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

LIS-Förderprogramm oder Ausschreibung: e-Research-Technologien

Neuantrag (alt: Antragsvorgangsnummer 20161222575850616203 vom 22.12.2016)

- » Barbara Schneider-Kempf, Berlin, Generaldirektorin, Staatsbibliothek zu Berlin - Preußischer Kulturbesitz (SBB-PK)
- » Dr. Georg Rehm, Berlin, Forschungsbereich Sprachtechnologie, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
- » Univ.-Prof. Dr. med. Heiner Fangerau, Düsseldorf, Direktor, Institut für Geschichte, Theorie und Ethik der Medizin, Heinrich-Heine-Universität Düsseldorf (HHU)
- » Prof. Dr. phil. Marian Dörk, Potsdam, Forschungsprofessur Informationsvisualisierung, Institut für angewandte Forschung Urbane Zukunft, Fachhochschule Potsdam (FHP)
- » Univ.-Prof. PhD Vivien Petras, Berlin, Lehrstuhl für Information Retrieval, Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin (HU)

Hypothesen sind Netze; nur der wird fangen, der auswirft.
(Novalis 1798)

Beschreibung des Vorhabens

1 Ausgangslage und eigene Vorarbeiten

Die Beziehungen zwischen Menschen etwa in Familien, Organisationen oder Märkten bilden das Gewebe sozialer Ordnungen. Beziehungen konstituieren Möglichkeiten und Zwänge; sie beeinflussen den Zugang zu sozialem Kapital und damit Handlungs- und Wahloptionen (Lin 2001). Die Analyse dieser Beziehungen ist wesentlich für das Verstehen und Erklären von sozialen Phänomenen. Mit der Sozialen Netzwerkanalyse (SNA) entwickelte vor allem die sozialwissenschaftliche Forschung auf der Grundlage der Graphentheorie geeignete Methoden und anschlussfähige empirische Theorien zur Beschreibung und Erklärung dieser Strukturen (Jansen 2007). Die methodischen und theoretischen Ansätze der SNA etwa für die Untersuchung von sozialen Positionen oder der Fähigkeit von Akteuren, soziales Kapital für individuelle Ziele einzusetzen (Kadushin 2012), finden zunehmend auch in Verbindung mit historischen Fragestellungen Anwendung (Bauerfeld/Clemens 2014, Düring et al. 2016). Forschung in diesem Bereich ist aber mit zwei grundlegenden Herausforderungen konfrontiert: Zum einen ist Erhebung und Aufbereitung von Daten für Analysen aus den dezentral, teilweise verstreut überlieferten Archiv- und Bibliotheksbeständen aufwendig. Zum anderen ist die Nutzung der einmal erhobenen Daten für neue Forschungsfragen oder auch nur die Überprüfung der Ergebnisse quantitativer historischer Analysen bisher vor allem von persönlichen Faktoren wie der Kenntnis über Datenbezug, -format und -auswahl sowie technische Verfahren abhängig. Die „quantitative Arbeit“, so Moretti (2009), „ist [...] ohne *Kooperation* gar nicht denkbar, und zwar nicht nur, weil es endlos lange dauern würde, Datensätze allein auf sich selbst gestellt anzulegen, sondern, weil Daten im Idealfall unabhängig von ihrem Rechercheur sind.“ Im Ergebnis der Digitalisierungskampagnen unserer Kultur- und Wissenschaftseinrichtungen stehen nun erstmals signifikant große, vielfältig repräsentative Datenkorpora bereit. Durch stete Innovation und Standardisierung in der Aufbereitung digitaler Bestände – beispielhaft genannt seien Optical Character Recognition (OCR) für die Konversion von Bilddaten in maschinell prozessierbare Texte oder Named Entity Recognition (NER) für das Erkennen von Entitäten – gewinnen diese Daten auch das Interesse einer noch jungen quantitativen Perspektive auf historische Phänomene. Doch trotz der erheblichen Potenziale beruhen bisherige Angebote in erster Linie auf den Logistik- und Nutzungskonzepten für analoge Bestände: So erfolgt die Datennutzung und -generierung über Kataloge, Discovery-Systeme oder digitale Sammlungen einzelner Einrichtungen für die ebenso konventionelle Beschäftigung mit Einzelobjekten. Einrichtungsübergreifende Aggregatoren wie die Deutsche Digitale

Bibliothek (DDB) optimieren zwar den zeit- und ortsunabhängigen Zugang, aber die quantitative Verwertung der Daten bleibt hinter den Möglichkeiten zurück.

An dieser Stelle setzt das Vorhaben **SoNAR (IDH)**, Interfaces to Data for Historical Social Network Analysis and Research, an: Das **anwendungsbezogene Forschungs- und Entwicklungsprojekt** soll systematisch forschungsorientiert das Aufbereiten, Bereitstellen und Analysieren von Massendaten für den Aufbau einer Forschungstechnologie für die Historische Netzwerkanalyse (HNA), die als ein Zweig der SNA historische Fragestellungen untersucht, erprobt werden. Ausgangspunkte für das Datenmaterial sind:

» Kalliope, KPE (Archiv- und archivähnliche Bestände wie Nachlässe und Autographen),
» Zeitschriftendatenbank, ZDB (fortlaufende Sammelwerke wie Zeitungen und Zeitschriften),
» Gemeinsame Normdatei, GND (Entitäten wie Personen, Körperschaften und Orte) sowie
» exemplarische Brief- (Edition Berliner Intellektuelle) und Zeitungsvolltexte (ZEFYS).

Die Antragsteller sind überzeugt, dass diese Repositorien von überragender Bedeutung für die Historische Netzwerkanalyse (HNA) sind. In Vorbereitung dieses Antrags wurden die Repositorien formal ausgewertet (Formate, Schnittstellen, Umfang und Art der Daten (Anhang 1)) und ihr Potenzial bewertet (Anhang 2). Die Potenzialbewertung stärkt die Ausgangshypothese, dass diese Datenbestände eine kritische Menge von vielfältigen Beziehungsinformationen enthalten und dass ihre differenzierte Datenstruktur die Verwertung durch Sekundäranalysen (Kromrey 2002) nachhaltig ermöglicht; denn

» es sind *empirische* Daten über soziale Ereignisse: der Brief von A an B, die Einträge von A in die Stammbücher von B, C und D, die Herausgabe einer Zeitung im Verlag M durch C, das Anlegen einer Akte von M über B. Ein soziales Ereignis ist "das (sozial kleinstmögliche) Temporalatom", das als ein nicht weiter differenzierter Punkt in einem Raum-Zeit-Diagramm abbildbar ist (Luhmann 1987). Die Daten verbinden Ereignisse durch Kontextualisierung wie Entstehung (z.B. Autor, Ort, Zeit), Sache (z.B. Anlass, Thema) und Überlieferung (z.B. Bestandsbildner, Aufbewahrung, Eigentümer).

» es sind *regelbasierte* und *eindeutig* erfasste Daten über Entitäten, also konkrete oder abstrakte Merkmalsträger wie Akteure oder Konzepte. Erschließungsregeln, z.B. die Resource Description and Access (RDA) für ZDB und GND, sichern die Vergleichbarkeit der Daten; sie regeln die Art und Weise der Erfassung von Ausprägungen der Entitäten, z.B. mithilfe von kontrollierten Vokabularien (z.B. MARC Code List for Relators) und Normdatenreferenzierungen.

» es sind *persistent* adressierte Daten mit eindeutigen, systemunabhängigen Identnummern und URI. Die Datensätze können so systemübergreifend für die Beschreibung von Entitäten in Beziehung gesetzt werden. Datenmodifikationen werden zudem sekundengenau gespeichert, sodass die für das Projekt wichtige Versionierung unterstützt wird (Kapitel 2.3, AP1).

Das entstandene und **stetig expandierende referenzielle System dieser verteilten Datenangebote** bietet der Wissenschaft die Chance, mit statistischen und visuellen Mitteln einen breiten, tiefen Einblick in Genese und Konstellation vergangener sozialer Beziehungen zu gewinnen. Das Potenzial von Metadaten deutet KPE mit visualisierten Korrespondenznetzwerken[1], die ZDB mit visualisierten Titelbeziehungen[2], aber etwa auch die DDB mit exemplarisch visualisierten Häufigkeitsverteilungen[3] an. Einzelne wissenschaftliche Arbeiten zeigen sehr überzeugend, aber notgedrungen in reduzierter und abstrahierter Form den Wert quantitativer Methoden anhand von Korrespondenzen aus Archivbeständen wie sie in KPE erfasst sind und belegen das enorme Erkenntnispotenzial für die historische Forschung (z.B. Mücke und Schnalke 2009, Boschung et al. 2002, Dauser 2008, Fangerau 2010, 2013). Die Titeldaten der ZDB flankieren Aussagen über soziale Netze (KPE, GND) mit Aussagen über Produktions- und Distributionskonstellationen (z.B. Verlag, Herausgeber, Verbreitung, Sprache). Durch das Aufbereiten von Entitäten in Volltexten von Briefen oder Zeitungsartikeln ist es möglich, die formalisierten Aussagen der Metadaten von KPE und ZDB substanziell zu erweitern.

[1] http://kalliope.staatsbibliothek-berlin.de
[2] http://zdb-katalog.de/
[3] https://uclab.fh-potsdam.de/ddb/ (Projektarbeit unter der Leitung von Prof. Dr. Marian Dörk)

Das Vorhaben steht so auch im Kontext einer Reihe verwandter Initiativen. Beispielhaft und repräsentativ zu nennen sind: Das DFG-Projekt „Gelehrte, Ausgräber und Kunsthändler: Die Korrespondenz des Instituto di Corrispondenza Archeologica als Wissensquelle und Netzwerkindikator“[4] des Deutschen Archäologischen Instituts (DAI) erfasst seit 2016 zunächst 17.300 Korrespondenzen formal und inhaltlich in der Kalliope-Verbunddatenbank. Die Metadaten werden im Anschluss für ein einzelnes Forschungsvorhaben, das die Betrachtung sozialer Netzwerke einschließt, aufbereitet. Einen alternativen Weg geht die Historische Kommission bei der Bayerischen Akademie der Wissenschaften mit dem DFG-Projekt zur „Entwicklung eines zentralen historisch-biographischen Informationssystems für den deutschsprachigen Raum“[5]: Es verlinkt Informationsangebote mithilfe der GND-Identnummern, u.a. mit Beteiligung von Kalliope, und visualisiert Personenbeziehungen auf der Basis von Lexikonartikeln der Deutschen Biographie (ADB/NDB). Erwähnenswert ist ebenfalls das von der Europäischen Union im Rahmen des COST-Programms geförderte Projekt „Reassembling the Republic of Letters“ (ISCH COST ACTION IS1310)[6]. Es verfolgt das Ziel, die europäische wissenschaftliche Gemeinschaft des 15. bis 18. Jahrhunderts, die „Res publica litteraria“, in ihren Korrespondenzen zu identifizieren und visuell zu rekonstruieren. Gefördert wird der Fachaustausch, an dem auch die Antragsteller von SBB und FHP teilnehmen. Von besonderer Bedeutung ist aufgrund des vergleichbar generischen Ansatzes das Projekt „Social Networks and Archival Context“ (SNAC)[7]. Es wird vom National Endowment for the Humanities (NEH) und der Andrew W. Mellon Foundation seit 2010 mit dem Ziel gefördert, eine Referenzdatei für Entitäten (Körperschaften, Personen, Familien) und deren Relationen zu den a) überlieferten Quellen, b) sozialen Beziehungen sowie c) bio- und historiographischen Kontexten aufzubauen. Da eine Forschungstechnologie für die HNA auf Daten angewiesen ist, ist mit dem Projekt „SNAC“ vereinbart, Best Practice-Ansätze und Kooperationsoptionen zu erörtern (Kapitel 5.3.1).

Trotz unterschiedlicher Ausgangsbedingungen und Zielstellungen sind diesen Initiativen die Nutzung von Metadaten oder Volltexten historischer Quellen für die HNA gemein. SoNAR (IDH) soll und kann für den breiten, fächerübergreifenden Bedarf Einzellösungen durch ein standardisiertes Angebot ersetzen und so Hürden für die Arbeit mit Methoden der HNA signifikant reduzieren. Notwendig ist das Vorhaben aufgrund der Komplexität des Aufbaus einer offenen, standardbasierten e-Research-Technologie. Im Ergebnis dieses Vorhabens werden die Leistungsfähigkeit bestehender Frameworks und Werkzeuge in einer Prozesskette zur Datenaufbereitung und -bereitstellung sowie die Chancen neuer Visualisierungs- und Interfacekonzepte für eine Forschungsumgebung demonstriert. Mit einem Implementierungs- und Betriebskonzept werden geeignete Ansätze und Konditionen für Aufbau und Betrieb der Forschungstechnologie aufgezeigt. Die Ergebnisse sollen Basis für ein Folgeprojekt sein, um SoNAR (IDH) produktiv zu verstetigen. Damit knüpft dieses Vorhaben an Konzepte der Informetrie, vor allem der Biblio- und der Szientometrie an, wobei jedoch weniger Fragen nach Trends (Tunger 2009), Impact (Hirsch 2005), Wachstum oder Marktwert (Haustein und Tunger 2013, Umstätter 2004) im Vordergrund stehen, sondern z.B. Figurationen (Elias 1970) oder die räumliche und zeitliche Evolution von sozialen Beziehungen und Kontexten (z.B. Themen). Es wird dabei der Umstand berücksichtigt, dass die Ausgangsdaten nicht für nur ein Forschungsthema erhoben sind, sondern vielfältig nutzbar gemacht werden können. Daher gilt es aber auch, belastbare Aussagen über den Umgang mit fehlenden oder fehlerhaften Daten zu treffen. Die Antragsteller greifen für ihre Aufgaben im Vorhaben auf fundierte Erfahrungen zurück: Projektmanagement, Daten und Betrieb von Forschungsdiensten (SBB-PK), Frameworks und Werkzeuge zur Datenaufbereitung und -bereitstellung (DFKI), Forschungsprozess und Methoden der HNA (HHU), Visualiserungen und Interfacedesign (FHP) sowie Evaluierung von Datenqualität, Forschungsprozess und Nutzerinterfacen (HU). Gemeinsam werden Antragsteller und assoziierte Partner interdisziplinär den Aufbau und Betrieb einer innovativen Forschungstechnologie erarbeiten. Erstmals kann ein standardisiertes Instrumentarium zur Verfügung stehen, um mit großen aufbereiteten Datenmengen und einer Forschungsumgebung etwa komplexe, multimodale sozio-historische Kontexte zu untersuchen und Erkenntnisse nach wissenschaftlichen Kriterien in Forschungsprozesse zu integrieren.

[4] http://gepris.dfg.de/gepris/projekt/318512975
[5] http://gepris.dfg.de/gepris/projekt/213818920
[6] http://www.republicofletters.net/
[7] http://socialarchive.iath.virginia.edu/

1.1 Projektbezogene Publikationen

[Entfällt]

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

Baierer, Konstantin/Dröge, Evelyn/Petras, Vivien/Trkulja, Violeta: Linked Data Mapping Cultures. An Evaluation of Metadata Usage and Distribution in a Linked Data Environment. In: Proceedings of the International Conference on Dublin Core and Metadata Applications. (2014). 1-11

Chen, Ko-le/Dörk, Marian/Dade-Robertson, Martyn: Exploring the promises and potentials of visual archive interfaces. In: Proceedings of the 2014 iConference. iSchools. (2014). 736-741

Dörk, Marian/Comber, Rob/Dade-Robertson, Martyn: Monadic exploration. Seeing the whole through its parts. In: CHI '14. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2014. (2014). 1535-1544

Dörk, Marian/Knight, Dawn: Wordwanderer. A navigational approach to text visualisation. In: Corpora. 10 (2015), 1. 83-94

Dörk, Marian et al.: Pivotpaths. Strolling through faceted information spaces. In: IEEE Transactions on Visualization and Computer Graphics (TVCG). 18 (2012), 12. 2710-2719

Fangerau, Heiner/Halling, Thorsten (Hg.): Netzwerke. Eine allgemeine Theorie oder die Anwendung einer Universalmetapher in den Wissenschaften. Bielefeld, 2009

Fangerau, Heiner: Spinning the Scientific Web. Jacques Loeb (1859–1924) und sein Programm einer internationalen biomedizinischen Grundlagenforschung. Berlin, 2010

Fangerau, Heiner/Imhof, Christiane: Medizinische Spezialisierung. Wege der Urologie in beiden deutschen Staaten und die Gründung der Deutschen Gesellschaft für Urologie der DDR. In: Halling, Thorsten/Moll, Friedrich/Fangerau, Heiner (Hg.): Urologie 1945-1990. Entwicklung und Vernetzung der Medizin in beiden deutschen Staaten. Heidelberg, 2015. 21-34

Glinka, Katrin/Meier, Sebastian/Dörk, Marian: Visualising the un-seen. Towards critical approaches and strategies of inclusion in digital cultural heritage interfaces. In: Kultur und Informatik. (2015). 105-118

Krischel, Matthis/Halling, Thorsten/Fangerau, Heiner: Anerkennung in den Wissenschaften sichtbar machen. Wie die Bibliometrie durch die soziale Netzwerkanalyse neue Impulse erhält. In: Österreichische Zeitschrift für Geschichtswissenschaften. 23 (2012), 3. 179-206

Neudecker, Clemens et al.: Large scale refinement of digital historical newspapers with named entity recognition. Proceedings of the IFLA 2014 Newspaper Section Satellite Meeting, Geneva. 2014

Neudecker, Clemens: An Open Corpus for Named Entity Recognition in Historic Newspapers. Proceedings of the 10th Language Resources and Evaluation Conference, 23-28 May 2016, Portorož, Slovenia

Neudecker, Clemens/Rehm, Georg: Digitale Kuratierungstechnologien für Bibliotheken. In: 027.7 Zeitschrift für Bibliothekskultur. 2 (2016), 4. 104-116

Petras, Vivien/Stiller, Juliane: A Decade of Evaluating Europeana. Constructs, Contexts, Methods, and Criteria. In: Proceedings of the 21st International Conference on Theory and Practice of Digital Libraries (TPDL). Research and Advanced Technology for Digital Libraries. 10450 (2017). 233-245

Petras, Vivien/Stiller, Juliane/Gäde, Maria: Building for Success (?). Evaluating Digital Libraries in the Cultural Heritage Domain. In: Cool, Colleen/Ng, Kwong Bor (Hg.): Recent Developments in the Design, Construction and Evaluation of Digital Libraries. 2013. 141-163

Petras, Vivien: Methoden für die Evaluation von Informationssystemen. In: Umlauf, Konrad et al. (Hg.): Handbuch Methoden der Bibliotheks- und Informationswissenschaft: Bibliotheksforschung, Benutzerforschung, Informationsanalyse. Berlin, 2013. 368-386

Rehm, Georg et al.: Event Detection and Semantic Storytelling. Generating a Travelogue from a large Collection of Personal Letters. In: Caselli, Tommaso et al. (Hg.): Proceedings of the Events and Stories in the New Workshop. Annual Meeting of the Association for Computational Linguistics (ACL-2017). (2017). 42-51

Rehm, Georg et al.: Designing User Interfaces for Curation Technologies. In: Yamamoto, Sakae (Hg.): Human Interface and the Management of Information. Information, Knowledge and Interaction Design. 19th International Conference. HCI International 2017. 10273 (2017), 1. 388-406

Stahnke, Julian et al.: Probing projections. Interaction techniques for interpreting arrangements and errors of dimensionality reductions. In: IEEE Transactions on Visualization and Computer Graphics (TVCG). 22 (2016), 1. 629-638

Stiller, Juliane/Petras, Vivien: A Framework for Classifying and Comparing Interactions in Cultural Heritage Information Systems. In: Ruthven, Ian/Chowdhury, Gobinda (Hg.): Cultural Heritage Information. Access and Management. London, 2015. 153-176

Zellhöfer, David: Exploring Large Digital Libraries by Multimodal Criteria. In: Proceedings of the 20th International Conference on Theory and Practice of Digital Libraries (TPDL). Lecture Notes in Computer Science. 9819 (2016). 307-319

1.1.2 Andere Veröffentlichungen

[Entfällt]

1.1.3 Patente

1.1.3.1 Angemeldet

[Entfällt]

1.1.3.2 Erteilt

[Entfällt]

2 Ziele und Arbeitsprogramm

2.1 Voraussichtliche Gesamtdauer des Projekts

Die Projektlaufzeit ist für 24 Monate festgesetzt. Für diesen Zeitraum wird eine Förderung beantragt.

2.2 Ziele

SoNAR (IDH) erprobt den Aufbau einer Forschungstechnologie für die Historische Netzwerkanalyse (HNA). Das Vorhaben knüpft an die Verbundrepositorien von KPE, ZDB und GND an und schließt eine gravierende Angebotslücke für die historische Forschung. Forschende, die mit der HNA Themen und Forschungsfragen untersuchen, werden künftig auf ein standardisiertes, überregional verfügbares Angebot zugreifen können. Die Auseinandersetzung mit diversen fachspezifischen Datenformaten oder der immer wiederkehrende Aufbau spezifischer technischer Infrastrukturen wird nachrangig; die einmal in den Verbundrepositorien erfassten Daten stehen mit SoNAR (IDH) über Maschinen- und intuitive Nutzerschnittstellen zur Verfügung.

Herausforderungen, die im Vorhaben adressiert werden sollen, sind: Qualität und Aktualität aufbereiteter Daten, Transparenz der Verarbeitung und Reproduzierbarkeit der Ergebnisse, Nutz- und Erweiterbarkeit der Technologie, Visualisierung und Interaktion mit Daten, Datenerhebung und -korrektur, Implementierung und Betrieb sowie Qualitätssicherung der Ergebnisse. Die Herausforderungen sind in Teilzielen abgebildet:

Teilziel 1 (Datenaufbereitung): Die Aufbereitung der Ausgangsdaten (Meta- und Normdaten, Volltexte) für statistische und visuelle Verfahren der HNA ist in einem automatisierten Datenfluss zu erproben.

Teilziel 2 (Datenmanagement): Die aufbereiteten Daten sind gemäß den Anforderungen wissenschaftlicher Nutzung bereitzustellen (Reproduzierbarkeit, persistente Adressierbarkeit, Provenienz).

Teilziel 3 (Forschungsdesign): Es ist ein Modell einer strukturierten Datenanalyse für historische Netzwerke auf methodischer Grundlage der Sozialen Netzwerkanalyse (SNA) zu erarbeiten.

Teilziel 4 (Visualisierung und Interfacedesign): Für die strukturierte Analyse der aufbereiteten Daten sind innovative, intuitive Visualisierungs- und Interfacekonzepte (Forschungsumgebung) zu entwickeln.

Teilziel 5 (Qualitätssicherung): Die Eignung der Komponenten, Konzepte und Methoden sind qualitativ und quantitativ für eine breite wissenschaftliche Nutzung zu evaluieren und zu bewerten.

Teilziel 6 (Implementierung und Betrieb): Für Implementierung und Betrieb (inklusive Weiterentwicklung sowie Datenerhebung und -korrektur) der e-Research-Technologie ist ein Konzept zu erarbeiten.

Die mit den Teilzielen verbundenen Risiken und Herausforderungen müssen vor einer Implementierung von SoNAR (IDH) testorientiert untersucht werden. Das Risikomanagement ist Teil der Qualitätssicherung (Kapitel 2.3). Für die produktive Implementierung wird ein Folgeprojekt angestrebt.

2.3 Arbeitsprogramm und Umsetzung

Die Umsetzung der Ziele erfolgt in Arbeitspaketen unter der Federführung einzelner Projektpartner:

» **AP1 – Teilziele 1, 2 und 6 (Datenaufbereitung und -management) (Leitung: SBB-PK)**

Aufbau und Test einer Prozesskette für die maschinelle und teilautomatisierte Aufbereitung von Beziehungen zwischen Entitäten in Meta- und Normdaten, Volltexte (unidirektionaler Datenfluss), versionier- und überprüfbare Speicherung sowie Bereitstellung der aufbereiteten Daten für AP2 und AP3 sowie Erarbeitung eines Implementierungs- und Betriebskonzepts für SoNAR (IDH).

» **AP2 – Teilziel 3 (Forschungsdesign und Forschungstests) (Leitung: HHU)**

Ausarbeitung eines modellhaften Forschungsdesigns für die HNA, Formulierung funktionaler Anforderungen an AP1 und AP3, Durchführung qualitativer hypothesenbasierter Forschungstests, Auswertung der Testergebnisse der Nutzerstudie (AP4-4) zur Konsolidierung des Forschungsdesigns sowie fachwissenschaftliche Evaluierung der Chancen der e-Research-Technologie für Forschung.

» **AP3 – Teilziel 4 (Visualisierung und Interfacedesign) (Leitung: FHP)**

Iterative Ideation, Gestaltung und Erprobung innovativer Visualisierungs- und Interfacekonzepte für die explorative und deskriptive Analyse der aufbereiteten Daten (AP1), Test und Demonstration der Konzepte durch eine prototypische Anwendung sowie Auswertung der Ergebnisse der Nutzerstudie (AP4-5) für die Konsolidierung der Visualisierungen und des Interfacedesigns.

» **AP4 – Teilziel 5 (Qualitätssicherung) (Leitung: HU)**

Durchführung qualitativer und quantitativer Evaluierungen zur a) Ergebnisqualität der Aufbereitung der Daten (AP1), b) Angemessenheit des modellhaften Forschungsdesigns (AP2), c) Usability der Visualisierungen der Daten und des Interfacedesigns (AP3) sowie d) Bedarfs- und Umfeldermittlung für die e-Research-Technologie (Gesamtprojekt).

In den Arbeitspaketen werden die Teilziele eigenverantwortlich bearbeitet, aber die Gestaltung und Erprobung der e-Research-Technologie konkretisiert sich gemeinsam und übergreifend. Neuralgische Punkte zwischen den Arbeitspaketen werden iterativ mit agilen Methoden adressiert. Dies sind v.a. die

» Bereitstellung von Daten über Beziehungen, Merkmale und statistische Werte von AP1 für AP3
» Beratung über die verfügbaren oder potenziell verfügbaren Daten durch AP1 für AP2 und AP3
» Formulierung der funktionalen Anforderungen von AP2 an AP3
» Abstimmung der Anforderungen an die Datenaufbereitung zwischen AP1 und AP3
» Überführung des Forschungsdesigns in Visulisierungskonzepte durch AP3 für AP2 und AP4
» Evaluierung der Ergebnisse und Methoden von AP1, AP2 und AP3 durch AP4

Die projektweite Koordination erfolgt durch regelmäßige Telefonkonferenzen. Halbjährlich finden, auch zur Demonstration von Ergebnissen, Projekttreffen (4.1 Reisekosten) statt. Ergebnisse werden protokolliert.

Für die Aufgaben Datenaufbereitung, Forschungsdesign und Visualisierung werden generische Ansätze vom begründeten Datenpotenzial (Anlage 2) statt von einem aktuellen Datenangebot oder Thema her gewählt. Dies wird unterstützt durch die Erarbeitung eines modellhaften Forschungsdesigns, durch Forschungstests und offene fachwissenschaftliche Diskussion des Modells in der HNR-Workshopserie (AP2) sowie Befragung zu Anforderungen an die HNR (Nutzerstudie AP4-4). Die erhobenen und evaluierten Anforderungen fließen in die Forschungstechnologie ein. Zudem werden OpenSource-Komponenten eingesetzt und Optionen zur Erweiterbarkeit der Technologie und die Verfügbarkeit von Daten für Forschungsvorhaben mitbetrachtet.

Für eine qualifizierte Entscheidung über Implementierung und dauerhaften Betrieb von SoNAR (IHD) ist die Standardisierung der Prozesskette entscheidend. Dieses wird durch eine **Qualitätssicherung** überprüft, die aus drei Ebenen besteht:

Ebene 1: AP-Interne Evaluierung

Jedes AP durchläuft vier Phasen: (1) Vorauswahl von Komponenten, Definition von Anforderungen oder Entwicklung von Konzepten, (2) Tests und Evaluierung, (3) Entscheidung bzw. Auswahl der besten Lösung und (4) Konsolidierung bzw. Optimierung der besten Lösung für ein Teilziel.

Ebene 2: AP-Output Evaluierung

Neuralgische Punkte, die das Zusammenwirken der Komponenten, die wissenschaftlichen Anforderungen oder die Nutzung der Technologie betreffen, werden projektintern evaluiert. Dies sind: a) die Qualität der aufbereiteten Daten, b) die Angemessenheit des modellhaften Forschungsdesigns, c) die intuitive und erkenntnisfördernde Nutzung von Visualisierungen und Design sowie d) die Nachfrage (Bedarf und Umfeld).

Ebene 3: Bewertung der Projektergebnisse

Kapitel 2.4 dieses Antrags berücksichtigt einen Kriterienkatalog zur Bewertung der Projektergebnisse für die potenzielle Implementierung der e-Research-Technologie (s. auch Anhang 4). Das **Risikomanagement** für Projektverlauf und -ziele ist Teil der Projektorganisation (s. auch Anhang 5).

Die **Verankerung** des Projektes innerhalb der Fächerkultur ist integraler Bestandteil des Vorhabens: Dies erfolgt durch direkte Einbindung der Forschung in SoNAR (IDH) (AP2 und AP4). Die Projektpartner werden das Projekt zudem in Form von Vorträgen auf Konferenzen und Berichten in Fachzeitschriften präsentieren.

Umsetzung des Arbeitsprogramms

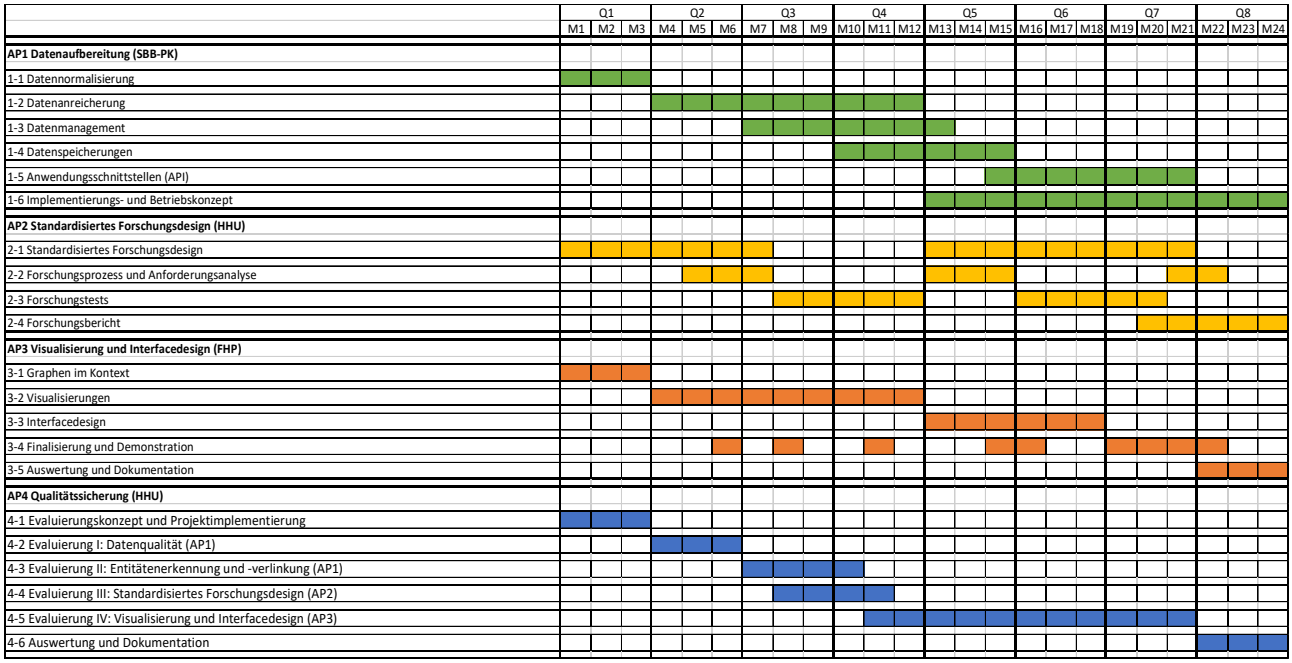

Abbildung 1: Gantt

AP1 Datenaufbereitung und -management (SBB-PK, DFKI)

Ziele

Disparate Datenquellen – Kalliope, Zeitschriftendatenbank, Gemeinsame Normdatei, Zeitungsvolltexte und Brieftranskriptionen – werden integriert, das heißt, Datenformate werden harmonisiert sowie Daten über Entitäten und ihre Relationen für die weitere Verarbeitung extrahiert. Eine Prozesskette wird erprobt, um (1) Verfahren zur Datennormalisierung sowie dem Datenmapping und -matching zu automatisieren und (2) Volltexte durch a) Texterkennung (OCR) (historische Berliner Zeitungen in ZEFYS) und b) Transkriptionen (Briefe und Texte aus dem Intellektuellen Berlin um 1800) linguistisch und semantisch (vor allem Named Entity Recognition, NER) so aufzubereiten, dass sie mit Normdaten (GND, WikiData) sowie in KPE und ZDB identifizierten Entitäten in Beziehung gesetzt werden können. Entitäten, Relationen und ihre Merkmale werden als Graphen modelliert und über eine Programmierschnittstelle (API) verfügbar gemacht.

Aufgaben und Umsetzung

AP1-1 Datennormalisierung

Ausgehend von einer detaillierten Auswertung der Ausgangssysteme KPE, ZDB und GND (Datenmodelle und -formate sowie Harvestingschnittstellen) wird für die Transformation und Bereinigung der Datenbestände ein ETL-Prozess (Extract, Transform, Load) aufgesetzt. Die in heterogenen Formaten vorliegenden Daten (EAD, MARC21, METS, etc.) werden in ein einheitliches, generisches Datenformat transformiert. Dadurch werden die Möglichkeiten zur Nutzung der Daten erweitert und die Interoperabilität der Datenbestände hergestellt. In der ersten Projektphase werden die Daten für AP3 als statischer Export bereitgestellt.

AP1-2 Datenanreicherungen

Um auch Volltexte für die Analyse nutzbar zu machen, werden mittels NER Entitäten (z.B. Personen, Orte, Organisationen) erkannt, disambiguiert und mit Normdaten (GND, WikiData) verlinkt[8]. Vorgesehen ist eine Prozesskette, in der in jedem einzelnen Verarbeitungsschritt Daten mit Annotationen versehen und diese zunehmend stärker strukturiert und semantisch aufgewertet werden, zu konzipieren und zu erproben.

[8] Für die NER kann auf Vorarbeiten des EU-Projektes Europeana Newspapers (https://github.com/EuropeanaNewspapers/) zurückgegriffen werden.

Dafür werden technische Frameworks (UIMA[9], Stanbol[10], FREME/DKT[11]) auf Funktionsumfang, Performanz und Erweiterbarkeit sowie Integrations- und Implementierungsaufwand hin getestet und verglichen. Dank eines modularen Ansatzes können einzelne Prozesse im Datenfluss flexibel adaptiert und wiederverwendet sowie weitere Funktionen, z.B. NLP-Verfahren, leicht in die Datenverarbeitungsschritte integriert werden.

AP1-3 Datenmanagement

Eine zentrale wissenschaftliche Anforderung ist die Reproduzierbarkeit der mit der Forschungstechnologie erzielbaren Ergebnisse. Die durch die vorgenannten Schritte aufbereiteten Daten zu Entitäten, Relationen und Merkmale müssen dafür versioniert sein. Durch Vergabe von persistenten Identnummern für die Daten in Kombination mit der Speicherung von Modifikationszeitpunkten, können Zustände eines bestimmten Zeitpunkts adressiert, reproduziert und so zitiert, d.h. in den Forschungsprozess integriert werden. Dafür ist zu prüfen, ob gängige Verfahren zur Versionsverwaltung (bspw. durch Protokollierung von Änderungen und Archivierung von Zuständen) für die hier untersuchten Datenbestände und deren erwartete Nutzung, ausreichende Methoden zur Verfügung stellen bzw. ob und inwieweit Anforderungen bereits durch existierende Standards (URN:NBN, DOI, DataCite) abgedeckt werden können. Zudem sind Provenienzdaten, die Auskunft über den Datenursprung (Ausgangsdaten) und technische Verfahren zur Datenanreicherung (z.B. Qualität der NER, Konfidenzwerte der Disambiguierung, Konfiguration), für die Forschung zu erhalten und transparent verfügbar zu machen. In diesem Kontext wird das Potenzial holistisch angelegter Konzepte zur Kodierung von Wissen wie Open Research Knowledge Graph (Auer 2018) mitbetrachtet.

AP1-4 Datenspeicherungen

Die in den Arbeitsschritten 1-1 und 1-2 aufbereiteten Daten werden für die wissenschaftliche Analyse (AP2) und Visualisierung (AP3) in einer Graphdatenbank gespeichert und verfügbar gemacht. Graphdatenbanken eignen sich besonders gut, um stark vernetzte und heterogene Informationen effizient zu speichern und die Datenbestände performant zu analysieren. Hierbei kann auf Methoden der Graphentheorie zurückgegriffen werden, um Merkmale wie topologische Besonderheiten oder Nachbarschaftsverhältnisse zu untersuchen. Es entsteht ein umfangreicher semantischer Kontext der Entitäten, sodass neue Angebote entwickelt oder bestehende verbessert werden können. Graphdatenbanken bieten gerade bei großen Datenbeständen und komplexen Abfragen, wie sie AP2 und AP3 vorsehen, Performanzvorteile gegenüber Triplestores.

AP1-5 Anwendungsschnittstellen (API)

Die Definition einer REST-basierten Programmierschnittstelle (API) soll Anforderungen zum dynamischen und selektiven Abruf von Daten durch Applikationen bedienen. Die API integriert den Datenabruf aus der Graphdatenbank mit Verfahren zur statistischen Auswertung und den Funktionen der Prozesskette (z.B. Reproduzierbarkeit). Die API stellt eine einheitliche Schnittstelle bereit, die die Verarbeitung und Analyse der Daten für Nutzer vereinfacht und die Anbindung externer Werkzeuge – z.B. R Statistics[12] oder Gephi[13] – ermöglicht. Der exakte Funktionsumfang der API wird in agilen Zyklen anhand der Anforderungen aus AP2 und den Visualisierungen von AP3 im Projektverlauf abgestimmt und mit Swagger[14] dokumentiert.

AP1-6 Implementierungs- und Betriebskonzept inkl. Dokumentation, Bedarfs- und Umfeldanalyse

Für die Vorbereitung einer Implementierung von SoNAR (IDH) werden die Ergebnisse – die Best Practice Technologie-Ansätze der AP1 und AP3 – evaluiert. Ziel ist es, ein Implementierungs- und Betriebskonzept mit Leistungsbeschreibung, Dokumentation sowie einer Bedarfs- und Umfeldanalyse zu erarbeiten. Dieses Konzept ist die Basis zur Bewertung der Implementierungs-, Betriebs- und Einsatzfähigkeit der e-Research-Technologie. Sukzessive fließen Ergebnisse und Schlussfolgerungen der AP1 bis AP4 in das Konzept ein. Das Konzept muss potenzielle Angebote zur Datenerhebung über die Ausgangssysteme für Forschungsprojekte,

[9] https://uima.apache.org/
[10] https://stanbol.apache.org/
[11] http://freme-project.eu, http://digitale-kuratierung.de
[12] https://www.r-project.org
[13] https://gephi.org
[14] https://swagger.io/

Personalqualifikation sowie Hardware- und Administrationskosten einschließen. Für die Bedarfs- und Umfeldanalyse arbeitet die SBB-PK eng mit der HU bei der Konzeption, Planung sowie Erhebung und Analyse der Daten im Kontext der Nutzerstudie (AP4-5) zusammen.

Ergebnisse AP1

1. Prototyp und Demonstration: Datenaufbereitung
2. Prototyp und Demonstration: Datenmanagement
3. Prototyp und Demonstration: Graphdatenbank
4. Prototyp und Demonstration: Anwendungsschnittstelle (API)
5. Konzept: Implementierung und Betrieb

AP2 Forschungsdesign und Forschungstests (HHU)

Ziele

Für die e-Research-Technologie SoNAR (IDH) ist ein modellhaftes Forschungsdesign auszuarbeiten. Es reflektiert methodische Schritte für die Analyse historischer sozialer Netzwerke (HNA) und ihrer Kontexte. Die Validität des sukzessiv zu entwickelnden Forschungsdesigns wird stets überprüft: (1) Durchführung von zwei verschiedenen Forschungstests, (2) öffentliche fachwissenschaftliche Überprüfung und Diskussion des Forschungsdesigns (HNR-Workshop) und (3) standardisierte Befragung zum Forschungsdesign (AP4-4). AP2 testet deskriptiv-statistische und netzwerkanalytische Maßzahlen (AP1). Diskursiv werden Anforderungen an a) die Visualisierung der Daten, b) die Interaktion mit Graphen sowie c) das Interfacedesign für eine optimale Integration von SoNAR (IDH) in Forschungsprozesse erarbeitet (AP3). Ein Forschungsbericht bewertet das Nutzungsspektrum der e-Research-Technologie für die fachwissenschaftliche Forschung.

Aufgaben und Umsetzung

AP2-1 Forschungsdesign

Das modellhafte Forschungsdesign beschreibt einen von Einzelthemen losgelösten Forschungsprozess der HNA in Anlehnung an die Methoden der SNA. Es orientiert sich an den Konditionen der historischen Forschung, da Daten nicht neu etwa durch Befragungen generiert werden können. Die HNA ist von den überlieferten Quellen abhängig. Zu berücksichtigen sind zudem die Potenziale der Verbundrepositorien, Daten (Merkmalsausprägungen) zu Entitäten (Merkmalsträger) in Beschreibungselementen (Merkmale) für quantitative Analysen und Visualisierungen zu erfassen (Anhang 2).

In einem ersten Schritt werden Entitäten und ihre Merkmale für die Visualisierung von Graphen betrachtet:

» Art der Beziehung: Diverse Quellenarten wie Briefe, Stamm-, Tage- und Adressbücher, Notizen und Mitschriften, Manuskripte und Aufsätze, Zeitungsartikel oder auch Findmittel, die auf archivischer Ordnungs- und Erschließungspraxis beruhen (Provenienzprinzip), enthalten explizit und implizit stärkere und schwächere Beziehungsinformationen. Die Qualität dieser Aussagen über die sozio-historischen Beziehungen muss bewertet, gewichtet und gefiltert werden können.
» Merkmale von Entitäten, z.B. Alter und Geschlecht von Personen, und Relationen, z.B. Zeitpunkt und Art, können mit deskriptiv-statistischen und netzwerkanalytischen Maßzahlen einzeln oder im Gruppenkontext beschrieben werden. Merkmalsausprägungen können zudem Filter für Graphen sein, z.B. Geschlecht, Alterskohorte, Herausgeberschaft, Beruf, Affiliationen, etc.
» Raum-Zeit-Modellierung: Für Akteure können raum-zeitliche Verläufe (z.B. Bewegungsprofile oder die Evolution der Netzwerke) etwa aus ego-zentrierter, dyadischer oder synoptischer Perspektive untersucht und verglichen werden. Die Optionen zur Periodisierung und diachronen Darstellung der Daten sind für historische Arbeiten grundlegend. Störgrößen wie Periodisierungsschnitte müssen ermittelt werden, da zu kurze oder zu lange Zeithorizonte arbiträre Ergebnisse haben können.

Die Identifikation der Analysepotenziale beruht auf den Elementen der Ausgangssysteme (Anhang 2). Das Forschungsdesign wird für quantitativ deskriptive und qualitativ interpretative Methoden der SNA für die

HNA entwickelt (Mixed-Method-Ansatz). Die Verwertung der Analyseergebnisse im Forschungsprozess, z.B. in Publikationen, muss den Kriterien von Nachvollziehbarkeit und Überprüfbarkeit genügen.

AP2-2 Forschungsprozess und Anforderungsanalyse

Iterativ und diskursiv werden mit der fortschreitenden Erarbeitung des Forschungsdesigns Anforderungen an AP1 (Datenaufbereitung) und AP3 (Visualisierung und Interfacedesign) gestellt: Ausgangspunkt ist eine Vorauswahl von Maßzahlen, die im Projektverlauf auf ihre wissenschaftliche Eignung im Kontext von SoNAR (IDH) überprüft werden. Die Datenbereitstellung über eine API (AP1) berücksichtigt deskriptiv-statistische und netzwerkanalytische Maßzahlen, die auch die Auswertung der Visualisierung (AP3) unterstützen.

Gemeinsam mit AP3 sind Konzepte für interaktive Visualisierungen zur Analyse der Daten und Integration in den Forschungsprozess zu entwickeln. Der Facettenreichtum der Daten und so der Auswertungspotenziale wird vom Einfachen zum Besonderen erprobt: Ausgehend von ego-zentrierten Korrespondenznetzwerken wird sukzessive die Vielschichtigkeit der Beziehungsinformationen durch eine intelligente Visualisierung für die diversen Datenbestände erarbeitet. Der Prozess wird unterstützt durch Forschungstests (AP2-3).

Zwischenergebnisse zum modellhaften Forschungsdesign werden nach dem ersten Forschungstest (AP2-3) mit Fachwissenschaftlern in einem Forschungsworkshop der HNR-Workshopserie (Kapitel 5.3.1) an der HHU erörtert und zugleich im Rahmen einer Nutzerstudie (AP4) untersucht. Die Ergebnisse fließen in das modellhafte Forschungsdesign (AP2-1) und darüber in die Anforderungsanalysen ein.

AP2-3 Forschungstests

Mit zwei Forschungstests wird die Eignung des modellhaften Forschungsdesigns überprüft. Sie werden nacheinander durchgeführt, damit Resultate und Erfahrungen aus dem ersten Test durch Optimierung der Datenbereitstellung (AP1) und der -visualisierung (AP3) bereits in den zweiten Test einfließen können. Es ist zu erwarten, dass die beiden Forschungstests auf vergleichbare Probleme stoßen: Ergebnisse können durch fehlende oder fehlerhafte Daten verzerrt werden und zu Fehlschlüssen (Bias) führen.

Die Forschungstests sind entscheidend dafür, das modellhafte Forschungsdesign methodenzentriert und offen für möglichst breite Forschungsfragen und -themen zu erarbeiten. Hierfür untersuchen die Tests zwei verschiedene fachwissenschaftliche Forschungsthemen mit sehr ähnlichen methodischen Ansätzen: Entlang des aktuellen (historiografischen) Forschungsstands werden für die zwei Forschungstests Hypothesen zur Wirkung von sozialen Beziehungen auf beobachtete Phänomene gebildet. Diese werden mit Ergebnissen der Forschungstechnologie SoNAR (IDH) verglichen. Die Resultate werden eingeordnet und bewertet: Als Bewertungskategorien dienen Gleichheit, Ungleichheit, Äquifinalität, Unbrauchbarkeit oder, im Idealfall, eine neue Erkenntnis. So deuten nachgewiesene, aber nicht erschlossene Nachlässe darauf hin, dass sich Netzwerke durch Datenzuwachs ändern können. Daher sind relevante Datenbestände zu identifizieren und dahingehend zu überprüfen, ob notwendige Daten in die Analyse eingeflossen sind oder aber etwa in den fehlenden Daten eine Begründung für die Differenz zwischen Forschungsstand (Literatur) und SoNAR (IDH) liegt. Es kann auch sein, dass Daten aus einer begrenzten Anzahl von Nachlässen durch Sättigungseffekte zu gleichen Ergebnissen führen, da die Daten repräsentativ im Sinne der Untersuchung sind. Ähnliches gilt für geographische Einschränkungen; so müssen fehlende Daten über transnationale Beziehungen auf ihre Bedeutung für das zu untersuchende Netzwerk geprüft werden. Erst mit den Erfahrungen aus diesen Praxistests kann der Forschungsbericht (AP2-4) erarbeitet werden.

Forschungstest 1: Das medizinisch-physiologische Forscherkollektiv des 19. und frühen 20. Jahrhunderts

Auf der Grundlage bisheriger Forschung erfolgt die Bildung von Hypothesen zu den sozialen Bedingungen der Entstehung von wissenschaftlichen Disziplinen, die von allgemeinen zu speziellen Annahmen reichen. Basis sind Personen- und Sachindizes einschlägiger Forschungsarbeiten zum Untersuchungsgegenstand. Es werden Stichwortlisten zu relevanten Wissenschaftlern und Forschungsfeldern erstellt. Die so gewonnenen Personen- und Themencluster sollen mit netzwerkanalytischen Methoden visualisiert, beschrieben und mit Ergebnissen aus dem Datenbestand von SoNAR (IDH) verglichen werden. Dies schließt Themencluster zur öffentlichen Wahrnehmung von Forschern und ihren Themen ein, die auf Inhalten von Zeitungsvolltexten

beruhen. Damit soll das „Public Engagement in Science“ in den genannten Epochen rekonstruiert sowie das Aufkommen und Verschwinden populärer Wissenschaftler mit „Massendaten“ analysiert werden.

Forschungstest 2: Die deutsche und österreichische Nationalökonomie im 19. und 20. Jahrhunderts

Beruhend auf dem aktuellen Forschungsstand werden wie im ersten Forschungstest zunächst Hypothesen über die Wirkung deutscher und österreichischer Nationalökonomen in der zweiten Hälfte des 19. und der ersten Hälfte des 20. Jahrhunderts gebildet. Das Thema eignet sich besonders zur Kontrolle der Methoden des modellhaften Forschungsdesigns. Der Untersuchungsgegenstand grenzt sich inhaltlich deutlich vom ersten Forschungstest ab, verfolgt aber ein vergleichbares historisch-hermeneutisches Erkenntnisinteresse: die Identifikation (der disziplinären Evolution) des Zusammenwirkens von Akteuren, ihre öffentliche Wahrnehmung und ihre Wirkung auf politische Entscheidungsprozesse. In den im Projekt zur Verfügung stehenden Daten sind robuste Ergebnisse zu erwarten. Aufgrund der gleichen Ansätze können Prozesse und Ergebnisqualität der Forschungstests verglichen und bewertet werden (AP2-4).

Die Forschungstests rücken zuletzt transnationale Kontexte in den Fokus: Durch Wissenschaftsmanagement und Reputationsstreben im einen und sozialpolitisches Engagement im anderen Fall gingen von deutschen und österreichischen Physiologen sowie Nationalökonomen Impulse auch nach dem Zweiten Weltkrieg aus (z.B. das Konzept „Soziale Marktwirtschaft“). Sie wurden international wahrgenommen, z.B. von den US-amerikanischen Physiologen, dem „Progressive Movement" oder etwa der „Chicago School of Economics“. Es ist zu erwarten, dass Grenzen der Daten nationaler Informationsinfrastrukturen sichtbar werden und so vom AP2 Anregungen für Kooperationen z.B. mit dem „SNAC“-Projekt (Kapitel 5.3.1) ausgehen.

AP2-4 Forschungsbericht

Der Nutzen von SoNAR (IDH) für die wissenschaftshistorische Forschung ist abschließend zu bewerten. Auf der Basis des Workshops und der Forschungstests werden Differenzen zwischen Potenzial und Ergebnissen, Art und Umfang der Funktionen für die HNA, sowie weitere Anwendungsgebiete aufgezeigt. Die Bewertung schließt ein: (1) Verifizierung der Wissensbestände, (2) Verwendung der visualisierten Daten und Analysen als heuristisches Instrument für die Generierung neuer Forschungsfragen und (3) die weitere Verarbeitung der Daten mit anderen Datenbeständen. AP4 unterstützt die Erstellung des Berichts.

Der Forschungsbericht ist zugleich die öffentlich zugängliche Dokumentation der Chancen und Grenzen von SoNAR (IDH) für Forscher, die mit der Forschungstechnologie arbeiten wollen.

Ergebnisse AP2

1. Modellhaftes Forschungsdesign
2. Anforderungskatalog für Visualisierung und Interfacedesign
3. Forschungsbericht mit Ergebnisprotokoll des HNR-Workshops

AP3 Visualisierung und Interfacedesign (FHP)

Ziele

Auf der Basis wissenschaftlicher Anforderungen an die Analyse der aufbereiteten Daten werden neuartige Visualisierungs- und Interfacedesignkonzepte entworfen und in einer prototypischen Anwendung getestet. Die Anforderungen variieren nach Fragestellung und können von einer hohen Komplexität sein. In enger Abstimmung mit AP2 werden visuelle und interaktive Repräsentationen der aufbereiteten Daten für methodisch vielfältige Untersuchungen erstellt. Die entstehenden Konzepte und Techniken werden mit AP2 und AP4 in iterativen Designzyklen entwickelt und evaluiert. Das Arbeitspaket umfasst Visualisierung und Interfacedesign als zwei Arbeitsbereiche, die in allen Phasen von Konzept, Prototyp und Evaluierung zusammen zu denken sind: Während Informationsvisualisierung die Überführung komplexer, abstrakter und umfangreicher Daten in graphische Form betrifft, widmet sich das Interfacedesign der Gestaltung einer integrierten Nutzerschnittstelle für die intuitive Interaktion mit den Daten. In einzelnen Projektphasen werden mittels Rapid Prototyping für sehr vielversprechende Visualisierungs- und Interaktionsansätze funktionale Prototypen für praxisnahe Tests entwickelt.

Aufgaben und Umsetzung

AP3-1 Graphen im Kontext

Initial werden Datenmodelle und -formate der Ausgangssysteme mit Unterstützung von AP1 betrachtet, um Aussagen über Beziehungen – Graphen – und ihre Dimensionen (zeitliche, räumliche, soziale und sachliche Kontexte) sowie Überlieferungsbezüge zu erfassen. Es werden anfangs kurze Zyklen zur Datenbereitstellung vereinbart. Die Art und Weise der Bereitstellung der Daten ändert sich mit dem Voranschreiten von AP1 und AP2; agil werden Anforderungen zwischen den Partnern abgestimmt. Für die optimale prototypische Umsetzung der Visualisierungs- und Interfacedesignkonzepte werden zudem existierende Werkzeuge und Frameworks untersucht, die bei der Erprobung von Visualisierungen Verwendung finden können.

AP3-2 Visualisierungen

Der Forschungsprozess für Entwurf und Entwicklung neuer Visualisierungskonzepte ist an Prinzipien agiler, iterativer und nutzerzentrierter Softwareentwicklung angelehnt. Das Forschungsdesign ist die Basis für die Entwicklung der Visualisierungen, deren neuartige Perspektiven auf die Daten wiederum eine fortwährende Justierung der Forschungsfragen erlauben werden. Dabei stellt die Verbindung von Anforderungsanalyse und Lösungsentwicklung die Koevolution von Forschungsdesign und Visualisierungskonzept dar.

In der frühen und mittleren Phase des Vorhabens werden kurze Zyklen für die Entwicklung und Erprobung von Visualisierungen angestrebt, um vielversprechende Konzepte schnell mit niederkomplexen Prototypen auf ihr epistemisches Potenzial prüfen zu können. Im letzten Drittel werden jene Techniken, die sich durch Evaluation (AP2, AP4) beweisen, konsolidiert und in einer experimentellen Umgebung zusammengeführt.

Die Kollaboration mit AP2 und AP4 ermöglicht eine Nutzerzentrierung, welche die Forschungssituation, ihre Anforderungen und Herausforderungen vor, während und nach der Visualisierungsentwicklung betrachtet. Am konkreten Beispiel der Forschungstests gilt es, für komplexe und mehrschichtige Akteursbeziehungen, in denen sowohl den Akteuren als auch den Beziehungen vielfältige Attribute zugeordnet sind, optimale Interaktionen im Umgang mit der Komplexität zu erproben. Die Akteursanalyse erfordert die Entwicklung von Visualisierungstechniken, welche die dynamische Sichtbarmachung, Erkundung und Filterung entlang von qualitativen und quantitativen Attributen sowie mit variierenden Granularitäten erlauben (AP 2-1).

AP3-3 Interfacedesign

In der zweiten Projekthälfte sollen zuvor isolierte Visualisierungsprototypen, die jeweils erkenntnisreiche Perspektiven auf die Datenbestände (z.B. Zeit, Raum, Semantik) anbieten, in einem modularen Interface zusammengeführt werden. Bei der interaktiven Gestaltung und prototypischen Umsetzung der Oberfläche und Interaktionstechniken liegt das Augenmerk auf der Behandlung verschiedener Analyseebenen (AP2-1) und Provenienzen (AP1-3). Neben einem Wechsel alternativer Repräsentationen soll das Potenzial visueller Gegenüberstellungen (Juxtaposition) und Überlagerungen (Superimposition) näher untersucht werden.

Es werden die folgenden Anforderungen an das Interfacedesign gestellt:

- Interaktivität (dynamische Arbeit mit Datenbestand – Visualisierung, Maßzahlen, Exploration)
- Nachvollziehbarkeit (Ausgabe der aufbereiteten Daten mit Rückbezug zu den Ausgangsdaten)
- Visuelle Konsistenz (Farbwahl für visuelle Variablen, Interaktionen für gleiche/ähnliche Funktionen)

Wesentlich für SoNAR (IDH) ist es, dass ein webbasiertes Interfacedesign entwickeln wird, das die Visualisierungs- und Analyseverfahren mit Interaktionstechniken auf intuitive Weise miteinander verknüpft.

AP3-4 Finalisierung und Demonstration

Die Entwicklung einer intuitiven webbasierten Forschungsumgebung ist eine Herausforderung. Daher ist der testbasierte Ansatz des Gesamtprojektes einerseits notwendig, andererseits aber gerade deshalb auch erfolgversprechend. Für die Visualisierung und das Interfacedesign werden umfangreiche Anforderungen gestellt, die intensiv getestet werden: in Forschungstests und einer Nutzerstudie. Die Ergebnisse sind aufzunehmen, mit AP2 und AP4 zu bewerten und in die Finalisierung der Oberfläche einzubeziehen.

Für die Demonstration der Leistungsfähigkeit von SoNAR (IDH) werden die Komponenten des Projektes in einer prototypischen Anwendung zusammengeführt. AP3 unterstützt AP1 beim Aufbau des Prototyps.

AP3-5 Auswertung und Dokumentation

Die Konzepte, Methoden und Erfahrungen werden ausgewertet und dokumentiert. Besondere Beachtung finden die Stichworte Flexibilität, Qualität und Komplexität für Visualisierung und Interfacedesign, die sich auf Leistungsfähigkeit und Funktionsspektrum von SoNAR (IDH) auswirken.

Die Auswertung findet Eingang in die Finalisierung des Implementierungs- und Betriebskonzeptes von AP1.

Ergebnisse AP3

1. Prototyp/Demonstration: Visualisierungen und Interface auf Basis der API von AP1
2. Wissenschaftliches Konzept für die Visualisierung von und Interaktion mit Graphen
3. Dokumentation der technischen Implementierung

AP4 – Qualitätssicherung (HU)

Ziele

Die Evaluierung beinhaltet die Triangulation verschiedener qualitativer und quantitativer Methoden, um die aufbereiteten Daten (AP1), die Ergebnisse der Datenanalyse (AP2) und die Datenvisualisierung (AP3) zu bewerten. Die Qualität der Metadaten, die durch eine automatische Aufbereitung erzeugt werden, sowie die extrahierten Daten aus den Volltexten (AP1) werden quantitativ mittels Gold Standards evaluiert. Die Entwicklung der interaktiven Visualisierungen (AP3) wird auf der Basis von Grounded Evaluation begleitend untersucht. Die Bedarfe der Fachwissenschaft spielen bei allen Evaluierungen eine zentrale Rolle (AP2). Sie werden durch zielgruppenorientierte Nutzerstudien untermauert. Die Arbeitsergebnisse, die kritisch für das Gesamtprojekt sind, werden sequenziell begutachtet.

Aufgaben und Umsetzung

AP4-1 Evaluierungskonzept und Projektimplementierung

Die Evaluierung erfolgt im Projektverlauf entlang der Arbeitspakete an den jeweils neuralgischen Punkten für das Gesamtprojekt: Qualität der ETL sowie Qualität der Disambiguierung und Verlinkung von Entitäten (AP1), Bewertung der Effektivität der durch die e-Research Technologie ermöglichten Analyseprozesse (AP2 und AP3) und schließlich die Evaluierung der Visualisierung und des Interfacedesigns im Rahmen einer umfassenden Nutzerstudie. Diese Vorgehensweise, die nicht die Ergebnisse (Umsetzung) der Arbeitspakete in den Blick nimmt, sondern Zwischenschritte, die der jeweils konkreten Umsetzung vorausgehen, erlaubt die Parallelität des sonst sequenziellen Ablaufs von Datenaufbereitung zu Datenanalyse zu Aus- und Bewertung. Das Evaluierungskonzept begleitet die jeweiligen Testphasen, identifiziert kritische Aspekte und stimmt Handlungsalternativen mit den Arbeitspaketen für die jeweils folgenden Projektphasen ab.

AP4-2 Evaluierung I: Datenqualität (AP1)

Hier wird die Qualität der aus Metadaten automatisiert aufbereiteten Daten bestimmt. Dies berücksichtigt die Algorithmen auf Basis impliziter Annahmen (AP2-1) für die Gewinnung von Beziehungsinformationen. Die Bewertung der Qualität der Algorithmen erfolgt mit einer gold-standard-basierten Evaluierung: erwartete und erzielte Ergebnisse werden verglichen und die Güte der Algorithmen quantitativ bestimmt. Dieses AP4-2 baut wie AP2-1 und AP3-1 auf Anhang 2 auf. Die Arbeitspakete haben jedoch verschiedene Perspektiven auf die Ausgangsdaten: fachwissenschaftliche Forschung, visuell-analytische Gestaltung und informationswissenschaftliche Qualitätssicherung. Ihre Erkenntnisse tauschen die Arbeitspakete aus.

AP4-3 Evaluierung II: Entitätenerkennung und -verlinkung (AP1)

Entitäten sind in den Metadaten der Ausgangssysteme ausgezeichnet, aber auch in Volltexten enthalten. Die Qualität der maschinellen Entitätenerkennung in Volltexten wird mithilfe eines manuell annotierten Testkorpus (Gold Standard) evaluiert. Maßzahlen wie Precision and Recall werden herangezogen, um die

Güte der NER-Algorithmen zu bestimmen. AP1 nutzt die Ergebnisse für Optimierungspotenziale sowie die Bewertung des Technikeinsatzes und so allgemein die Möglichkeit, Volltexte in der Forschungstechnologie berücksichtigen zu können. AP2 und AP3 setzen die Ergebnisse zur Justierung ihrer Prozesse ein.

AP4-4 Evaluierung III: Modellhaftes Forschungsdesign (AP2)

Evaluiert wird das modellhafte Forschungsdesign. Die Evaluierung umfasst sowohl die Ausschöpfung der Potenziale der bereitgestellten Daten für wissenschaftliche Untersuchungen als auch die Angemessenheit des Designs als ein modellhaftes Forschungsdesign. Für die erste Untersuchung werden Erkenntnisse über Potenziale der Daten in Bezug zu den im Forschungsdesign berücksichtigten Analysen (Operationalisierung der Merkmale der Merkmalsträger) verglichen. Die Angemessenheit des Forschungsdesigns wird im Kontext einer Nutzerstudie (AP4-5), die auch die Visualisierung evaluiert, untersucht. Die Evaluierung wird eng mit AP2, besonders den Forschungstests, abgestimmt und die Ergebnisse der Nutzerstudie fließen in den Forschungsbericht (AP2) ein.

AP4-5 Evaluierung IV: Visualisierung und Interfacedesign (AP3)

Die Entwicklung der Visualisierungen und des Interfacedesigns wird begleitend untersucht (Grounded Evaluation, vgl. Isenberg et al. 2008). Ziel der Evaluierung ist es, die Entwicklung eng am Forschungsprozess auszurichten und so dem Nutzungskontext zu entsprechen. Dies bedeutet auch, den Einsatz qualitativer Methoden, wie beispielsweise Beobachtungsstudien, schon früh in den Entwicklungsprozess einzubetten.

Die Validierung der Visualisierungen beruht auf dem Konzept von Tamara Munzner (2009). Die Evaluierung setzt auf ein tiefgehendes Verständnis der zugrunde liegenden Daten und des Forschungsprozesses. Die Evaluierung ist iterativ konzipiert. Dabei geht es darum, die vier aufeinander aufbauenden Ebenen einer erfolgreichen Visualisierung zu differenzieren und Gefahren für die Validität der Aussagen durch die Visualisierungen frühzeitig zu erkennen. Das modellhafte Forschungsdesign (Ebene 1) ist die Basis, auf der die Datentypen (Ebene 2) für die weitere Verarbeitung festgelegt werden. Die visuelle Kodierung dieser Daten (Ebene 3) ist der dritte Schritt gefolgt vom Design der Algorithmen (Ebene 4). Alle vier Ebenen werden in der Evaluierung ausdifferenziert und im Hinblick auf die Forschungsanforderungen untersucht.

Mit AP2 werden Beobachtungsstudien durchgeführt, um Arbeitsweise und -methoden (Forschungsprozess) sowie verwendete Daten zu verstehen. AP4 begleitet hierzu auch den öffentlichen fachwissenschaftlichen HNR-Workshop (AP2-2). Die Erkenntnisse fließen in die Visualisierungskonzepte von AP3 ein.

Das iterative Rapid Prototyping in AP3 wird evaluativ begleitet. Im Projektverlauf wird ab dem zwölften Monat eine qualitative und quantitative Nutzerstudie mit fachwissenschaftlichen und allgemein interessierten Nutzerinnen und Nutzern durchgeführt. Die diversen Gruppen sollen helfen, einen Eindruck vom Grad des intuitiven Zugangs und funktionalen Spektrums für die HNA zu gewinnen. Zudem wird der Entwicklungsprozess in AP3 durch Expertenevaluationen, wie heuristische Walk-Throughs, unterstützt.

Fachwissenschaftliche Nutzerinnen und Nutzer werden zudem zu Anforderungen an das Forschungsdesign für die HNA (AP4-4) befragt. Bestandteil der Nutzerstudie ist ebenfalls die Erhebung von Bedarf und Umfeld für die Forschungstechnologie SoNAR (IDH) (AP1).

AP4-6 Auswertung und Dokumentation

Die Resultate und Schlussfolgerungen der Evaluierungen der Arbeitspakete werden zusammengeführt und dokumentiert. Konsequenzen für die Implementierung und den Betrieb werden für AP1 aufbereitet.

Ergebnisse AP4

1. Dokumentation der Datenqualität (AP1)
2. Evaluierung des modellhaften Forschungsdesigns (AP2) sowie Bedarfs- und Umfeldanalyse (AP1)
3. Evaluierung der Visualisierungs- und Interfacedesignkonzepte (AP3)

2.4 Maßnahmen zur Erfüllung der Förderbedingungen und Umgang mit den Projektergebnissen

SoNAR (IDH) wird von der SBB-PK getragen. Es ist das Ziel, die Forschungstechnologie abhängig von den Ergebnissen des Vorhabens zu verstetigen. Bereits im Projektverlauf ist eine intensive Zusammenarbeit mit Wissenschaftlern geplant, die mit der Methode der Sozialen Netzwerkanalyse etwa historische Ereignisse untersuchen (AP2, AP4). Es ist zu erwarten, dass die HNA sich als Teil der SNA verstetigt. SoNAR (IDH) kann diesen Prozess unterstützen, indem der Zugang zu Daten erheblich simplifiziert wird und sie sowohl für bestehende Anwendungen wie Kataloge und Volltextdienste als auch für die HNA genutzt werden können.

Die Kooperationen etwa mit den in Punkt 1.1 genannten Initiativen, vor allem mit dem SNAC-Projekt, sowie die enge Einbeziehung der Forschung versprechen fachlich angemessen aufbereitete Daten und Interfaces für die digital gestützte Kollaboration im Forschungsprozess. Das Projekt ist bestrebt, Standards einzusetzen und Eigenentwicklungen zu vermeiden. Es werden nur Open Source und Standardformate genutzt. Die Veröffentlichung sowohl von Software als auch von Publikationen erfolgt Open Source bzw. Open Access.

Zudem beruhen die Ausgangsdaten auf persistenten Identnummern für eine eindeutige Adressierung der Entitäten. Die Bereitstellung aufbereiteter Daten erfolgt über die API etablierter Graphdatenbanken. Nötige Erweiterungen werden modular entwickelt, dokumentiert und zur weiteren Nutzung auf GitHub hinterlegt. Der Nachweis erfolgt in einschlägigen Datenbanken (z.B. RIsources, re3data.org).

Der Erfolg und damit die Entscheidung über die dauerhafte Implementierung der e-Research-Technologie, wird nach informationsfachlichen, wissenschaftlichen und wirtschaftlichen Kriterien bewertet (Anhang 4).

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

[Entfällt]

3 Literaturverzeichnis

Auer, Sören: Towards an Open Research Knowledge Graph. Zenodo. 2018 (http://doi.org/10.5281/zenodo.1157185)

Bauerfeld, Daniel/Clemens, Lukas: Gesellschaftliche Umbrüche und religiöse Netzwerke. In: Bauerfeld, Daniel/Clemens, Lukas (Hg.): Gesellschaftliche Umbrüche und religiöse Netzwerke. Analysen von der Antike bis zur Gegenwart. Bielefeld, 2014

Boschung, Urs et al. (Hg.): Repertorium zu Albrecht von Hallers Korrespondenz 1724-1777. Basel, 2002 (Studia Halleriana ; VII/1)

Dauser, Regina: Informationskultur und Beziehungswissen. Das Korrespondenznetz Hans Fuggers. Tübingen, 2008 (Studia Augustana. Augsburger Forschungen zur europäischen Kulturgeschichte ; 16)

Düring, Marten et al. (Hg.): Handbuch Historische Netzwerkforschung. Grundlagen und Anwendungen. Münster, 2016

Elias, Norbert: Was ist Soziologie? 1970 (Gesammelte Schriften in 19 Bänden, 5. Berlin. 2009)

Fangerau, Heiner: Spinning the Scientific Web. Jacques Loeb (1859-1924) und sein Programm einer internationalen biomedizinischen Grundlagenforschung. Berlin, 2010

Fangerau, Heiner: Evolution of knowledge from a network perspective. Recognition as a selective factor in the history of science. In: Fangerau, Heiner et al. (Hg.): Classification and Evolution in Biology, Linguistics and the History of Science. Concepts, Methods, Visualization. Stuttgart, 2013. 11-32

Haustein, Stefanie/Tunger, Dirk: Sziento- und bibliometrische Verfahren. In: Grundlagen der Praktischen Information und Dokumentation. Berlin, 2013. 479-492

Hirsch, Jorge E.: An index to quantify an individuals scientific research output. In: Proceedings of the National Academy of Science of the United States of America. 102 (2005), 46. 16569-16572

Isenberg, Petra et al.: Grounded Evaluation of information visualizations. In: ACM DL. BELIV '08 Proceedings of the 2008 Workshop on Beyond time and errors: novel evaluation methods for Information Visualization. 2008

Jansen, Dorothea/Wald, Andreas: Netzwerktheorien. In: Benz, Arthur et al. (Hg.): Handbuch Governance. Theoretische Grundlagen und empirische Anwendungsfelder. Wiesbaden, 2007. 188-199

Kadushin, Charles: Understanding Social Networks. Theories, Concepts, and Findings. Oxford, 2012

Kromrey, Helmut: Empirische Sozialforschung. Opladen, 2002

Lin, Nan: Social Capital. A Theory of Social Structure and Action. Cambridge, 2001 (2011)

Luhmann, Niklas: Soziale Systeme. Grundriß einer allgemeinen Theorie. Frankfurt am Main, 1987

Moretti, Franco: Abstrakte Kurven, Karten, Stammbäume. Abstrakte Modelle für die Literaturgeschichte. Frankfurt am Main, 2009

Munzer, Tamara: A Nested Model for Visualization Design and Validation. In: IEEE Transactions on Visualization and Computer Graphics (TVCG). 15 (2009), 6. 921-928

Mücke, Marion/Schnalke, Thomas: Briefnetz Leopoldina. Die Korrespondenz der Deutschen Akademie der Naturforscher um 1750. Berlin, 2009

Umstätter, Walter: Szientometrische Verfahren. In: Grundlagen der Information und Dokumentation. Berlin, 2004. 237-243

Tunger, Dirk: Bibliometrische Verfahren und Methoden als Beitrag zu Trendbeobachtung und Trenderkennung in den Naturwissenschaften. Jülich, 2009 (Schriften des Forschungszentrums Jülich. Reihe Bibliothek/Library, 19)

Anhang

Anhang 1 – Beschreibung der Repositorien

Die Beschreibung ist fokussiert auf Umfang, Inhalt, Qualität, Datenformate und Schnittstellen. Die SBB-PK verfügt über den Zugang zu den Repositorien des Kalliope-Verbundes, der Zeitschriftendatenbank und der Gemeinsamen Normdatei über verschiedene Schnittstellen und Formate.

Die Ausgangssysteme des Kalliope-Verbundes hostet die SBB-PK, die der ZDB und der GND die DNB. Für die Arbeit mit GND-Daten spiegelt die SBB-PK den Datenbestand der GND im MARC21-XML-Format.

Das Verfahren zur Spiegelung der GND kann auf die Daten der ZDB übertragen werden. Die Spiegel sind ausschlaggebend für Performance, Stabilität und flexible Selektion von Datensätzen und -elementen.

1-1 Kalliope-Verbund

Kalliope umfasst 4.5 Millionen Datensätze, davon 4.25 Millionen Metadaten- und 322.000 Normdatensätze von 971 Einrichtungen. Es sind im Kern Daten zu Nachlässen, Autographensammlungen und Verlagsarchive, die in Form von Findbüchern erschlossen wurden. Korrespondenzen dominieren mit über 2.8 Millionen Datensätzen den Datenbestand. 124.000 Objekte sind digitalisiert. Aktuell erfassen 57 Einrichtungen primär Daten zu ihren Beständen im Verbund mit einem direkten, redaktionell betreuten Zugang zur GND.

1-1.1 Datenformate und -schnittstellen

Offline, File Transfer Protocol (FTP):

- » Encoded Archival Description (EAD, XML - Metadaten)
- » Encoded Archival Context for Corporate Bodies, Persons, Families (EAC-CPF, XML - Normdaten)

Online, eXistDB Application Programming Interface (API)

- » Encoded Archival Description (EAD, XML - Metadaten)
- » Encoded Archival Context for Corporate Bodies, Persons, Families (EAC-CPF, XML - Normdaten)

Online, Solr Application Programming Interface (API)

- » Encoded Archival Description (EAD, JSON - Metadaten)
- » Encoded Archival Context for Corporate Bodies, Persons, Families (EAC-CPF, JSON - Normdaten)

Online, Search-/Retrieval via URL (SRU)

- » Metadata Object Description Schema (MODS, XML - Metadaten)
- » Dublin Core (DC, XML - Metadaten)

1-1.2 Datenqualität - Relationen und Referenzen

Jeder Metadatensatz des Kalliope-Verbundes verfügt über eine eindeutige, systemunabhängig persistente Identnummer sowie URI. Durch konsequente automatisierte Vergabe der ISIL je Metadatensatz, können sie überregional eindeutig einer bestandshaltenden Einrichtung zugeordnet werden. Die ISIL ist persistent. Entitäten, die mit einer Vorlage identifiziert werden können, sind über eine eindeutige GND-ID relationiert:

- » Individualisierte Personen
- » Körperschaften und Kongresse
- » Sachschlagwörter und Gattungen/Materialarten (Subset von Sachschlagwörtern)
- » Geographika (Entstehungsorte) und Werktitel

Angaben zu Datum (Zeit–Beziehung) erfolgen nach ISO 8601, zu Sprache nach ISO 639-2. Digitale Objekte, Bilddaten und Volltexte, werden i.d.R. standardisiert mit persistenten URI in den Metadaten hinterlegt. Dies gilt vorrangig für Einrichtungen, die Standardanwendungen wie Kitodo oder Visual Library einsetzen.

Zudem ist es möglich, Bestandsbildner und Provenienzgeschichte für Autographen- und Archivbestände wie Nachlässe in tabellarischer Form unter Einbeziehung der GND (Person, Körperschaft) normiert zu erfassen.

1-2 Zeitschriftendatenbank

Die Zeitschriftendatenbank umfasst 1.86 Millionen Titeldatensätze für fortlaufende Sammelwerke und über 19 Millionen Bestandsnachweise. Aktuell erfassen rund 4000 Einrichtungen primär Daten zu ihren Beständen im Verbund mit einem direkten, redaktionell betreuten Zugang zur GND.

1-2.1 Datenformate und -schnittstellen

Offline, File Transfer Protocol (FTP):

» Marc21 (Marc21, XML - Metadaten)

Online, Search-/Retrieval via URL (SRU)

» Marc21 (Marc21, XML - Metadaten)
» Marc21plus (Marc21 für Lokaldaten, XML - Metadaten)
» Dublin Core (DC, XML - Metadaten)
» PicaPlus (Pica+, XML - Metadaten)
» Resource Description Framework (RDF, XML - Metadaten)

Online, Open Archives Initiative (OAI)

» Marc21 (Marc21, XML - Metadaten)
» Dublin Core (DC, XML - Metadaten)

Online, Linked Open Data (LOD)

» Resource Description Framework (RDF, XML - Metadaten)
» Terse RDF Triple Language (Turtle, N3 - Metadaten)
» JavaScript Object Notation - Linked Data (JSON-LD - Metadaten)

Offline, Linked Open Data (LOD)

» Resource Description Framework (RDF, XML - Metadaten)
» Terse RDF Triple Language (Turtle, N3 - Metadaten)
» JavaScript Object Notation - Linked Data (JSON-LD - Metadaten)

1-2.2 Datenqualität - Relationen und Referenzen

Die Metadaten der ZDB verfügen über eine persistente Identnummer und bestandshaltende Einrichtungen werden über die ISIL eindeutig den Titeln zugeordnet. Für die Erfassung von Personen, Körperschaften und Sachschlagwörtern werden die entsprechenden GND-Datensegmente eingesetzt. Aktiv unterstützt die ZDB die Körperschaftsnormdatenredaktion. Angaben zu Datum (Zeit–Beziehung) erfolgen nach ISO 8601, zu Sprache nach ISO 639-2. Digitale Objekte, Bilddaten und Volltexte, werden i.d.R. standardisiert mit persistenten URIs in den Metadaten hinterlegt. Zudem ordnet die ZDB Metadaten den DDC-Sachgruppen zu. Die ZDB entwickelt sich zu einem Steuerungsinstrument für die Digitalisierung im Zeitungssegment.

1-3 Gemeinsame Normdatei

Seit der Zusammenlegung von PND, GKD und SWD 2012 ist eine selbstreferenzielle Datei für Entitäten entstanden. Die GND enthält 16 Millionen Datensätze (Personennamen, individualisierte Personen, Körperschaften, Geographika, Sachschlagwörter und Werke). Bibliotheksverbünde in Deutschland, Österreich, Schweiz, etc. sowie auch Editoren der Wikipedia unterstützen die Redaktionsarbeit.

1-3.1 Datenformate und -schnittstellen

Offline, File Transfer Protocol (FTP):

» Marc21 (Marc21, XML - Normdaten)

» Resource Description Framework (RDF, XML - Normdaten)

Online, Search-/Retrieval via URL (SRU)

» Marc21 (Marc21, XML - Normdaten)
» Resource Description Framework (RDF, XML - Normdaten)

Online, Open Archives Initiative (OAI)

» Marc21 (Marc21, XML - Normdaten)
» Resource Description Framework (RDF, XML - Normdaten)

Online, Linked Data (LD)

» Resource Description Framework (RDF, XML - Metadaten)

Offline, Linked Data (LD)

» Resource Description Framework (RDF, XML - Metadaten)

1-3.2 Datenqualität - Relationen und Referenzen

Die GND findet in Datenbeständen inner- und außerhalb der Kulturerbeeinrichtungen Anwendung. Es ist ein selbstreferenzielles System, das heißt, dass alle Entitäten (Personen, Körperschaften, Kongresse, Werke, Orte, Sachschlagworte) miteinander im Datenbestand verlinkt werden können. Zudem werden Segmente wie Personen DDC-Sachgruppen zugeordnet (weitere Aspekte mit Relevanz für SoNAR (IDH) s. Anhang 2).

1-4 Externe Datenangebote

Daten, die nicht über Protokolle überregionaler Verbundstrukturen zur Verfügung stehen (Zeitungsartikel der SBB-PK, ALTO XML und Briefkorpora, Dr. Anne Baillot, TEI XML), werden offline bezogen. Es werden vor allem Volltexte berücksichtigt, die über eine persistente Metadatensatz-Identnummer von KPE und ZDB eindeutig identifiziert und so Ergebnisse der Datenaufbereitung (NER, AP1) mit Entitäten der Metadaten (s. auch Anhang 2) korreliert werden können.

Anhang 2 – Potenzialanalyse (Tabelle Relationen)

Die Potenzialanalyse umfasst drei Aspekte: a) die qualitative Potenzialanalyse beschreibt Datenfelder der Repositorien zur Referenzierung von Entitäten, b) die quantitative Potenzialanalyse bildet Teilmengen der Repositorien auf der Grundlage der GND-Identnummer und c) die implizite Potenzialanalyse weist auf Erschließungstraditonen hin, die Annahmen über potenzielle soziale Beziehungen unterstützen.

a) Qualitative Potenzialanalyse (Entitäten-Relationen-Beschreibungselemente und Vokabularien)

In den folgenden Tabellen sind die relevanten Beschreibungselemente der Datenbanken KPE, ZDB und GND zum Erfassen maschinell eindeutig interpretierbarer Entitäten-Relationen aufgelistet. Nicht berücksichtigt sind Rollen- bzw. Funktionszuschreibungen etwa für die Akteursbeziehungen wie Verfasser, Adressat oder Herausgeber. Sie sind im ETL-Prozess zu normalisieren, da die Werte zwar innerhalb der Datenbanken normalisiert erfasst sind, jedoch untereinander variieren können.

Datenbank	Merkmalsträger	Merkmal	Bemerkung
Kalliope	**Nachlassquellen**	Kalliope-ID	Persistente Identnummer
		Kalliope-URI	Persistente URI
		Sprache der Vorlage	ISO 639-2
		Transaktionsdatum	ISO 8601 Timestamp
		Entstehungsdatum	ISO 8601 Datum Vorlage
		Personen	GND-Tp-Teildatenbestand
		Körperschaften	GND-Tb-Teildatenbestand
		Gattung/Materialart	GND-Ts-Teildatenbestand
		Entstehungsorte	GND-Tg-Teildatenbestand
		Sachschlagworte	GND-Ts-Teildatenbestand
		Werktitelreferenz	GND-Tu-Teildatenbestand
		Bestandshaltende Einrichtung	ISIL-Datenbestand

Datenbank	Merkmalsträger	Merkmal	Bemerkung
ZDB	**Fortlaufende Sammelwerke**	0100 ZDB-ID + URI	Persistente Identnummer und URI eines Datensatzes
		0210 Modifikation Datensatz	ISO 8601 Timestamp
		0500 Gattung	kontrolliertes Vokabular
		1100 Zeitangaben	ISO 8601
		1133 Zielgruppe	GND-Ts-Teildatenbestand
		1140 Veröffentlichungsart	Kontrolliertes Vokabular, z.B. überregional, regional, lokal
		1500 Sprache	ISO 639-2
		1700 Erscheinungsland	ISO 3166-2
		3000/3010-3018 Personen	GND-Tp/Tn-Teildatenbestand
		3100/3110 Körperschaften	GND-Tb-Teildatenbestand
		4000 Vorlagentitel	
		4045 Herstellerangabe	nicht normiert (Verlag)
		4800 Bestandshaltende Einrichtung	ISIL-Verknüpfung

Datenbank	Merkmalsträger	Merkmal	Bemerkung
GND	Tp Person Indiv.	035$a(DE-588) ID	GND-ID
		024$aURI	GND URI
		100 Personenname	Ansetzungsform
		375 Geschlecht	1 (male) / 2 (female)
		400 Verweisung	
		500 Person Relation	GND-Tp-Teildatenbestand
		510 Körperschaft Relation	GND-Tb-Teildatenbestand
		511 Konferenz Relation	GND-Tb-Teildatenbestand
		530 Werktitel Relation	GND-Tu-Teildatenbestand
		548 Datum Relation	ISO 8601
		550 Tätigkeit-Sachbegriff Relation	GND-Ts-Teildatenbestand
		551 Ort Relation	GND-Tg-Teildatenbestand

Datenbank	Merkmalsträger	Merkmal	Bemerkung
GND	Tb Körperschaft	035$a(DE-588) ID	GND-ID
		024$aURI	GND URI
		110 Körperschaft	Ansetzungsform
		410 Verweisung	
		500 Person Relation	GND-Tp-Teildatenbestand
		510 Körperschaft Relation	GND-Tb-Teildatenbestand
		511 Konferenz Relation	GND-Tb-Teildatenbestand
		530 Werktitel Relation	GND-Tu-Teildatenbestand
		548 Datum Relation	ISO 8601
		550 Sachbegriff Relation	GND-Ts-Teildatenbestand
		551 Ort Relation	GND-Tg-Teildatenbestand

Datenbank	Merkmalsträger	Merkmal	Bemerkung
GND	Tf Konferenz	035$a(DE-588) ID	GND-ID
		024$aURI	GND URI
		111 Konferenz	Ansetzungsform
		411 Verweisung	
		500 Person Relation	GND-Tp-Teildatenbestand
		510 Körperschaft Relation	GND-Tb-Teildatenbestand
		511 Konferenz Relation	GND-Tb-Teildatenbestand
		530 Werktitel Relation	GND-Tu-Teildatenbestand
		548 Datum Relation	ISO 8601
		550 Sachbegriff Relation	GND-Ts-Teildatenbestand
		551 Ort Relation	GND-Tg-Teildatenbestand

Datenbank	Merkmalsträger	Merkmal	Bemerkung
GND	Tu Werktitel	035$a(DE-588) ID	GND-ID
		024$aURI	GND URI
		130 Werktitel	Ansetzungsform
		430 Verweisung	
		500 Person Relation	GND-Tp-Teildatenbestand
		510 Körperschaft Relation	GND-Tb-Teildatenbestand
		511 Konferenz Relation	GND-Tb-Teildatenbestand
		530 Werktitel Relation	GND-Tu-Teildatenbestand
		548 Datum Relation	ISO 8601
		550 Sachbegriff Relation	GND-Ts-Teildatenbestand
		551 Ort Relation	GND-Tg-Teildatenbestand

Datenbank	Merkmalsträger	Merkmal	Bemerkung
GND	Ts Subject	035$a(DE-588) ID	GND-ID
		024$aURI	GND URI
		150 Sachschlagwort	Ansetzungsform
		450 Verweisung	
		500 Person Relation	GND-Tp-Teildatenbestand
		510 Körperschaft Relation	GND-Tb-Teildatenbestand
		511 Konferenz Relation	GND-Tb-Teildatenbestand
		530 Werktitel Relation	GND-Tu-Teildatenbestand
		548 Datum Relation	ISO 8601
		550 Sachbegriff Relation	GND-Ts-Teildatenbestand
		551 Ort Relation	GND-Tg-Teildatenbestand

Datenbank	Merkmalsträger	Merkmal	Bemerkung
GND	Tg Geographika	035$a(DE-588) ID	GND-ID
		024$aURI	GND URI
		034 Geokoordinaten	mit Zeitangaben
		151 Ortsnamen	Ansetzungsform
		451 Verweisung	
		500 Person Relation	GND-Tp-Teildatenbestand
		510 Körperschaft Relation	GND-Tb-Teildatenbestand
		511 Konferenz Relation	GND-Tb-Teildatenbestand
		530 Werktitel Relation	GND-Tu-Teildatenbestand
		548 Datum Relation	ISO 8601
		550 Sachbegriff Relation	GND-Ts-Teildatenbestand
		551 Ort Relation	GND-Tg-Teildatenbestand

b) Quantitative Potenzialanalyse (empirischer IST-Datenbestand)

Während der vorangehende Abschnitt die qualitativen Potenziale der Datenbanken beschreibt, Entitäten-Relationen auf der Basis von Quellen zu erfassen, skizziert die folgende quantitative Potenzialanalyse die bereits erfassten Entitäten-Relationen. Hierfür wurden Teilmengen anhand der GND-ID gebildet:

- » Menge der Referenzen in ZDB und KPE auf Körperschaften (Tb) und Personen (Tp) der GND
- » Menge der in ZDB und KPE referenzierten Tb- und Tp-Normdatensätze der GND
- » Durchschnitt der sowohl in ZDB als auch KPE referenzierten Tb- und Tp-Normdatensätze der GND
- » Menge der Tp-Normdatensätze in der GND mit Referenz auf weitere Tp-Normdatensätze der GND
- » Menge der Tb-Normdatensätze in der GND mit Referenz auf weitere Tb-Normdatensätze der GND

Die Menge der Titeldaten für ZDB und KPE in der Spalte „Menge der Referenzen zur GND (Tp+Tb)“ und die Ergänzung der relevanten Grundmenge der Tp- und Tb-Datensätze der GND (Zeile 4) sind Vergleichswerte:

	Menge der Referenzen zur GND (Tp+Tb)	**Menge der referenzierten Tp-/Tb-GND-Datensätze**	**Menge Tp-Tp-Referenzen in der GND (Kategorie 500)**	**Menge Tb-Tb-Referenzen in der GND (Kategorie 510)**
ZDB	1.860.000 Titeldaten enthalten: 1.342.064 Referenzen zur GND	372.685 (= 363.834 Tb + 8.851 Tp)	8.851 Tp enthalten 1534 Referenzen zu weiteren Tp-Datensätzen	363.834 Tb enthalten 181.222 Referenzen zu weiteren Tb-Datensätzen
KPE	4.211.000 Titeldaten enthalten: 6.527.345 Referenzen zur GND	320.578 (=45.713 Tb + 274.874 Tp)	274.874 Tp enthalten 29.533 Referenzen zu weiteren Tp-Datensätzen	45.713 Tb enthalten 14.889 Referenzen zu weiteren Tb-Datensätzen
ZDB/KPE	n.a.	25.675 (= 20.695 Tb + 4.469 Tp)	4.469 Tp enthalten 1.291 Referenzen zu weiteren Tp-Datensätzen	20.695 Tb enthalten 10.577 Referenzen zu weiteren Tb-Datensätzen
GND	*n.a.*	*5.794.737 (= 1.493.359 Tb + 4.301.378 Tp)*	*4.301.378 Tp enthalten 278.419 Referenzen zu weiteren Tp-Datensätzen*	*1.493.359 Tb enthalten 505.685 Referenzen zu weiteren Tb-Datensätzen*

Die ermittelten Daten unterstützen die Erwartungen: Von den Datenbanken ZDB und KPE geht eine hohe Anzahl von Referenzen auf Körperschafts- und Personennormdatensätzen (Tb) der GND aus, die aktuell auf 372.685 (ZDB)/320.578 (KPE) eindeutige Körperschaften (Tb) und Personen (Tp) verweisen. Die Menge der sowohl in der ZDB als auch in KPE referenzierten GND-ID überrascht ebenfalls nicht: In KPE dominieren eher Personen (z.B. Nachlässe, Autographen) und in der ZDB eher Körperschaften (z.B. Verlage, Herausgeber). Überraschend ist dagegen zunächst die relativ geringe Menge der Referenzen zwischen Personen-Personen (Tp-Tp) und Körperschaften-Körperschaften (Tb-Tb) der GND (Tp-Tb und Tb-Tp wurden nicht ausgewertet). Diese Beobachtung kann v.a. auf drei Faktoren zurückgeführt werden:

Der erste Faktor ist das Speicherverfahren: Wird eine Referenz im Datensatz A gespeichert, wird diese nicht maschinell reziprok im referenzierten Datensatz ergänzt. SoNAR (IDH) wird die reziproken Beziehungen im Graphen je Entität sichtbar machen. Der zweite Faktor ist, dass die GND bisher v.a. durch bibliothekarische Katalogisierung Daten gewinnt, die oft nur rudimentär individualisierende Daten erfasst (v.a. Geschlecht, Beruf, Lebensdaten). Die GND gewinnt jedoch durch die zunehmende Kooperation etwa mit Archiven und Forschung, da z.B. durch archivische Erschließung oder Editionsvorhaben erheblich besser kontextualisierte Daten z.B. über familiäre und freundschaftliche Beziehungen oder Affiliationen zu Körperschaften erhoben werden. Ein dritter Faktor ist, dass die Möglichkeit, Relationen in der Normdatei zu erfassen, erst seit der Zusammenführung von PND, GKD und SWD zur GND 2012 besteht. Es kann erwartet werden, dass die GND-internen Referenzen in den kommenden Jahren noch erheblich zunehmen.

c) Implizite Potenzialanalyse (Datenkontexte)

KPE unterstützt mit der Verbunddatenbank das Verzeichnen von Archiv- und archivähnlichen Beständen wie Nachlässe, Autographensammlungen und Verlagsarchive. Die findbuchorientierte Erfassung der Daten

erfolgt nach dem Provenienzprinzip. Diese Ordnungs- und Erschließungstradition führt zu der begründeten Annahme, dass neben den explizit erfassten Beziehungen wie Korrespondenzpartner eines Briefes in einem Datensatz eines Findbuchs, implizite Beziehungen etwa zwischen Akteuren über die einzelnen Datensätze hinaus allein durch den Bestandskontext angenommen werden können; denn alle in einem Findbuch eines Archivbestandes erfassten Akteure können über den Bestandsbildner direkt oder indirekt miteinander in einer Beziehung gestanden haben. Diese Annahme über implizite Beziehungen zwischen Akteuren aufgrund des Provenienzprinzips ist z.B. ein Teil der Algorithmen des SNAC-Projekts zum Extrahieren, Identifizieren und Verlinken von Akteuren, die in Findbüchern erwähnt sind.[15] Diese Potenziale (Wahrscheinlichkeiten) sollen im Kontext des Projektvorhabens SoNAR (IDH) auf ihre Brauchbarkeit hin überprüft werden.

[15] Pitti, Daniel: Leveraging VIAF in Social Networks and Archival Context. Authority Data on the Web. IFLA, 2016 (hier besonders: Prinzipien archivischer Beschreibung (Records in Context) – „Though interrelations have been more implicitly than explicitly documented (in finding aids, GM), among other objectives, SNAC aspires to make the interrelations explicit." https://www.oclc.org/content/dam/oclc/events/2016/IFLA2016/presentations/Leveraging-VIAF-in-the-Social-Networks-and-Archival-Content.pdf