Evaluierungskonzept für SoNAR (IDH)

Sina Menzel, Sandra Balck, Hannes Schnaitter & Vivien Petras

Humboldt-Universität zu Berlin

- Version III: Dezember 2020 -

Abstract

Dieses Dokument beschreibt das Konzept zur Qualitätssicherung im Projekt SoNAR (IDH) in der Version III. Dabei werden Ziele und methodisches Vorgehen der Evaluierungen im Projektverlauf beschrieben. Änderungen gegenüber der Version II wurden jeweils am Kapitelanfang festgehalten.

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

Keine Änderungen gegenüber der früheren Version.

2.1 Ziel

Ausgangsdaten

Die Evaluierung I bezieht sich auf die Ergebnisse des Arbeitsschrittes AP1-1. Sie beantwortet die Frage, ob die Transformation der für SoNAR (IDH) vorliegenden Daten in ein einheitliches Datenformat sowie deren Migration in eine projektinterne Graphdatenbank über die Software neo4j[1] ohne Daten- und Potenzialverluste erreicht wurde. Als Potenzial wird die Anzahl der vorhandenen, gültigen Referenzen innerhalb der Ausgangsdatenbestände verstanden[2]. Damit werden konkret die Ergebnisse des *Extract-Transform-Load-Prozesses* (ETL) evaluiert. Ausgangsdatenbestände sind Datendumps vom Kalliope Verbund (KPE), der Zeitschriftendatenbank (ZDB), der Deutschen Nationalbibliothek (DNB) und der Gemeinsamen Normdatei (GND) aus dem Juni 2019. Im Zuge des AP1-1 wurden in Rücksprache mit allen Projektpartnern Metadatenfelder festgelegt, die für die HNA maßgebliche Informationen enthalten (*Datenmodell*). Dieses Datenmodell ist die Grundlage für die Transformation und Potenzialanalyse ist das projektintern[1]. Das Modell ist selbst nicht Bestandteil der Evaluierung I.

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 gemäß des Datenmodells bereitstellen. Diese werden differenziert nach:

Für die Ausgangsdaten:

- M1 Anzahl der Datensätze je Datendump und insgesamt.
- M2 Anzahl der gemäß dem Datenmodell relevanten Datenfelder je Datendump. M2beinhaltet dabei keine Datensatz-Dubletten, aber zählt ungültige Datenfelder und

[1] Vgl. https://onedrive.live.com/?authkey=%21AH5z6Zyly9DgH80&cid=3B9129E4B3C7C3C9&id=3B9129E4B3C7C3C9%21403&parId=3B9129E4B3C7C3C9%21368&o=OneUp (Projektinterner Zugang, Dokumentveröffentlichung am 13.11.19)

Datenfeld-Dubletten, wie in M5 und M7 ausdifferenziert. Außerdem fasst M2Datenfelder mit ein, die implizite Entitäten aus Zeitausdrücken enthalten (Chron-Terme). Diese ergeben sich aus dem Feld 548 im GND-Datensatz.

- M3 Anzahl der gültigen Referenzen je Datenfeld und Datendump ohne Dubletten auf Ebene der Datensätze oder Datenfelder. M3 beinhaltet damit allein diejenigen Referenzen, die in die Zieldatenbank übertragen werden sollen (Token).
- M4 Anzahl der gemäß M3 referenzierten, gültigen Entitäten je Entitätentyp und Datendump (Types).
- M5 Anzahl der ungültigen Referenzen je Datendump auf Ebene eines Datenfeldes nach Fehlertyp.
- M6 Anzahl der Dubletten je Datendump auf Ebene eines Datensatzes. Konkret zählt M6 die zur DNB übertragenen ZDB-Datensätze im SoNAR-Datenpool.
- M7 Anzahl der Dubletten je Datendump auf Ebene eines Datenfeldes/Unterfeldes.

Für die Zieldaten:

- N1 Anzahl der Knoten insgesamt.
- N2 Anzahl der Knoten je Datendump.
- N3 Anzahl der Knoten je Entitätentyp.
- N4 Anzahl der Knoten je Relationentyp und Datendump.
- N5 Anzahl der Kanten insgesamt.
- N6 Anzahl der Kanten je Datendump.
- N7 Anzahl der Kanten je Relationentyp.
- N8 Anzahl der Kanten je Relationentyp und Datendump.

Sich ergebende Datenreduktionen und -bereinigungen durch AP1 werden abgestimmt und festgehalten. Dies betrifft auch bereits vorhandene Inkonsistenzen und Fehler in den Ausgangsdaten, 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. Für die vordefinierten Metadatenfelder aus dem Datenmodell muss in den Zieldaten weiterhin eine Entsprechung vorliegen. Hierüber erfolgt eine qualitative Überprüfung. Anschließend werden die Auszählungen der Ausgangs- und Zieldaten von AP1 aus der Vorbereitung miteinander abgeglichen. Dies lässt Schlüsse über mögliche Datenverluste zu (quantitative Überprüfung).

Die Evaluierung I wurde im April 2020 abgeschlossen. Der Bericht ist hier projektintern einsehbar:

https://onedrive.live.com/?authkey=%21AH5z6Zyly9DgH80&cid=3B9129E4B3C7C3C9&id=3B9129E4B3C7C3C9%21878&parId=3B9129E4B3C7C3C9%21357&o=OneUp

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

Keine Änderungen gegenüber der früheren Version.

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. Die aktuellste Version der Richtlinien kann stets eingesehen werden unter: https://github.com/qurator-spk/neat/blob/master/Annotation_Guidelines.pdf

3.2 Vorbereitung

Beachtet werden nach aktuellem Stand und gemäß der vorliegenden Normdaten: Personen, Geografika, Körperschaften, Werke, Konferenzen und sonstige Ereignisse (Events). Ausgezeichnete Entitäten, die keiner dieser Klassen zugeordnet werden können, werden als Sonstige gekennzeichnet. Von der NER betroffen sind die historischen Volltextdaten in SoNAR. 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. Das zugrundeliegende Datenformat für die Annotation der Dokumente ist dem IOB-Format aus dem GermEval Task 2014 angelehnt[2].

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[3] werden anschließend Schlüsse auf die Performanz der automatisiert erzeugten NE gezogen. Die Berechnung erfolgt gemäß der

[2] "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)

[3] 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

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 hinzugezogen, 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*.

Im Fall, dass die Auszeichnung auf zwei Ebenen erfolgt, also eingebettete und sublexische Entitäten[4] 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

Keine Änderungen gegenüber der früheren Version.

4.1 Ziel

Die Evaluierung II-b 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[5] 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. Dafür wird ebenfalls das o.g. In-House Annotationstool genutzt; dessen Anpassung an den Evaluationsschritt II-b erfolgt nach Rücksprache ebenfalls durch die SBB.

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.

[4] 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).

[5] 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.

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;*
- Falsch verlinkte korrekt ausgezeichnete Entitäten (spurious: SPU), *false positives*[6].

Die errechneten Kennzahlen geben Aufschluss über die Erfolgsrate der angewandten NEL-Algorithmen. Die Ergebnisse und Optimierungspotenziale werden mit AP1 diskutiert. Veröffentlichung von Teilergebnissen in einem Beitrag zu einem Sammelband zum Thema „Qualität der Inhaltserschließung“; wird projektintern zur Verfügung gestellt.

5. AP4-4 Evaluierung III: Modellhaftes Forschungsdesign

Keine Änderungen gegenüber der früheren Version.

5.1 Ziel

Die Evaluierung III zielt auf die Ergebnisse des Arbeitspaketes AP2-1, dem modellhaften 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

Im Vorfeld der Evaluierung III wurden Erhebungen des Visualisierungsworkshops von AP3 systematisch ausgewertet(codiert), deren Ergebnisse zur Konzipierung der Leitfäden für die Phasen der 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.

5.3 Vorgehen

Das modellhafte Forschungsdesign wird durch eine mehrphasige Interviewstudie evaluiert, die nach Möglichkeit, aber nicht zwingend, gemeinsam mit der Evaluierung der Visualisierung und des Interfacedesigns durch Anwendungstests (AP4-5) 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

[6] 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.

von teilstrukturierten Leitfadeninterviews[7] 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 individuelle Sitzungen mit den ExpertInnen, die in einen Interviewteil und einen Testteil zur Auswertung des aktuellen Visualisierungsprototyps aufgeteilt sind (vgl. 6.3). Die zeitliche Aufteilung richtet sich dabei nach den Leitfäden und der Kommunikationsform. Da durch das SARS-CoV-2 die persönliche Interaktion mit ProbandInnen zuweilen ausgeschlossen, zuweilen nur unter Auflagen zulässig ist, wird die Interviewstudie phasenweise aufgeteilt und das Studiendesign nach jedem einzelnen Analyse-Zyklen angepasst. Es werden keine signifikanten Änderungen der Testergebnisse zwischen Remote- und Laborumgebung erwartet (vgl. Greifeneder 2012).

Als ExpertInnen gelten Personen, die wissenschaftliche Arbeiten im Bereich HNA vorweisen. Die Akquise erfolgt 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 die wissenschaftliche Disziplin der ProbandInnen wird ein möglichst diverses Sample angestrebt.

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

Die sieben Interviews liegen in transkribierter Form vor. Die Analyse der Evaluierung III geht besonders auf Abweichungen zum theoretischen Gerüst des Forschungsdesigns von AP2 ein (“Gaps or holes in existing theory”, Ridder 2017, S. 287 ff.). Die Codierung und Analyse konzentriert sich dabei auf Aussagen zu konkreten Forschungsprozessen, angewandten Methoden und Tools sowie Wünschen und Anforderungen, die an Tools und Daten gestellt werden. Ergebnisse werden AP2 rückgemeldet und ggf. Änderungsbedarf diskutiert.

Für zusätzliches Feedback werden die wöchentlichen Einführungs-Meetings von Eva Holly und Mark-Jan Bludau dokumentiert und ausgewertet, um so Aussagen über Fortschritte, aufkommende Fragen und Erwartungen an SoNAR während des Einarbeitungsprozesses beobachten zu können.

6. AP4-5 Evaluierung IV: Visualisierung und Interfacedesign

Änderungen gegenüber der früheren Version:

- Ausarbeitung der Durchführung der Nutzertests

6.1 Ziel

Die Evaluierung IV 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 technischen Komponenten der nutzerzentrierten Designpraxis, indem zunächst Anforderungen gesammelt werden (u.a. durch AP2 und Co-Design-Workshops), auf deren Basis anschließend aufeinander folgende technische Komponenten erstellt werden. Diese werden iterativ in Reaktion auf die Rückmeldungen der

[7] Interviewleitfaden projektintern einsehbar im Ordner: 2019-2020 SoNAR (IDH)/03 Arbeitspakete/AP4 - Evaluierung

NutzerInnen optimiert[8]. Die Erhebung und fundierte Analyse dieser Rückmeldungen ist Ziel der Evaluierung IV.

Beantworten die entwickelten Tools die in der HNA gestellten Fragen und unterstützen sie den Forschungsprozess?

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 ausdifferenziert. Dies wird in SoNAR (IDH) durch die Vorarbeit aus AP2-1, den Co-Design-Workshop inAP3-1, sowie der Evaluierung III (AP4-4) gewährleistet. Anschließend werden aufgabenbasierte Anwendungstests mit ProbandInnen durchgeführt, diese setzen sich aus Interview- sowie möglicherweise Workshop-Teilnehmenden aus der HNA-Forschung zusammen. Dafür wird die Methode des cognitive walk-through angewandt. Dies erfolgt in Kombination mit der think aloud-Methode, bei welcher die ProbandInnen gebeten werden, während oder direkt nach der Erfüllung einer Aufgabe ihre Gedanken laut auszusprechen (vgl. Eccles/Arsal 2017, S. 514) wodurch die Konstruktion von mentalen Modellen im Umgang mit den Visualisierungskomponenten erfasst wird. (vgl. Mayr et al. 2016, S. 99 f.) Gegenstand der Tests in Evaluierung IV sind die zum Testzeitpunkt entwickelten Visualisierungskomponenten. Mit Einverständnis der ProbandInnen werden die Tests zusätzlich zum schriftlichen Protokoll durch Audioaufnahmen und Screen-Recording dokumentiert. Den Anwendungstests liegt ein einheitlicher Leitfaden zugrunde.

6.3 Vorgehen

Im Sinne der *Grounded Evaluation* nach Isenberg et al. 2008 bilden bereits die Interviewergebnisse aus der Evaluierung III sowie das modellhafte Forschungsdesign aus AP2-1 die Grundlage für Evaluierung IV, da sie den Nutzungskontext abstecken. In den ExpertInneninterviews werden Abfragen zu aktuell genutzten Software-Lösungen zur HNA-Visualisierung einbezogen (z.B. Gephi oder NodeXL) sowie deren Vor- und Nachteile aus Sicht der ExpertInnen erfragt. Darüber hinaus findet bereits in den Interviews eine gezielte Abfrage der individuellen Strategien der ExpertInnen im Umgang mit Unsicherheiten in den Daten (bspw. ungenaue temporäre Angaben) statt.

1. Basierend auf den Ergebnissen von Evaluation III werden Anforderungen an SoNAR beschrieben und definiert
2. *Entwicklung von simulated work tasks* (Beispielhafte Forschungsfragen) auf Grundlage der zuvor definierten Anforderungen. Testleitfäden und das Erhebungsinstrument (Software und Hardware) werden im Vorfeld durch einen Pretest validiert und ggf. angepasst (siehe 5.3).

[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).

3. *Evaluierung der Teilkomponenten* (ausgewählte Visualisierungen von Teilbeständen) anhand von simulated work tasks, welche "kognitiv durchwandert" werden (siehe 6.2). Ziel ist es, die intuitive Bedienbarkeit des Systems mit größtmöglicher Objektivität einzuschätzen (vgl. Blackmon 2004). Hierfür wird ein Use case mit einem Nutzungsszenario beschrieben, in dem eine bestimmte simulated work task durchgespielt wird. Getestet werden 5-6 Komponenten, diese müssen nicht direkt zusammenhängen (ausgehend von den vorliegenden Visualisierungen).
4. *Beobachtung* des Umgangs mit den Visualisierungen durch eine Kombination von Think Aloud (siehe 6.2) und konkreten Fragen zu Visualisierungskomponenten und Nutzungsszenario. Im Anschluss werden die Teilnehmenden befragt, wie sie die Interaktion erlebt haben und was ihnen an den Komponenten gefallen oder nicht gefallen hat. Die *Dokumentation* erfolgt über Aufzeichnung von Zoom + Bildschirmübertragung.
5. *Codierung und Auswertung* der Aufnahmen und Protokolle der Anwendungstests erfolgt mithilfe eines qualitativen Analysetools. Zwischenergebnisse werden AP3 mitgeteilt und Optimierungspotenzial abgestimmt. Die Frequenz dieser Abstimmungen und damit die einzelnen Zyklen und Gegenstände der Visualisierungstests werden im weiteren Projektverlauf geklärt.

Die Nutzerstudie wird über Zoom oder eine vergleichbare Videokonferenz-Lösung geführt. Die ProbandInnen erhalten ein Dokument, im dem der Ablauf des Versuchs beschrieben ist. Dieses Dokument beinhaltet eine „Storyline", in die die jeweiligen Aufgaben eingebettet sind. Die dafür zu benutzenden Visualisierungen sind im Dokument verlinkt. Die ProbandInnen teilen durchgehend ihren Bildschirm, so dass die Durchführenden die Handlungen der ProbandInnen beobachten und aufzeichnen können.

Um robuste Ergebnisse zu erhalten, werden die einzelnen zu testenden Komponenten, soweit sinnvoll, für jeden Durchlauf unterschiedlich angeordnet.

Die für die Tests benötigten Visualisierungen müssen öffentlich zugreifbar sein, damit sie keine technischen Barrieren für die Tests darstellen. Die Tests werden soweit möglich mit Live-Daten durchgeführt. Sollte dies technisch / organisatorisch noch nicht möglich sein, so werden sie auf Basis eines festgelegten Datensatzes durchgeführt. Dies wird bei der Ausgestaltung der Aufgaben berücksichtigt und den ProbandInnen mitgeteilt.

Benötigte Dokumente:

- Anschreiben ProbandInnen
- Einverständnis- und Datenschutzerklärung
- Ablaufdokument pro ProbandIn

Zeitplan:

04.01.-15.01.2021: Zusätzliche ProbandInnen suchen
11.01.-15.01.2021: Einladungen für ProbandInnen versenden
31.01.2021: Einfrieren der Visualisierungen für Nutzstudie

15.02.-26.02.2021: Nutzerstudie durchführen

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 potenzieller NutzerInnen von SoNAR - 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	Zeitpunkt
Quantitativer Teil				
AP4-2	I	AP1-1	Statistische Analyse	April 2020
AP4-3	II-a	AP1-2	Gold Standards Guideline-Entwicklung NER Testing Performanzmessung	August 2019 – März 2021 (Zwischenergebnisse werden kontinuierlich mitgeteilt)
	II-b	AP1-2	Gold Standards NEL Testing Performanzmessung	
Qualitativer Teil				
AP4-4	III	AP2-1	Case Studies Interview	ab Mai 2020 (phasenweise Durchführung)
AP4-5	IV	AP3-2; AP3-3	Case Studies Cognitive Walkthrough	ab Januar 2021

			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

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.