.3390/informatics6030029.