SoNAR /
Social Network Analysis and related Research

Implementierungs- und Betriebskonzept

Stand: 7. Oktober 2021

gefördert von der Deutschen Forschungsgemeinschaft, Programm e-Research-Technologien, https://gepris.dfg.de/gepris/projekt/414792379, 2019-2021

Vorwort

Das Implementierungs- und Betriebskonzept ist ein Ergebnis des Projekts “Interfaces to Data for Historical Social Network Analysis and Research, SoNAR (IDH)”. Es wurde von der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz zur Vorbereitung des Aufbaus einer Forschungstechnologie SoNAR für die Historische Netzwerkforschung (HNR) erarbeitet.

Das Konzept bewertet und integriert die Ergebnisse der Projektpartner. Hierzu zählen:

- *Fachanforderungen* der Heinrich-Heine-Universität Düsseldorf, Institut für Geschichte, Theorie und Ethik der Medizin (UDK), Prof. Dr. Heiner Fangerau,
- *Visualisierungskonzepte* der Fachhochschule Potsdam, Institut für Angewandte Forschung Urbane Zukunft (FHP), Prof. Dr. Marian Dörk und
- *Nutzerstudien* der Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft (HU), Prof. Vivien Petras, Ph.D.

Das Deutsche Forschungszentrum für Künstliche Intelligenz, Abteilung Sprachtechnologie (DFKI), Prof. Dr. Georg Rehm, entwickelte modellhaft eine *Datenprozesskette* für praxisorientierte Tests und Evaluierungen (Forschungs-, Visualisierungs- und Nutzerstudien). Die *Anforderungsanalyse* führte die Fa. effective WEBWORK GmbH durch. Sie unterstützte auch die *Aufwandsabschätzung* für die Implementierung und den Betrieb der Forschungstechnologie SoNAR.

Die Projektergebnisse sind mit Ausnahme der Aufwandsabschätzung auf der Plattform GitHub öffentlich dokumentiert: https://github.com/sonar-idh

Berlin, Oktober 2021

Executive Summary

Das Implementierungs- und Betriebskonzept für den Aufbau einer Forschungstechnologie für die Historische Netzwerkforschung (HNR) schließt an die Ergebnisse des DFG-Projekts SoNAR (IDH) an. Dessen Kernannahme ist es, dass Daten von Bibliotheken und Archiven, speziell Norm- und Metadaten, Aussagen über soziale Beziehungen enthalten. Durch das Projekt konnte nun auf der Grundlage von Semantic Web Technologien zur Integration heterogener Datenquellen und zur Aufbereitung der Daten als Netzwerkdaten eine Datenprozesskette einschließlich Zugangswege erarbeitet werden. Begleitet wurde die Erprobung durch eine Umfeldanalyse sowie Studien über ein modellhaftes Forschungsdesign. Dadurch wurde deutlich, dass ein konkreter Bedarf in Daten für visuelle und quantitative historische Netzwerkanalysen besteht. Die Methoden und Theorien der HNR gelten als wirkungsvoll und die HNR kann bereits auf vielfältige Analyseanwendungen zurückgreifen. Die Datenerhebung ist dagegen aufwändig und nicht selten prohibitiv.

Die Studien zum modellhaften Forschungsdesign belegen, dass Daten von Kultureinrichtungen für die HNR gewinnbringend sind. Analysen zu Datenkategorien zeigen ein hohes Potenzial, um differenziert Akteure und soziale Beziehungen zu beschreiben. Einschränkungen zeigen sich in der Praxis, da Daten oftmals nur zur Identifikation von Entitäten erfasst werden. Dennoch ist die Datenmenge der Normdateien und Verbunddatenbanken, die sich auch neuen Nutzergruppen und so neuen Anforderungen öffnen, sowie das Potenzial, weitere Datenquellen integrieren zu können, enorm. Das Erprobungsprojekt fand auch international Interesse, sodass Konzepte u.a. mit dem „Archives nationales“ und der Kooperative „Social Network and Archival Context“ erörtert und Kooperationen für eine Implementierung abgestimmt wurden.

SoNAR greift wissenschaftliche Anforderungen an Daten und den Zugang zu Daten auf: Neben der Aufbereitung von Netzwerkdaten zählt hierzu die Sicherung der Transparenz der Herkunft, Verarbeitung und Überprüfbarkeit der Datenverarbeitung. Der Datenzugang wird über moderne Schnittstellen für nachnutzende Systeme zur Datenanalyse und ein User Interface (UI) möglich. Letzteres berücksichtigt klassische Retrieval-Methoden und explorative Instrumente, zu denen die Graph-Visualisierung von Teildatenmengen zählt. Die Konzepte für die Visualisierung von und die Interaktion mit den Netzwerkdaten dienen der Prüfung und Selektion verfügbarer Daten für Datenanalysen. Darüber hinaus können sie frühe Phasen von Forschungsprozessen unterstützen, speziell die Entwicklung von Forschungsfragen und Formulierung von Hypothesen. Grundsätzlich wird SoNAR in der Lage sein, aufbereitete Netzwerkdaten auf die Ursprungsdaten und so auf die bezeugenden Quellen zurückzuführen, sodass eine Datenprüfung stets gewährleistet ist.

SoNAR kann eine international ausgerichtete Infrastruktur für die HNR werden und substanziell durch die Datenbereitstellung für wissenschaftliche Analysen zum Erkenntnisgewinn beitragen.

Inhalt

1. Einführung

Der Zugang zu Daten über historische Netzwerke ist steinig. Historische Quellen[1], die Netzwerke und Akteure dokumentieren, sind häufig unikal und an unterschiedlichen Orten überliefert. Die Datenerhebung anhand ermittelter Quellen ist zeitintensiv. Chancen, bereits erhobene Daten für Sekundäranalysen nutzen zu können, sind aufgrund mangelnder Standards bei Dokumentation, Format und Zugang gering. Obwohl die Historische Netzwerkforschung (HNR) vielversprechende methodische Ansätze und theoretische Perspektiven zur Analyse vergangener Ereignisse bietet, bleibt sie wegen dieses begrenzten Zugangs zu quantifizierbaren Daten hinter den Möglichkeiten zurück. An diesem Punkt schließt das Konzept zu einer Forschungstechnologie für die HNR nun an. Sie wird im Folgenden als SoNAR – ***So****cial* ***N****etwork* ***A****nalysis and related* ***R****esearch* – bezeichnet. Das Konzept operationalisiert den Umstand, dass Bibliotheken, Archive und Museen umfassend Daten in Verbindung mit aufbewahrten Quellen erheben bzw. erzeugen: für den Nachweis in Katalogen und Findbüchern, durch die Digitalisierung der Quellen, in kooperativen Projekten mit Forschung und Gesellschaft, uvm. Diese Daten, die oft verteilt in heterogenen Datenrepositorien gespeichert sind, enthalten explizit und implizit Aussagen über Akteure und Beziehungen. SoNAR wird die wissenschaftliche Arbeit der HNR und verwandter Forschungen unterstützen können, indem durch standardisierte Prozesse und Einsatz offener Standards heterogene Daten diverser Datenrepositorien integriert und als Netzwerkdaten[2] aufbereitet werden. Erstmals wird die HNR auf einen umfassenden themen-, zeit- und ortsübergreifender Datenbestand mit Semantic Web Technologien zugreifen können. Das SoNAR-Konzept ist an einem langfristigen Betrieb orientiert, das heißt, dass regelmäßige Datenaktualisierungen und -erweiterungen wesentlicher Bestandteil der Serviceleistungen sein werden. SoNAR wird zudem als international-orientierte Infrastruktur entwickelt, um einen möglichst breiten Zugang zu Daten für Forschungsvorhaben organisieren zu können. Das Besondere ist jedoch nicht nur die Integration umfangreicher Datenbestände und die Aufbereitung zu einem sozialen Netzwerkgraphen, sondern die Sicherung wissenschaftlicher Anforderungen an die Daten: die Transparenz der Herkunft und Verarbeitungsschritte sowie der Zugang zu den Quellen, die die Akteure und ihre Beziehungen bezeugen. Dadurch ist SoNAR nicht nur ein Angebot für die HNR, sondern zugleich ein alternativer, ein akteurszentrierter Einstieg in konventionelle Bibliothekskataloge und Archivfindmittel.

Das vorliegende Konzept für die Implementierung und den Betrieb einer Forschungstechnologie SoNAR ist das Ergebnis des Erprobungsprojekts SoNAR (IDH)[3]. Es umfasst drei Teile: (1) Analyse von Bedarf und Umfeld, (2) Beschreibung und Abgrenzung des SoNAR-Systems und (3) Aufgaben und Aufwände für Implementierung und Betrieb. Der erste Teil (Kapitel 2) betrachtet Anwender, Anforderungen und Datenquellen. Schlussfolgerungen zu Bedarf und Umfeld beruhen auf der Auswertung von Projektberichten, Publikationen und etablierten Softwarelösungen zur Analyse und Visualisierung von Netzwerkdaten (Anhang 1). In die Betrachtung sind auch Aussagen von Interviewpartnern zu Forschungskontexten im Rahmen einer Studie der Humboldt-Universität (Balck/Menzel/Petras 2021) zu einem modellhaften Forschungsdesign der HNR (Fangerau et al. 2021) eingeflossen. Beide Studien sowie Analysen zu Visualisierungs- und Interaktionskonzepten

[1] Quelle wird Synonym zum RDA-Begriff Ressource verwendet (s. Regelwerk Resource Description and Access, RDA)
[2] Als Netzwerkdaten werden in diesem Konzept sowohl die sozialen Beziehungen zwischen Akteuren, z.B. familiäre Beziehung oder Korrespondenzbeziehung, als auch die Daten über Akteure, z.B. Alter oder Beruf, definiert.
[3] https://gepris.dfg.de/gepris/projekt/414792379

(Bludau/Dörk 2021) und die Evaluierung ihrer prototypischen Implementierung[4] (Schnaitter et al. 2021) waren Grundlage zur bedarfsorientierten Ermittlung von Anforderungen. Während die Studien den gesamten Forschungsprozess ausgehend vom modellhaften Forschungsdesign in den Blick nahmen, wurden für das Konzept die Anforderungen aufgegriffen, die die Aufbereitung und den Zugang zu den Daten für wissenschaftliche Analysen betreffen. Hintergrund ist der durch die Umfeldanalyse ermittelte Bedarf nach Daten sowie die Abgrenzung des SoNAR-Systems zu Anwendungen für Datenanalysen und -visualisierungen. Im Unterkapitel Datenquellen werden diese systematisiert und Vorbedingungen für deren Integration anhand der Erfahrungen mit der exemplarischen Datenprozesskette spezifiziert[5]. Auch wurden Datenkategorien von Ontologien analysiert, um einen Eindruck über das Spektrum von Werten aus Norm- und Metadaten zu gewinnen (Anhang 3). Neben der Aufbereitung heterogener Datenquellen als Netzwerkdaten wird im Ergebnis der Analysen SoNAR wissenschaftliche Kernanforderungen adressieren:

- Bereitstellung von *Provenienzdaten* zur Herkunft und Verarbeitung der Input-Daten, aus denen Aussagen über Netzwerke und Akteure gewonnen werden,
- Unterstützung der *Reproduktion* von Forschungsprozessen durch einen langfristigen Zugang zu den In- und Output-Daten sowie den Transformationsmodellen,
- Zugang zu einer webbasierten *Nutzerschnittstelle*, um die Verfügbarkeit von Daten für ein Forschungsthema prüfen sowie Fragestellungen entwickeln zu können, und
- Zugang zu einer *Programmierschnittstelle (API)*, um Daten in Forschungsumgebungen für wissenschaftliche Datenanalysen übernehmen zu können.

Der zweite Konzeptteil (Kapitel 3) enthält die Beschreibung des SoNAR-Systems. Sie beruht auf der Analyse der wissenschaftlichen Anforderungen innerhalb des zuvor identifizierten Bedarfs. Die Beschreibung umfasst: Anwendungsfälle, Systemanforderungen sowie die Beschreibung der Kernkomponenten und -prozesse. Die Anwendungsfälle und Systemanforderungen sind im Detail beschrieben (Anhang 2). Der dritte Teil des Konzepts (Kapitel 4) systematisiert die Aufgaben und Aufwände im Zusammenhang mit Implementierung und Betrieb. Die Aufwandsschätzung ist nach Aufgaben von Arbeitspaketen systematisiert (Anhang 4). Sie ist der Ausgangspunkt für die Projektierung der Implementierungsphase. Diese wird im letzten Kapitel skizziert.

2. Forschungskontext

2.1 Ausgangssituation

Die Historische Netzwerkforschung (HNR) ist ein interdisziplinäres Forschungsparadigma, das die Soziale Netzwerkanalyse (SNA) auf historische Fragen anwendet (Kerschbaumer et al. 2020, 282). Ihre Prämisse ist, dass “Beziehungen zwischen Entitäten erklärungsmächtig sind” (Düring et al. 2016, 6). Mit ihren Methoden und Hypothesen werden historische Entwicklungsprozesse nachgezeichnet, um “Strukturen zu entdecken, die nicht von allen [...] Akteuren erkannt werden, aber deren Form uns über zugrunde liegende soziale Mechanismen unterrichtet” (Lemercier 2012, 21). Datenerhebungsmethoden wie Interviews oder Beobachtungen sozialer Interaktionen sind in der Regel ausgeschlossen. Die HNR ist so auf die Quellen von Bibliotheken, Archiven und Museen angewiesen (Fangerau et al. 2021, 1). Trotzdem, oder gerade deswegen, ist die HNR eine akzeptierte Methode der Geschichtswissenschaft und mit breiten Themen, Fragestellungen und

[4] Video zu den Konzepten: https://sonar.fh-potsdam.de/assets/videos/sonar_prototype-demo_komp.mp4
[5] Grundlage der Prozessanalyse: https://github.com/sonar-idh/Transformer

wichtigen Beiträgen zur Methodenentwicklung vertreten (vgl. Rehbein 2020, 256/Ahnert et al. 2020). Die Anwendung netzwerkanalytischer Methoden nahm in den vergangenen zwei Dekaden stetig zu. Dennoch bleibt sie im Vergleich zu klassischen Methoden der historischen Forschung eine Nische (Rehbein 2020, 259). Dies hat forschungspraktische Ursachen, zu denen besonders die Datenerhebung zählt; denn Aufwände zur Sichtung vieler, dezentral überlieferter Quellen ist oft prohibitiv. Robert Gramsch-Stehfest konstatiert: "Zwar gibt es auch 'kleine Formen' der […] Netzwerkforschung, und […] didaktisch hat [sie] mit ihren Visualisierungstechniken viel zu bieten. Doch solange [...] massenhaft Daten manuell erhoben und verwaltet werden müssen, kann die Methode ihr Potential zweifellos nicht voll entfalten" (2020, 9). In nur zehn Jahren wurde aber dennoch aus einem methodischen Ansatz der Geschichtswissenschaft (Reitmayer/Marx 2010, Düring/Keyserlingk 2015) eine bedeutende Säule von Digital Humanities, Digital History sowie der historischen Informationswissenschaft (Rehbein 2020, 277)[6]; denn durch die HNR werden soziale Strukturen sichtbar, die durch klassische Textanalysen schnell übersehen werden können (Balck/Menzel/Petras 2021, 4 f.). Soziale Graphen sind daher ein Instrument, um Kontexte für die Quelleninterpretation aufzudecken (ebd., s.a. EGAD 2021, 6).

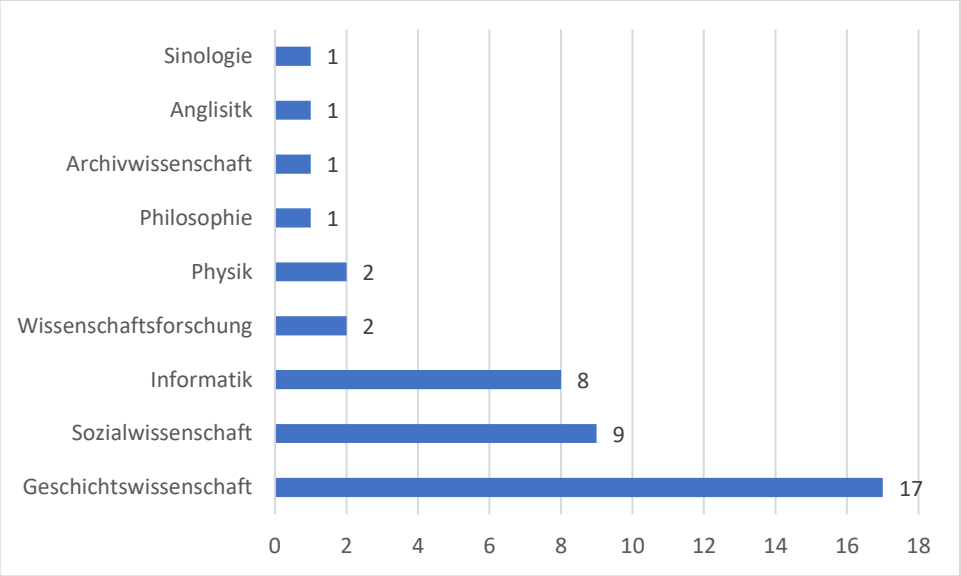

Abbildung 1: Fachlicher Hintergrund Vortragender auf der HNR+ResHist Conference, Juni 2021

Augenfällig ist die Diversität; die HNR umspannt ein breites inhaltliches Spektrum von der Antike bis zur Zeitgeschichte. Die Forschung ist oft international und interdisziplinär. Neue Plattformen wie historicalnetworkresearch.org fördern rege internationale Diskurse, z.B. über Publikationen, Workshops und Konferenzen. Die Historical Network Research Conference[7] wird jährlich an europäischen Universitäten drittmittelfinanziert ausgerichtet (2013 Hamburg, 2014 Ghent, 2015 Lissabon, 2017 Turku, Brno 2018). Ihre Schwerpunkte sind Methoden, Themen und Quellen. Das Forschungsfeld gewinnt durch Interdisziplinarität, die in der Teilnahme von Vertretern diverser Disziplinen auf Konferenzen (Abbildung 1[8]) und in Forschungsprojekten zum Ausdruck kommt. Studien (Interviews zu Forschungsdesign, Anforderungen und Visualisierungskonzepten) belegen die Fächervielfalt (Balck/Menzel/Petras 2021, 32; Schnaitter et al. 2021, 39).

[6] vgl. auch: Missionstatement der AG Graphen & Netzwerke des Verbands Digital Humanities zu Graphen und Netzwerke, https://graphentechnologien.hypotheses.org/ueber-das-blog (2021-09-06)
[7] https://historicalnetworkresearch.org/hnr-events/hnr-conferences/ (2021-09-06)
[8] Von 42 Referenten der HNR+ResHist Conference 2021 wurden die erstgenannten in der Verteilung berücksichtigt.

Zur Verbreitung der HNR trägt die Open-Science-Kultur bei. Diese zeigt sich in Aktivitäten wie die Gründung von Open-Access-Journals, z.B. „Journal of Historical Network Research“, die Pflege einer HNR-Bibliografie[9], der Fachaustausch über (Micro-) Blogs[10] und Mailinglisten[11] sowie neue Modelle des dynamischen Publizierens wie kollaboratives Schreiben (Rehbein 2020, 264). Ein regelmäßiger Programmteil von HNR-Konferenzen ist das Vorstellen und Bewerten von Software für die HNR. Eine systematische Desktop-Recherche hat zu dem Ergebnis geführt, dass aktuell mindestens 28 Anwendungen der HNR-Forschung zur Verfügung stehen (Anhang 1, Tab. 1). Sie können nach ihrem Schwerpunkt unterteilt werden: 1) Analyse und Visualisierung, 2) nur Analyse oder 3) nur Visualisierung. Viele sind kostenfrei und plattformunabhängig (Anhang 1, Tab. 2-4). Sie unterstützen den Im- und Export in gängigen Formaten, z.B. CVS oder JSON. Mit Ausnahme von SplitsTree können Daten als Graphen visualisiert und lokal als Grafik gespeichert werden. Die Analysetools bieten vielfältige Funktionen für statistische Berechnungen. Neue Workshops und Tutorials, bspw. auf der Website „The Programming Historian“[12], tragen zur Qualifizierung im Umgang mit Daten, statistischen Modellen und Anwendungen bei. Um eine Vorstellung über die Verbreitung von Anwendungen zu gewinnen, wurden alle 26 Beiträge des „Journal of Historical Network Research“ (2018-2020) ausgewertet: In 23 Artikeln wird der Einsatz von Anwendungen im Forschungsprozess erwähnt, die aber nur in 18 explizit benannt sind: Gephi (8) und Visone (4) dominieren die Verteilung. Genannt wurden weiterhin: VennMaker (2), Node XL (1), Nodegoat (1), Pajek (1), Palladio (1), Cytoscape (1) (Anhang 1, Tab. 5). Die Auswertung von Monografien und Sammelbänden, die nach der HNR-Bibliography Vol. 7, ausgewählt wurden, untermauern den Trend: Gephi ist das „bekannteste und vielseitigste“ Tool (Düring 2016, 175). Das „Handbuch Historische Netzwerkforschung“ empfiehlt für Datenanalysen ebenfalls: Nodegoat, VennMaker, NodeXL und Palladio sowie Pajek und UCInet (ebd., 177).

Die Auswertung von 36 Projekten hat gezeigt, dass sie hinsichtlich ihres Schwerpunkts analog zu Software gruppiert werden können: (1) Analyse oder (2) Visualisierung sozialer Netzwerke (Anhang 1, Tab. 6-8). Beiden Ansätzen ist die Datengrundlage gemein: Archivalien, z.B. Briefe aus Nachlässen, oder Texte, z.B. Zeitungen oder Biografien: Das Projekt „Tudor Networks“ beruht auf Briefen, das Projekt „Kindred Britain“ nutzt Biografien und „Mapping Notes and Nodes“ integriert heterogene (Meta-) Datenreihen (Alvarez/ van der Heuvel, 2014). Einzelne Projekte wie „Hidden Perspectives“ extrahieren Ereignisse, Personen und Körperschaften aus digitalen Archiven mit Algorithmen. Die Mehrzahl der ausgewerteten Projekte bieten webbasierte Nutzerschnittstellen an, sodass die zusammengetragenen Daten exploriert werden können. Doch nur sieben Projekte bieten ihre Daten auch für den Download an (Anhang 1, Tab. 7, 8). Einen anderen Ansatz in Bezug auf Datenerhebung und -speicherung wählte das Projekt „Agents of Change: Women Editors and Socio-Cultural Transformation in Europe, 1710-1920“: Das ERC Starting Grand Projekt griff auf die Wikidata zurück und erhob lediglich fehlende Daten, die in der Wikidata ergänzt wurden.

Die tendenzielle Einteilung von Tools und Projekten nach einem eher statistisch-mathematischen und einem eher visuell-explorativen Fokus kann mit Beobachtungen des Instituts für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin zusammengebracht werden:

[9] https://historicalnetworkresearch.org/bibliography (2021-07-05)
[10] Nur exemplarisch: Quantitative Methods in the Humanities: https://quanthum.hypotheses.org; Archeological Networks: https://archaeologicalnetworks.wordpress.com (2021-07-05)
[11] https://historicalnetworkresearch.org/external-resources (2021-07-12)
[12] The Programming Historian. Lessons: https://programminghistorian.org/en/lessons/ (2021-07-05)

> „Es sind [...] zwei Gruppen [...]: [...] die explorativ, prozesshaft und [...] die [...] quantitativ, sozialwissenschaftlich Forschenden. Die erste Gruppe möchte [...] Themen erschließen, auf neue Fragestellungen stoßen oder einfach Unbekanntes entdecken und greift dabei gerne auf vorgefertigte Visualisierungen (mit [...] Dokumentation [...]) zurück. Die zweite Gruppe [...] konzentriert sich [...] auf die Datengrundlage und sieht einen großen Nutzen in […] standardisierten Daten aus verschiedenen Datenquellen. Wobei der Vorteil weniger in der Aufbereitung [...] mittels Visualisierung gesehen wird als vielmehr in der Möglichkeit des Exports der Daten [...] für [...] Analysen. Visualisierungen werden bevorzugt selbst gebaut und erst im Anschluss an die Analyse verwendet, um eine Beeinflussung durch bildliche Verzerrungen zu verhindern.“ (Balck/Menzel/Petras 2021, 23)

Gerade die zweite Gruppe greift eher auf generische Forschungsumgebungen wie R oder Jupyter Notebooks zurück. Die Wahl von Forschungsansätzen ist abhängig von der Forschungsfrage und Fachwissenschaftler wechseln entsprechend zwischen den Ansätzen (ebd., 8). Ähnlich verhält es sich mit der Entscheidung für eine oder mehrere Anwendungen: Sie ist nicht nur von Funktionen, sondern von Usability, Support und Performance beeinflusst. Ein Faktor ist aber die Qualifikation: fehlen Statistik- und grundlegende Programmierkenntnisse, ist die Anwendung von quantitativen Ansätzen, das heißt, auch die Nutzung von eher generischen Forschungsumgebungen gering. In Interviews wurden aber auch Aussagen getroffen, dass gerade die jüngere Generation bereit ist, Data Science und fragliche Anwendungen zu vertiefen (Balck/Menzel/Petras 2021, 18). Die Stärke moderner Forschungsumgebungen für das Analysieren und Visualisieren, Dokumentieren und Publizieren von Forschungsergebnissen demonstriert eine Fallstudie der Fachhochschule Potsdam zu einem Curriculum[13] für die HNR (Bludau/Dörk 2021, 37 f.). Dieses zeigt zugleich die Breite der Anwendungsmöglichkeiten generischer Forschungsumgebungen und das hohe Maß an Individualisierbarkeit gerade auch in Bezug auf individuelle Analyseanforderungen.

2.2 Anforderungen

Die Ausgangssituation vermittelt ein differenziertes Bild über den Stand der HNR. Das Spektrum der Methoden reicht von qualitativ-rekonstruktiven bis zu quantitativ-standardisierten Ansätzen. Die Entscheidung für einen methodischen Ansatz und die Wahl der Instrumente ist abhängig vom Forschungsinteresse. Für die Analyse von Daten – ob primär mit Visualisierungen, mit statistisch-mathematischen Maßzahlen oder durch Ergänzung beider Richtungen – stehen schon heute eine Vielzahl von Softwarelösungen zur Verfügung. Der begrenzende Faktor der HNR ist der Zugang zu Daten, die visuell aufbereitet resp. statistisch ausgewertet werden können. SoNAR kann an diesem Punkt anknüpfen, das heißt, dass der Schwerpunkt auf der Integration von Daten und ihre Aufbereitung als Netzwerkdaten liegen wird. Die Datenanalyse erfolgt je nach Anforderung mit geeigneten nachnutzenden Systemen. Die Anforderungen innerhalb dieses konkretisierten Rahmens können in zwei Gruppen unterteilt werden: (a) Anforderungen an die Daten und (b) Anforderungen an den Datenzugang. Tabelle 1 führt die zentralen Anforderungen auf:

(a) Anforderungen an Daten	(b) Anforderungen an den Datenzugang
1 Datenqualität	1 Zugriff für nachnutzende Systeme
2 Datenprovenienz	2 Dokumentation verfügbarer Daten
3 Reproduzierbarkeit	

[13] https://github.com/sonar-idh/jupyter-curriculum

Tabelle 1: Gruppierung der Anforderungen an SoNAR

a) Anforderungen an Daten

Eine Anforderung nimmt Bezug auf die *Datenqualität*, die SoNAR für die wissenschaftliche Arbeit mit nachnutzenden Systemen gewährleisten muss (Balck/Menzel/Petras 2021, 21). Als Qualität wird die systemunabhängige Eindeutigkeit der Daten definiert, das heißt, die Daten müssen ohne Aufbereitungsaufwände für visuelle und statistisch-mathematische Analysen mit nachnutzenden Systemen bereitstehen. Die Anforderung adressiert somit die Aufbereitung der Input-Daten von anbietenden Systemen. Sie erfordert Maßnahmen für die persistente Adressierbarkeit der Elemente eines Graphens und das Zusammenführen von gleichen Entitäten, die in Input-Daten durch Identifier diverser Wissensbasen wie GND, Wikidata, ISNI oder VIAF referenziert sind.

Die Kernanforderung betrifft die Transparenz der Herkunft und Verarbeitungsschritte der für wissenschaftliche Analysen bereitgestellten Daten (*Datenprovenienz*). Die Anforderung resultiert aus den allgemeinen Grundsätzen wissenschaftlichen Arbeitens und wurde in Studien über ein modellhaftes Forschungsdesign als grundlegend für SoNAR identifiziert (Fangerau et al. 2021, 15, Fachanforderungen R029, R030, R031, R032; Bludau/Dörk 2021, 4 f., 24 f.; Balck/Menzel/Petras 2021, 20 f.; Schnaitter et al. 2021, 32 f.). Provenienzdaten sollen automatisch mit den einzelnen Phasen der Datenprozesskette bei einer Implementierung von SoNAR erzeugt werden. Dies greift aber auch auf anbietende Systeme vor, indem Anforderungen an Input-Daten gestellt werden. Dies schließt Datenstandards ein (s. Kapitel 2.3). Hierzu zählen u.a., dass anbietende Systeme, die datenerfassenden und bestandshaltenden Einrichtungen nach ISO 15511 identifizieren. Hierzu zählt auch, dass maschinelle Methoden, die anbietende Systeme einsetzen, nach dem Standard PROV-O[14] (Algorithmen, Konfidenzwerte[15]) dokumentiert und übertragen werden.

Als weitere Anforderung wurde die *Reproduzierbarkeit* identifiziert. Diese Anforderung ist hier ein Unterfall der Anforderung nach Datenprovenienz und bezeichnet die Option, auf vergangene Datenstände und Transformationsmodelle zurückgreifen zu können. Die Anforderung resultiert aus dem Bedarf zur Dokumentation von Datenquellen, zur Transparenz der Datenaufbereitung und zu einem standardisierten Zugang zu Daten für Replikationsstudien oder sekundäre Nutzung gleicher Daten für neue Forschungsperspektiven (Balck/Menzel/Petras 2021, 25).

b) Anforderungen an den Datenzugang

Als Forschungstechnologie legt SoNAR den Schwerpunkt auf die Integration und Aufbereitung von Daten. Dies schließt Anforderungen an den Datenzugang ein, die aus drei verschiedenen Perspektiven resultieren: (1) der Zugriff auf das SoNAR-System über *nachnutzende Systeme* für Datenanalysen, (2) die Dokumentation der im SoNAR-System *verfügbaren Daten* sowie (3) der Zugang zu vergangenen Datenstände z.B. für Replikationsstudien reproduzieren zu können.

[14] https://www.w3.org/TR/prov-o/

[15] Im Projekt SoNAR (IDH), in dessen Rahmen das hier vorliegende Konzept erarbeitet wurde, konnten maschinelle Verfahren zur Erprobung von Named Entity Recognition (NER) und Named Entity Linking (NEL) erprobt werden. Die gewonnenen Daten wurden einerseits in ein Graphdatenbank zur Erprobung integriert (https://sonar:sonar2021@h2918680.stratoserver.net:7473/browser/), andererseits evaluiert (Menzel et al. 2021 https://doi.org/10.1515/9783110691597-012). Für NEL-Algorithmen wurden Konfidenzwerte in Input-Daten ergänzt, z.B. https://github.com/sonar-idh/Goldstandard/blob/main/02_ocr_corrected-EL-NER/27646518_1897-05-05_26_225_020-NER-EL.tsv (Spalte: conf).

Die erste Anforderung besteht darin, dass die Daten – Netzwerkdaten, Provenienzdaten – nach den FAIR-Prinzipien[16] unabhängig vom Anbieter der Forschungstechnologie (Service-Provider) in offenen Formaten und über eine offene Schnittstelle (Machine-2-Machine) abgefragt und in das *nachnutzende System* direkt übernommen werden können. Als Formate können berücksichtigt werden: RDF-Formate wie JSON-LD oder CSV (Bludau/Dörk 2021, 36; Balck/Menzel/Petras 2021, 11 f.). Eine Analyse gängiger Anwendungen für die HNR weist auf die starke Verbreitung von CSV, JSON und XML hin (Anhang 1). Die *Programmierschnittstelle* (API) wird vorrangig von quantitativ-orientierten Anwendern präferiert (Kapitel 2.1). Ergänzend wird ein Download von Daten über eine webbasierte *Nutzerschnittstelle* nachgefragt (Schnaitter et al. 2021, 25 f.).

Eine zweite Anforderung ist die Dokumentation der durch das SoNAR-System bereitgestellten Daten. In diesem Kontext werden zwei Fachanforderungen aufgegriffen: die *Dokumentation der Datenherkunft und -verarbeitung* (Provenienzdaten) und *die Dokumentation der verfügbaren Datenkategorien und Werte* (*Netzwerkdaten*). Die Dokumentation beider Datentypen unterstützt Fachwissenschaftler bei der Sichtung der Daten (Fangerau et al. 2021, 2 ff.). Die Dokumentation erfolgt durch die Nutzerschnittstelle und ist ein Element zur Prüfung und Selektion von Daten für den Download. Es umfasst klassische, klassisch-explorative Ansätze (Suchformular, Facetten) und innovative Visualisierungs- und Interaktionskonzepte zur Darstellung von Netzwerkgraphen.

Durch die Erprobung von SoNAR konnten in diesem Zusammenhang Synergieeffekte zwischen den Anforderungen herausgearbeitet werden. Eine Nutzerschnittstelle zur Dokumentation und Selektion von Daten ist zugleich ein niedrigschwelliges Angebot für explorativ-prozesshafte und quantitativ-sozialwissenschaftlich orientierte Anwender, um insbesondere frühe Phasen eines Forschungsprozesses zu unterstützen (Tabelle 2). Die Unterstützung der Datenerhebung ist die Kernfunktion von SoNAR und erfordert Maßnahmen zur Übernahme von Daten in nachnutzende Systeme zur Analyse. Die Visualisierungs- und Interfacekonzepte der Nutzerschnittstelle können aber Aktivitäten eines Forschungsprozesses, speziell bei der Planung und Vorbereitung fördern; „Visualisierungen dienen neben der Exploration auch zur Fragen- und Hypothesengenerierung sowie einer ersten Interpretation relevanter Daten […]“ (Schnaitter et al. 2021, 21). In Tabelle 2 sind die durch Synergie der Anforderungen zusätzlich gewonnenen Leistungen hellgrün markiert.

1 Planen	2 Vorbereiten	3 Durchführen
Forschungsfrage entwickeln	Forschung operationalisieren	Daten analysieren
Informationen recherchieren	Datenquellen ermitteln	Ergebnisse beschreiben
Erklärungsansätze ermitteln	Daten erheben	Forschung publizieren

Tabelle 2: Einordnung von SoNAR-Anwendungsszenarien im Forschungsprozess[17]

So zeigt auch die Befragung zur Umsetzung des modellhaften Forschungsdesigns (Fangerau et al. 2021) in der prototypischen Demonstration[18] (Bludau/Dörk 2021), dass Anwendungsszenarien „sämtlich […] als realistische und sehr geläufige Forschungsfragen und -prozesse bezeichnet“ werden (Schnaitter 2021, 35). Zudem wurden „die für die Exploration von großen Datenmengen

[16] https://www.go-fair.org/fair-principles/
[17] Zum (idealtypischen) Forschungsprozess: Balck/Menzel/Petras 2021, 6 ff.; Fangerau 2021, 2 ff.)
[18] https://sonar.fh-potsdam.de/prototype/

konzipieren Visualisierungsprototypen von allen Proband:innen als nützlich und sinnvoll für die Hypothesenbildung und Exploration bewertet“ (ebd., 36). Die Hypothesenbildung ist in Tabelle 2 in „Forschung operationalisieren“ enthalten.

Durch die Studien zum modellhaften Forschungsdesign des Projekts SoNAR (IDH) wurden weitere Anforderungen formuliert. Sie betreffen die Datenanalyse oder Publikation von Ergebnissen und sind daher an nachnutzende Systeme gerichtet. Sie sind hier soweit aufgegriffen, wie den Zugang zu Daten für die weiteren Schritte eines Forschungsprozesses betreffen. Das Implementierungs- und Betriebskonzept nimmt explizit folgende Anwendungsfälle (Use Cases, UC) auf (Anhang 2):

- UC1: Daten importieren und in einem HNR-Datenmodell bereitstellen (*Aufbereitung*)
- UC2: Einen Teildatenbestand selektieren (*Auswahl und Dokumentation*)
- UC3: Die Visualisierungsform auswählen (*Explorieren und Dokumentation*)
- UC4: Daten zur externen Analyse herunterladen (*Zugang Download*)
- UC5: Automatisch Daten herunterladen (*Zugang Programmierschnittstelle*)
- UC6: Weitere Aktionen mit dem SoNAR System durchführen (*Tutorials*)

Die Anwendungsfälle sind einschließlich Unterfälle beschrieben. Der erste Anwendungsfall greift das Kernziel der Forschungstechnologie SoNAR auf: Aufbereitung und Bereitstellung von Daten. Der zweite Anwendungsfall berücksichtigt die Anforderungen nach Dokumentation und Auswahl der Daten über eine Nutzerschnittstelle mit Retrieval-Funktionen. Der dritte Anwendungsfall unterstützt die Exploration der Datenmenge und Präzisierung der Selektion. Er berücksichtigt das Entwickeln einer Forschungsfrage und die Formulierung von Hypothesen. Die Anwendungsfälle vier und fünf beschreiben den Bezug der Daten für Datenanalysen mit nachnutzenden Systemen. Unter dem sechsten Anwendungsfall sind Anforderungen zusammengefasst, die die praktische Arbeit fördern: auf Originaldaten zugreifen, die Arbeit nach Unterbrechung im SoNAR-System fortführen sowie eine Einführung in die Arbeit mit dem SoNAR-System. Bereits die prototypische Demonstration setzt den Anwendungsfall um, auf Originaldatensätze zugreifen zu können und demonstriert die Chance, mit SoNAR Datenquellen ergänzend zu etablierten Katalogen ermitteln zu können – aus der Akteursperspektive.

2.3 Datenquellen

Der „Rohstoff“ der Forschungstechnologie SoNAR sind Daten. Die Input-Daten (Eingangsdaten) beschreiben Quellen (Metadaten) oder Akteure (Normdaten), die mit den Quellen identifiziert werden. Hierzu zählen u.a. Bestandsbildner, Urheber, Adressaten oder Personen, die auf einer Quelle abgebildet oder aber ihr Gegentand sind. Quelle kann jede Ressource sein, z.B. Akte oder Korrespondenz, Fotographie oder Video, Tagebuch oder Protokoll, Monographie oder Artikel. Sie sind oft die einzigen Quellen für das Erheben von Netzwerkdaten. Kataloge von Bibliotheken oder Archiven können ebenfalls Datenquelle für die HNR sein (Fangerau et al. 2021, 4). Verschiedene Quellentypen dokumentieren unterschiedliche Arten von Beziehung wie Finanzbeziehungen, die Transmission von Ideen und Wissen, das Zusammenwirken in Organisationen wie Vereinen und Parlamenten (ebd., 7). Eine Analyse der Datenkategorien der Gemeinsamen Normdatei (GND) und Expert Group Archival Description (EGAD/ICA) zum Erschließungsmodell Records in Context (RiC-O) zeigt das breite Spektrum von potenziell verfügbaren, regelbasierten und strukturierten Daten über Quellen und Akteure (Anhang 3). In begleitenden Studien des Erprobungsprojekts SoNAR (IDH) wurde das Potenzial der Daten der Gemeinsamen Normdatei, der bibliographischen Daten der Zeitschriftendatenbank, der Kataloge der Deutschen Nationalbibliothek und der

Staatsbibliothek zu Berlin – Preußischer Kulturbesitz sowie des Kalliope-Verbunds (Archivdaten) empirisch überprüft. Das Ergebnis ist differenziert:

Normdaten der Gemeinsamen Normdatei (GND) beschreiben Akteure und oft qualitative soziale Beziehungen, z.B. familiäre oder professionelle Relationen. Die Daten der GND sind oft verlinkt und durch Nutzung der GND-ID zur Identifikation der Akteure in Metadaten, sind Datenquellen überregional vernetzbar. Die Datenerfassung ist vorrangig an die Bedürfnisse bibliothekarischer Identifikation von Personen und Organisationen, nicht aber auf ihre differenzierte Beschreibung ausgerichtet (ebd., 34). Die Ergänzung von Normdaten über Akteure und soziale Beziehungen mit *Metadaten* erweitert jedoch ihre Beschreibung teils signifikant. Bibliographische Metadaten ergänzen Aussagen über Themen, die u.a. allgemeine Berufsbezeichnungen der GND präzisieren: Während in der GND die Berufsbezeichnung oft nur mit einem allgemeinen Sachbegriff wie Arzt verlinkt ist, können Sachbegriffe zu den Publikationen der Person den Arbeitskontext bedeutend erweitern, z.B. Urologie, Physiologie etc. (ebd., 22 f.). Ko-Autoren- und Ko-Herausgeberschaften erweitern in einem bedeutenden Umfang professionelle Kollaborationen. Eine herausragende Datenquelle sind die Archivdaten, speziell des Kalliope-Verbunds. Das Potenzial der Daten für ein Forschungsprojekt hängt aber von der Erschließungstiefe ab (Konvolut vs. Einzeldokument) ab (ebd., 23). Trotz der Einschränkungen bietet allein der Testdatenbestand einen Zugang zu den sozialen Beziehungen von rund 2,5 Millionen Personen und 300.000 Organisationen seit 1750[19]. Norm- und Metadaten sind so „grundsätzlich geeignet“ (ebd., 34), aber ihre Potenziale als eine Datenquelle für die HNR ist oft unbekannt (Schnaitter 2021, 23).

Norm- sowie bibliographische und archivische Metadaten sind so vielfältig wie die Einrichtungen und ihre Bestände. Für eine optimale Versorgung setzt SoNAR auf Verbunddatenbanken. Deren Daten bilden den Kerndatenbestand. Speziell die Öffnung der GND[20] sowie die Möglichkeit einer forschungsgeleiteten, normdatenbasierten Erschließung von Archivquellen im Kalliope-Verbund können Forschungsprojekte bereits jetzt bei der Datenerhebung zielgerichtet unterstützen. Darüber hinaus wurde in Interviews die Anforderung formuliert, Forschungsdaten in SoNAR hochladen zu können, die außerhalb von Bibliotheks- und Archivinformationssystemen erhoben wurden (Schnaitter et al. 2021, 31). Diese Forderung wird aufgegriffen und ein Datenupload aus einzelnen Datenrepositorien unterstützt. Datenquellen können daher bspw. auch intellektuell oder maschinell erzeugte Annotationsdaten digitaler Editionen und Volltextrepositorien (Menzel et al. 2021) oder auch Informationsangebote wie Professoren- und Matrikelportale[21] sein. Eine notwendige Voraussetzung ist die Konformität mit Datenstandards und -formaten. Hierzu zählt, dass die Daten bereits strukturiert und verlinkt in einem RDF-Format aufbereitet sind und für die Verlinkung etablierte Wissensbasen wie GND, Wikidata oder VIAF Verwendung finden.

Verbunddatenquellen	Spezialisierte Datenquellen
Normdaten	Volltextrepositorien
Metadaten	Digitale Editionen
	Fachinformationen

Tabelle 3: Übersicht der Art der Datenquellen

[19] vgl. Statistiken für Personen und Organisationen https://sonar.fh-potsdam.de/prototype/
[20] https://gnd.network/Webs/gnd/DE/Projekte/projekte_node.html
[21] http://matrikel.uni-rostock.de/

Bereits mit einer Implementierung von SoNAR wird die Integration internationaler Angebote zu berücksichtigen sein; die Erforschung von historischen Netzwerken erfordert den Zugang zu den Quellen und Daten außerhalb des deutschsprachigen Raums. Die Implementierung von SoNAR wird daher die Kooperative „Social Network and Archival Context“ (SNAC)[22]. SNAC ist eine seit 2010 aufgebaute und nunmehr auch international etablierte Normdatei für die Erschließung von Archivbeständen. Der SNAC-Datenbestand erweitert den Zugang zu Quellen in Archiven und Bibliotheken insbesondere im anglo-amerikanischen Raum in bedeutendem Umfang.

Da die Studien zum modellhaften Forschungsdesign die Eignung der Verbunddatenbank gezeigt haben, wird eine erste Implementierung folgende Datenquellen berücksichtigen:

- Gemeinsame Normdatei (GND)
- Kalliope-Verbunddatenbank (KPE)
- Zeitschriftendatenbank (ZDB)
- Gemeinsamer Bibliotheksverbund (GBV)[23]
- Social Networks and Archival Context (SNAC)[24]

Zur exemplarischen Integration einzelner Datenquellen durch Upload wird bei der Projektierung eine geeignete Partnereinrichtung berücksichtigt. Dies könnten bspw. Anbieter von Professoren- und Matrikelportalen, die mit GND-Identnummern referenziert sind, sein.

3. Implementierung

3.1 Kernkomponenten

Die Umfeldanalyse und die Betrachtung von Phasen von Forschungsprozessen der HNR haben zu präzisen Anforderungen an SoNAR geführt. Der Fokus wird auf der Integration, Aufbereitung und Bereitstellung von Daten für wissenschaftliche Analysen der HNR liegen. Vier Eckpunkte sind für eine Implementierung als Forschungstechnologie konstitutiv: Maßnahmen (1) zur Sicherung der des Zugangs zu Ursprungsdaten und Quellen durch Provenienzdaten, (2) zur Reproduzierbarkeit durch die Sicherung der Input- und Output-Daten einschließlich öffentlicher Versionierung der Transformations- und Datenmodelle, (3) für eine niedrigschwellig zugänglichen Dokumentation und Selektion von Daten für Datenanalysen über eine Nutzerschnittstelle sowie (4) zum direkten Zugang zu Netzwerk- und Provenienzdaten über nachnutzende Systeme.

Die Datenprozesskette zur Integration der Input-Daten und Aufbereitung als Netzwerkdaten wird auf dem Standard „Resource Description Framework“, RDF, und entsprechende Semantic Web Technologien beruhen. Durch den Ansatz können insbesondere heterogene Datenbestände standardbasiert integriert und zugleich eine Community-orientierte Weiterentwicklung des HNR-Datenmodells gefördert werden (Basis ist die Web Ontology Language, OWL). Offene Standards sind für die Wirtschaftlichkeit von SoNAR – für den Betrieb, die Anschlussfähigkeit an anbietende und nachnutzende Systeme sowie für internationale Kooperationen – eine Voraussetzung.

Die Kernkomponenten des SoNAR-Systems sind aufgeteilt auf zwei Module: Das Backend, das für die Sammlung, Transformation, Anreicherung und Distribution der Daten zuständig ist, und das

[22] Letter of Intent im Anhang zum DFG-Abschlussbericht https://snaccooperative.org/
[23] Bereits der Teildatenbestand der SBB-PK im GBV hat sich als gleichwertig zum DNB-Datenbestand gezeigt. Bei der Projekttierung einer Implementierungsphase wird im Vorfeld die Quelle für bibliographische Metadaten noch einmal zu prüfen sein. Eine zu prüfende Option ist Culturegraph.org.
[24] Ein Letter of Intent liegt dem Projektabschlussbericht bei.

Frontend, das die Recherche und Exploration der aufbereiteten Daten sowie die Nachnutzung über verschiedene Kanäle ermöglicht.

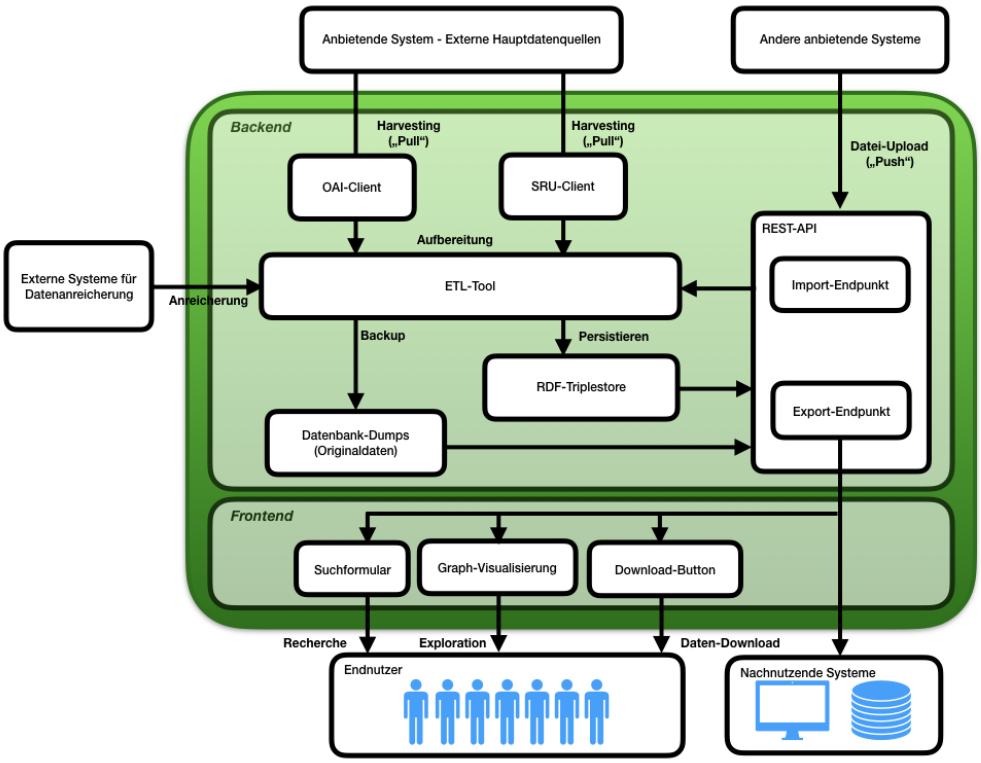

Abbildung 2: Schematische Darstellung der Kernkomponenten und -prozesse

Das Backend stellt vor allem eine Datenbank zur Verfügung, in der die gesammelten Daten gesichert, transformiert und für den Zweck der HNR aufbereitet werden, während das Frontend eine Website mit einer Nutzerschnittstelle (User Interface, UI) zur Interaktion mit dem System und einer Visualisierungskomponente zur Verfügung stellt.

Es lassen sich folgende Kernkomponenten des Systems identifizieren (Abb. 2):

- **OAI-Client:** OAI (Open Archives Initiative) definiert eine offene Schnittstelle, um Metadaten anbietender Systeme zu beziehen[25]. Die Kommunikation erfolgt zwischen Datenlieferanten (Data Provider) und Dienstanbieter (Service Provider) automatisiert. SoNAR tritt als Dienstanbieter auf, der den Datenbestand von Datenlieferanten bezieht.
- **SRU-Client:** Über SRU (Search/Retrieval via URL) kann mit Suchbegriffen gezielt in indexierten Daten eines Datenlieferanten gesucht werden[26]. Die mit der Abfrage identifizierte Teildatenmenge kann übernommen werden. SoNAR als Service Provider kann SRU als eine zweite, optionale Methode neben OAI bedarfsabhängig nutzen.
- **Datenbank-Dumps / Datenspeicher:** … ist der Ablageort für Input-Daten inklusive der technischen Metadaten. Sie werden als komprimierte Archivdateien nach dem Liefer- bzw. Harvesting-Datum aufbewahrt. Die Erstellung von Archivdateien ist Teil des ETL-Prozesses (s.u.) und Voraussetzung für die Reproduktion eines ETL-Prozesses zur Unterstützung von Forschungsprozessen.

[25] Open Archives Initiative: http://www.openarchives.org/ (2021-09-23)
[26] Search/Retrieval via URL: https://www.loc.gov/standards/sru (2021-09-23)

- ETL-Tool: Für die prototypische Demonstration erfolgte die Aufbereitung XML-basierter Eingangsdaten mit Python-Skripten[27]. Für die produktive Implementierung werden für den Aufbereitungsprozess etablierte Tools für „Extract, Transfer, Load (ETL)"-Prozesse eingesetzt, die eine kontrollierte, konsistente Nachbearbeitung von Datenlieferungen und die Datenablage in der Zieldatenbank unterstützen. Sie sind erheblich einfacher und kostengünstiger zu implementieren und reduzieren signifikant das Risiko eines unzureichenden Wissenstransfers bei personellen Änderungen (s. Kapitel 3.3).
- RDF-Triplestore: … ist das Ziel des Aufbereitungsprozesses von Daten anbietender Systeme. Es ist eine NoSQL-Datenbank und priorisiert im Vergleich zu relationalen Datenbanksystemen Beziehungen zwischen den Daten. Dadurch wird die Abbildung hierarchischer und vernetzter Strukturen vereinfacht. Während andere Datenbanken zur Abfragezeit Beziehungen durch aufwändige Join-Operationen (SQL) berechnen, speichert ein RDF-Triplestore Verbindungen der Daten im Modell. Der Zugriff auf Knoten und Kanten in einem nativen RDF-Triplestore ist so eine Operation mit einer konstanten Laufzeit und ermöglicht es, schnell Millionen von Kanten pro Sekunde zu durchlaufen. Das macht sie zur effizientesten Lösung für SoNAR.
- Suchformular: … ist eine zentrale Komponente des UI. Hier laufen die Interaktionen zur Datenselektion zusammen. Sie kann zur gezielten Suche nach Werten genutzt werden. Es unterstützt die einfache Suche, die Suche mit und in Facetten sowie eine Expertensuche (Retrieval-Sprache: SPARQL)
- Visualisierung: … ist eine weitere zentrale Komponente des UI. Mit ihr können die in den Daten identifizierten Netzwerke in unterschiedlichen Formen dargestellt werden. Sie kann wie das Suchformular für einen explorativen Einstieg in die Daten genutzt werden. Die Netzwerke lassen sich in einfacher und komplexer Form darstellen. Die Darstellung kann um weitere Kriterien erweitert werden (z.B. Zeit, Raum, Klassen von Entitäten). Knoten und Kanten sind mit den Ausgangsdatensätzen verlinkt. Zusätzlich werden Häufigkeitsverteilungen von Merkmalsausprägungen von Knoten und Kanten angezeigt. Es können Vermittler/Hubs identifiziert werden. Akteure eines Netzwerks mit gemeinsamen Merkmalen können hervorgehoben werden, z.B. Themen, Orte und Affiliationen mit Organisationen. Die Reduktion eines visualisierten Graphens mehrerer Akteure auf ein egozentrisches Netzwerk eines Akteurs ist möglich. Visualisierungen können in den Formaten SVG und PNG gespeichert werden (Kapitel 3.2).
- Download-Button: … ist eine Komponente des Frontend, um den Datenexport über das UI auszulösen. Er löst eine mehrstufige Interaktion aus, um den Export in Bezug auf den Umfang und die strukturelle Form zu parametrisieren (z.B. Datenkategorien, -format).
- Export-Endpunkt: … bietet nachnutzenden Systemen Zugang zu den Daten von SoNAR über eine REST-API. Durch Parametrisierung der Abfrage sollen Einschränkungen in Bezug auf Teilnetzwerke analog zum Download-Button möglich sein.
- Import-Endpunkt: … bietet anbietenden Systemen über eine REST-API die Möglichkeit, Daten für die Prozessierung durch das ETL-Tool und die entsprechende Anreicherung der Datenbank zur Verfügung zu stellen. Er erfordert eine Authentifizierung an SoNAR. Dieser Prozess sieht einen vorgelagerten Prozess außerhalb des SoNAR-Systems vor: eine datenanbietende Stelle schickt eine Anfrage an den Betreiber von SoNAR (per E-

[27] https://github.com/sonar-idh/Transformer

Mail o.ä.). Der Betreiber prüft die Konformität der anbietenden Stelle mit Vorgaben für SoNAR (Datenstandards). Im Fall einer positiven Prüfung erhält die anbietende Stelle einen Authentifizierungstoken (Access-Token). Dieser wird beim Kontaktaufbau an den Import-Endpunkt übergeben und von der API validiert.

3.2 Kernprozesse und Funktionen

Die in 3.1 beschriebenen Kernkomponenten sind Voraussetzung zur Durchführung der zentralen Systemprozesse (Abb. 2). Die Prozesse sind: (1) Datenintegration und -aufbereitung (Backend) sowie (2) Bereitstellung der Daten (Frontend).

(1) Datenintegration und -aufbereitung (Backend)

Die Forschungstechnologie SoNAR bezieht Daten zur Integration für die HNR aus sehr diversen Quellen. Sie müssen integriert und aufbereitet werden. Der erste Schritt ist das Aggregieren der Daten über einschlägige Schnittstellen (OAI, SRU). Für Hauptdatenquellen (Kapitel 2.3) geschieht dies regelmäßig automatisch. Weitere, von einem Metadatenmanager zertifizierte anbietende Systeme stoßen die Übertragung von Daten an SoNAR selbständig an.

Die Input-Daten werden im Datenspeicher abgelegt. Bei Hauptdatenquellen werden Formate, die vom Anbieter und dessen Kunden standardmäßig produktiv genutzt werden, auch für SoNAR präferiert. Dadurch soll SoNAR unabhängig von alternativen Angeboten des Datenanbieters, z.B. Datendumps, auf einen aktuellen, stabilen Datenstand der Hauptdatenquelle zugreifen können. Zertifizierte Anbieter, die selbständig die Übertragung anstoßen, bieten Daten im RDF-Format an (unter Beachtung vorgegebener Datenstandards).

Für jede Datenquelle wird immer eine neue Archivdatei angelegt, die den kompletten, aktuellen Datenbestand eines Zeitpunkts enthält (Input-Daten). Der Zeitpunkt entspricht dem Zeitpunkt der Aktualisierung der Netzwerkdaten (Output-Daten). Die Archivdateien werden versioniert und sind online verfügbar. Zudem wird eine Version der Netzwerkdaten mit jeder Aktualisierung als Archivdatei online bereitgestellt[28]. Die Versionsbeschreibung der Archivdatei der Output-Daten enthält a) die URL der jeweiligen Archivdateien der Input-Daten, die in eine Version der Output-Daten eingegangenen sind, sowie b) die URL der Konfiguration der ETL-Komponente (z.B. des GitHub-Repositoriums). Die URL der Versionsbeschreibung kann zur Zitation genutzt werden[29].

Die Archivierung und Versionierung der Input- und Output-Daten sowie der ETL-Konfiguration adressiert die Anforderung nach Transparenz und Reproduzierbarkeit von Forschungsprozessen. Die Lösung dokumentiert die Genese der Ausgangsdaten und ermöglicht die Wiederherstellung von SoNAR zu einem Zeitpunkt X, um Probleme der Datentransformation ex-post zu erkennen[30].

Regelmäßig werden aktuelle Input-Daten in mehreren Schritten über das ETL-Tool aufbereitet. In einem ersten Schritt werden Validierungs- und Konsistenzprüfungen durchgeführt, um die

[28] Bei der Inbetriebnahme nach einer Implementierung wird zunächst ein monatlicher Takt angestrebt.
[29] Alternativ kann bei einer Implementierung für jede Versionsbeschreibung eine DOI erzeugt werden.
[30] Eine Alternative ist die Abbildung der Zustände von Knoten und Kanten im Graph-Datenmodell (s. bspw. https://medium.com/neo4j/keeping-track-of-graph-changes-using-temporal-versioning-3b0f854536fa). Sie ist jedoch zum aktuellen Zeitpunkt und mit Blick auf Datenmodell und Betrieb zu experimentell und komplex.

Datenqualität zu sichern, und, wenn sie in einer XML-Struktur vorliegen, in RDF transformiert[31]. In einem zweiten Schritt werden Daten maschinell aufbereitet. Hierzu zählen:

1) Gleiche Entitäten werden zusammengeführt. Input-Daten verschiedener Datenquellen können Entitäten durch Identifier (ID) diverser Normdateien identifizieren, z.B. Wikidata, GND, VIAF, ISNI oder SNAC. Über eine ID-Konkordanz, z.B. Lobid.Org, Wikidata oder VIAF, können gleiche Entitäten unter einer ID in SoNAR zusammengeführt werden.
2) Zusätzlich werden Beschreibungen zu Personen und Organisationen der GND um Daten für ausgewählte Datenkategorien aus der Wikidata maschinell ergänzt. Für Entitäten, die nicht in der GND beschrieben, sondern nur über Wikidata, SNAC oder VIAF identifiziert sind, werden Beschreibungen dieser Normdateien abgefragt und in SoNAR integriert.

In einem dritten, letzten Schritt erfolgt die Abbildung der Input-Daten auf das Datenmodell der HNA (Anhang 3). Ihre Transformation erfolgt automatisiert durch Skripte des ETL-Tools, die nicht durch Code, sondern eine domänenspezifische Sprache konfiguriert werden. Die gewonnenen Daten – Netzwerkdaten, Provenienzdaten – werden an den RDF-Triplestore übertragen.

Die Aufbereitung der Input-Daten erfordert eine systembibliothekarische Betreuung, sodass die Rolle eines Metadatenmanagers definiert wurde, um die Transformationsregeln des ETL-Tools festzulegen und anzupassen. Zur Tätigkeit zählt auch die Zertifizierung anbietender Systeme und die Sicherung der Datenprovenienz und -verarbeitung der über SoNAR bereitgestellten Daten.

(2) Bereitstellung der Inhalte (Frontend)

Für die Recherche und Interaktion mit dem RDF-Triplestore steht ein Web-Frontend mit diversen Funktionalitäten zur Verfügung. Über das Web-Frontend wird das Angebot im Web auffindbar. Es ist die Schnittstelle für Recherche, Exploration und Download, und es wird, soweit möglich, responsiv umgesetzt. SoNAR soll auf klassischen stationären und mobilen Rechnern als auch auf Tablets und Smartphones genutzt werden können[32]. Für Recherche, Exploration und Download im UI wird ein Dashboard mit drei Arbeitsflächen bereitgestellt: (1) Datenauswahl (Suche, Facetten), (2) Graph-Visualisierung ausgewählter Datensegmente sowie (3) Dokumentation. Abhängig von der Bildschirmgröße des Endgeräts können die Bereiche nebeneinander dargestellt oder durch Wechsel eines Bereichs ausgewählt werden.

Der Bereich **Datenauswahl**[33] enthält Funktionen zur Suche (s. Kernkomponente *Suchformular*). Diese umfasst die einfache Suche, Suche mit und in Facetten sowie die Expertensuche (SPARQL). Facetten (Anhang 3) sind ein Ansatz für das explorierende Browsen. Werte einer Facette können alphabetisch oder nach Häufigkeit sortiert werden. In den Facetten kann nach Werten gesucht werden. Ein oder mehrere Werte einer Facette oder eine Kombination von Werten mehrerer Facetten können zur Bildung von Teildatenmengen markiert oder ausgeschlossen werden. Das Ergebnis einer Suche ist eine Liste von Datensätzen oder eine interaktive Graph-Visualisierung. Zwischen beiden Ansichten kann gewechselt werden.

---

[31] Dieser Schritt erfolgt nur für Hauptdatenquellen, die automatisch abgerufen werden. Es besteht eine Präferenz für produktiv genutzte Formate der Hauptdatenquellen, um Effekte der volatilen technischen Entwicklung für den produktiven Dienst SoNAR zu reduzieren.

[32] Aufgrund der teilweise noch geringen Rechenleistung bei mobilen Endgeräten (Smartphones oder Tablets) müssen speziell bei der Graph-Visualisierung Leistungen eingeschränkt werden.

[33] Beispielkonzept für Datenauswahl mit explorativen Methoden: https://github.com/sonar-idh/visualization-prototypes/blob/main/img/prototype01.jpg

Der **Visualisierungsbereich** (s. Kernkomponente *Visualisierung*) bildet die Graphen visuell ab. Sie kann direkt auf Selektionen im Bereich Datenauswahl reagieren. Die Visualisierung unterstützt die Exploration des Datenbestands mit dem Ziel, einerseits Fragestellungen und Hypothesen zu entwickeln, und andererseits Daten zum Download für die Datenanalyse zu selektieren. Durch interaktive Schaltflächen können Knoten und Kanten hinzugefügt oder entfernt werden. Für das Hinzufügen schlägt das SoNAR-System Akteure vor, die mit dem Graphen assoziiert sind, aber aufgrund der Suchkriterien nicht berücksichtigt sind. Datenauswahl und Visualisierung sind so zwei Seiten einer Medaille: für die Recherche und Exploration der SoNAR-Daten.

Demgegenüber unterstützt der dritte Bereich, die **Dokumentation**, nicht die aktive Selektion von Daten, sondern informiert über die ausgewählte Datenmenge. Hierzu zählen:

- Liste der Output-Daten im Graphen mit einem Link zum anbietenden System[34]
- Maßzahlen der deskriptiven und Netzwerkstatistik (Häufigkeiten, Gatekeeper etc.)[35]

Die Liste der Output-Daten enthält stets den Zeitstempel der Datenintegration und den Link zum anbietenden System. Die Liste kann nach auszuwählenden Datenkategorien gruppiert werden. Wird ein Datensatz markiert, werden korrespondierende Kanten und Knoten im visualisierten Graphen hervorgehoben. Dasselbe gilt für Merkmalsausprägungen, die in der Ansicht Maßzahlen angeklickt werden, sodass Knoten und Kanten einfach im Grafen identifiziert werden können.

Für die Visualisierung der Netzwerkgraphen sind drei Konzepte von Bedeutung. Ihnen liegen die Visualisierungsstudien der Fachhochschule Potsdam zugrunde[36]:

Fächer für multimodale Beziehungen zwischen zwei Akteuren[37]

Zwischen zwei Akteuren können eine oder mehrere Formen sozialer Beziehung bestehen, z.B. familiäre und berufliche Beziehungen, Korrespondenzbeziehungen und Affiliationen oder aber allgemeinere Beziehungen wie: Jemand kennt wahrscheinlich einen anderen Akteur („knows of“, s. Anhang 3, Types of Relationships). Um mehrere Formen sozialer Beziehungen zwischen zwei Akteuren in einem Graphen abzubilden, kann die Kante zwischen zwei Akteuren durch Anklicken aufgefächert werden. Jeder Stab eines Fächers repräsentiert einen Beziehungstyp, der wiederum angeklickt werden kann. In der Dokumentation werden die korrespondierenden Output-Daten zu dem jeweiligen Beziehungstyp aufgelistet. Die Output-Daten sind mit den Repräsentationen der anbietenden Systeme durch einen persistenten Link verbunden.

Hervorhebung von Akteuren mit gemeinsamen Merkmalen

Durch die Auswahl eines oder mehrerer Werte oder Output-Daten in der Dokumentation werden Akteure und Kanten im Graphen hervorgehoben, sodass mit dieser Funktion Zusammenhänge zwischen Akteuren eines Graphens sichtbar werden. Dies sind bspw. ego-zentrierte Netzwerke, familiäre Beziehungen, das Wirken an einem Ort zu einem gemeinsamen Zeitpunkt oder für eine Körperschaft oder aber persönliche Merkmale wie Geschlecht, Sprache, Religion oder Herkunft.

Bildung von merkmalsbezogenen Clustern

[34] Liste der Output-Daten: https://github.com/sonar-idh/visualization-prototypes/blob/main/img/prototype04.jpg
[35] Beispiel Datenkategorie: https://github.com/sonar-idh/visualization-prototypes/blob/main/img/prototype03.jpg
[36] Visualisierungskonzepte der Fachhochschule Potsdam: https://github.com/sonar-idh/visualization-prototypes
[37] Beispiel Fächer: https://github.com/sonar-idh/visualization-prototypes/blob/main/img/17.jpg

Das SoNAR-System kann visualisierte Graphen nach Merkmalen zusammenfassen. Hierzu können ein oder mehrere Merkmale ausgewählt werden. So werden Akteure bspw. nach räumlichen und zeitlichen Werten (Geokoordinaten des Wirkungsorts, Wirkungsdaten), nach Beruf und Themen, mit denen sich Personen beschäftigt haben, gruppiert. Cluster sind eine ergänzende Methode zur Hervorhebung von Akteuren mit gemeinsamen Merkmalen: Bei der Hervorhebung werden Gemeinsamkeiten sichtbar, aber die Anordnung der Knoten und Kanten wird nicht beeinflusst. Bei Clustern werden Akteure dagegen im Graphen nach ihren Gemeinsamkeiten gruppiert.

Die Hervorhebung und das Clustern von Akteuren nach Merkmalen unterstützt die Selektion von Daten. Die Daten können für Analysen in nachnutzenden Systemen in verschiedenen Formaten mit technischen Metadaten (Provenienzdaten), in den Formaten RDF (JSON-LD, XML) oder CSV heruntergeladen werden. (s. Kernkomponente *Download-Button*). Die Output-Daten sind stets zur Nachprüfung im Datenspeicher abrufbar; auf Output-Daten, die durch eine Aktualisierung des SoNAR-Systems entfernt wurden, weist das SoNAR-System den Anwender hin.

Rechercheergebnisse können zudem zwischengespeichert werden, um die Recherche zu einem späteren Zeitpunkt fortzuführen. Über das Web-Frontend stehen zudem Informationen über das System, die Nutzungsmöglichkeiten sowie Tutorials zur Verfügung.

3.3 Implementierungsempfehlung

Das SoNAR-System wurde im Projekt SoNAR (IDH) prototypisch implementiert, um Erfahrungen über Datenprozesse zu sammeln, Anforderungen fachlicher Nutzung anhand von Fallbeispielen zu identifizieren sowie Visualisierungs- und Interfacedesignkonzepte zu erarbeiten[38]. Besonders relevant ist, dass durch das Datenvolumen der Performanz für Aufbereitung und Visualisierung eine zentrale Bedeutung zukommt und die Datenmenge auch nach einer Inbetriebnahme stetig zunehmen wird. Daher wird empfohlen, mit Beginn der Implementierungsphase sogenannte Durchstichimplementierungen mit den potenziellen Kandidaten, die im Folgenden genannt sind, durchzuführen. Eine Continous-Integration-Pipeline sollte neben den üblichen Komponenten- und Integrationstests automatisch auch die Performanz messen, um Herausforderungen bei der weiteren Entwicklung frühzeitig identifizieren zu können. Ein weiteres Kriterium zur Evaluierung bzw. Entscheidung für eine einzelne Komponente ist ihre Wartbarkeit nach der Inbetriebnahme. Hierzu zählt das Lizenzmodell, das in die Beurteilung etwaiger Limitierungen einfließen sollte. Die technischen und betrieblichen Anforderungen legen in der Tendenz den bevorzugten Einsatz von Open-Source-Lösungen nahe. Anhand der Projekterfahrungen können folgende Empfehlungen für Kandidaten für einzelne Komponenten und ihr Zusammenspiel abgeleitet werden (Abb. 3):

Catmandu und Metafacture sind geeignete Kandidaten für die Konfiguration und Durchführung von ETL-Prozessen. Catmandu ist ein CLI-Werkzeug (Command Line Interface). Es ist eine Open-Source-Lösung und ermöglicht die Konfiguration und Durchführung von ETL-Prozessen. Die SBB verfügt über Erfahrungen mit der Anwendung, da es u.a. bei der Zeitschriftendatenbank (ZDB) eingesetzt wird. Da Catmandu bereits einen OAI-PMH- und einen SRU-Client zur Aggregation von Daten umfasst und gängige Bibliotheksmetadaten wie MARC21 und MODS in RDF transformieren kann, erfüllt es viele Anforderungen an ein ETL-Tool für SoNAR und ist eine solide, Communtiy-basierte Lösung. Metfacture ist ein alternatives, in Java geschriebenes Open-Source-Tool für ETL-Prozesse. Es kann eigenständig als CLI-Werkzeug eingesetzt oder als Java-Bibliothek in Projekten

[38] https://sonar.fh-potsdam.de/prototype/; https://sonar:sonar2021@h2918680.stratoserver.net:7473/browser/

wie SoNAR eingebunden werden. Metafature ist modular und erlaubt flexible Konfigurationen für den optimalen Einsatz auch bei variierenden Anforderungen. Die Deutsche Nationalbibliothek (DNB) setzt Metafacture in ihrem Linked Data Service seit mehreren Jahren erfolgreich ein.

Blazegraph und Neo4j mit der „neosemantics"-Erweiterung sind als Graphdatenbank besonders geeignet. Beide unterstützen sowohl das RDF-Datenmodell als auch mit Gremlin und Cypher je eigene Graph traversal Abfrageprachen. Die Standardabfragesprache SPARQL wird allerdings nur von Blazegraph unterstützt. Blazegraph ist eine in Java geschriebene Open-Source-Lösung, die auf Performanz und Skalierbarkeit ausgelegt ist. Sie kann direkt in einer Anwendung eingebettet oder als eigenständiger Datenbankserver betrieben werden. Die Anwendung verfügt bereits über eine integrierte REST-Schnittstelle, um Daten durch nachnutzende Systeme abzufragen. Neo4j ist dagegen eine proprietäre Graphdatenbank, die als eigenständiger Datenbankserver betrieben wid und mit dem Object Graph Mapper eine gute Java-einbindung ermöglicht.

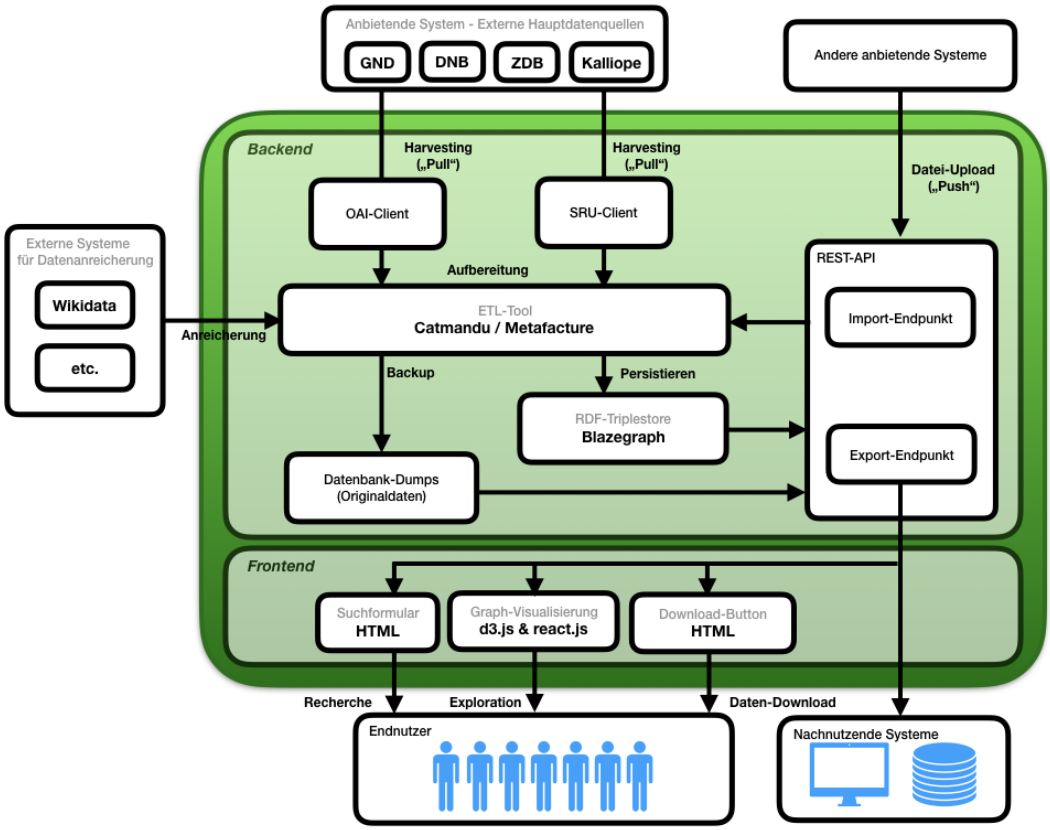

Abbildung 3: Implementierungsempfehlung für Kernkomponenten

D3.js und WebGL sind Kandidaten zur Implementierung der Graph-Visualisierungen. D3.js ist eine JavaScript-Bibliothek für dynamische Visualisierungen. Sie unterstützt Vektorgraphiken und eignet sich für die Darstellung und Speicherung von Graph-Daten im Web-Frontend von SoNAR. WebGL ist im Unterschied zu D3.js eine JavaScript-API für die native hardwarebeschleunigte Darstellung von interaktiven Graphiken in einem Webbrowser. Sie bietet anders als D3.js keine Abstraktionsschicht für Datenvisualisierungen, hat aber Vorteile hinsichtlich der Performanz und Individualisierbarkeit (Bludau/Dörk 2021, 5 und 38).

react.js und Vue eigenen sich gleichermaßen für die Implementierung des Web-Frontend. Beides sind gängige JavaScript-Frameworks für Web-Anwendungen. Zusammen mit modernen Web-Technologien wie HTML5 und CSS3 können sie Grundlage zur Entwicklung eines modernen UI für SoNAR sein. Bedeutender als das Framework ist jedoch die Architektur: infrage kommen sowohl die Implementierung einer Single-Page-Applikation, die nur im Browser ausgeführt werden kann, als auch eine hybride Architektur mit server-seitigem Pre-rendering. Sinnvolle Kandidaten für die Implementierung von letzterem sind node.js oder Java Springe Framework.

Erst durch eine Durchstichimplementierung zum Beginn eines Implementierungsprojekts können Kandidaten und Architektur für eine Implementierung konkretisiert werden.

4. Ausblick

Mit der Implementierung von SoNAR wird eine infrastrukturelle Lücke der Forschung adressieren und ein eindeutig identifizierter Bedarf bedient. Durch das Erprobungsprojekt SoNAR (IDH) liegen fundierte Erkenntnisse zum Umfeld und zu den Anforderungen an SoNAR vor. Der ganzheitliche Ansatz zur Erprobung ließ eine substanzielle, am Bedarf orientierte Abgrenzung von Leistungen einer Forschungstechnologie SoNAR in Abgrenzung vor allem zu den nachnutzenden Systemen für wissenschaftliche Datenanalysen zu. Durch die Schwerpunktsetzung auf die Integration und die Aufbereitung von Daten, die im Zusammenhang mit bzw. auf der Grundlage von Quellen von Bibliotheken, Archiven und Museen erhoben werden, kann ein für die HNR spezialisiertes und anknüpfungsfähiges Angebot entstehen: indem umfangreich Datenbestände integriert, Daten zu Akteuren und ihren Beziehungen extrahiert und über leistungsstarke Schnittstellen bereitgestellt werden. Die Anbindung an Datenrepositorien von Kultureinrichtungen führt dazu, dass Aussagen über Akteure immer auf einen Datenbestand und immer auf den Kontext einer beschriebenen Quelle zurückgeführt werden können. Der Linked-Data-Ansatz für eine HNR-Datenmodellierung ist entscheidend, um heterogene Datenquellen mit vertretbarem Aufwand integrieren und auch standardbasiert abfragen zu können. SoNAR wird sukzessiv das bekannte, dezentrale Wissen von Kultureinrichtungen über Akteure zusammenführen. Verbunddatenbanken tragen erheblich als Hauptdatenquellen zu einem HNR-Kerndatenbestand bei. Einzelne digitale Repositorien können den Kernbestand substanziell erweitern. Die FAIR-Prinzipien sind als Leitbild (und -planken) für SoNAR entscheidend, um wissenschaftlichen Anforderungen an Transparenz zu genügen. Durch das Erprobungsprojekt wurde deutlich, dass FAIR nicht nur die Forschungsdaten, sondern auch Provenienzdaten, das Datenmodell sowie die Modelle zur Transformation und Aufbereitung der Input-Daten im ETL-Prozess einschließt.

Auf der Basis dieses Implementierungs- und Betriebskonzepts ist vorgesehen, vorbehaltlich der Ergebnisse der Begutachtung im Programm e-Research-Technologien die Implementierung zu beantragen. Die Projektierung für einen Antrag einer Implementierungsphase wird u.a. folgende Schritte beinhalten:

Erstens. Die in einer ersten Implementierung zu berücksichtigenden Datenquellen sind mit den Stakeholdern abzustimmen. Dies schließt u.a. eine Erörterung des Zugangs zu bibliographischen Metadaten über den Dienst Culturegraph mit der Deutschen Nationalbibliothek oder alternativ der Zugang zu den bibliographischen Metadaten des Gemeinsamen Bibliotheksverbunds ein.

Zweitens. Ebenfalls ist in diesem Zusammenhang die produktive Nutzung von Entity Facts resp. Lobid.Org mit der Deutschen Nationalbibliotheken und dem Hochschulbibliothekszentrum in NRW als eine Quelle für ID-Konkordanzen abzustimmen.

Drittens. Mit der Kooperative „Social Network and Archival Context" ist die Einbindung sowohl der Daten als auch der fachlichen Entwicklung der Datenmodellierung zu erörtern. SNAC ist eine bedeutende Säule für die internationale Ausrichtung von SoNAR und der breiten Verankerung.

Viertens. Die Projektierung muss den Aspekt der Performanz sowohl für die Datenaufbereitung als auch den Zugang über die beiden alternativen Schnittstellen (API, UI) berücksichtigen. Dies schließt die Ermittlung der Kosten für die IT-Infrastruktur für die Implementierungsphase und die spätere Inbetriebnahme ein.

Das vorliegende Konzept geht von einer Implementierung und Inbetriebnahme von SoNAR durch die SBB aus. Im Rahmen der Projektierung wird dieser Sachverhalt geprüft. Gedanklich kann eine Ansiedlung bei der Deutschen Digitalen Bibliothek vorgestellt werden, wobei Einschränkungen in der nationalen Orientierung dieses Angebots liegen. Auch Bibliotheksverbünde sind vorstellbar. Für eine Ansiedlung bei der SBB-PK spricht die Ansiedlung oder aber bedeutende Beiträge zu den zentralen Verbunddatenbanken, die zugleich Datenquelle sind (ZDB, KPE, GND). Dies schließt die internationale Vernetzung etwa mit SNAC ein. Die SBB verfügt in den Abteilungen Überregionale Bibliographische Dienste und Informations- und Datenmanagement über die Erfahrungen, um die technischen, organisatorischen und fachlichen Aspekte von SoNAR abzudecken.

Mit der Implementierung wird, ausgehend vom Konsortium, der Aufbau eines Netzwerks von Stakeholdern (Service- und Datenanbietern sowie wissenschaftlicher Anwendergemeinschaften) angestrebt. Durch SoNAR kann die HNR und angrenzende Disziplinen nachhaltig infrastrukturell gestärkt werden. Hierzu trägt bei, dass die Datenbasis auch nach einer Implementierung stetig erweitert werden kann. Indem die Zugangswege zu den Daten niedrigschwellig gestaltet werden, können einzelne resp. kleinere Forschungsprojekte auf HNR-Methoden zurückgreifen, sodass der wissenschaftliche Diskurs und in dessen Folge der Erkenntnisgewinn bedeutend verbreitert wird.

Literatur

Ahnert, Ruth/ Ahnert, Sebastian E./ Coleman, Catherine Nicole/ Weingart, Scott: The Network Turn. Changing Perspectives in the Humanities. Cambridge, 2020

Allemang, Dean/ Hendler, Jim: Semantic Web for the Working Ontologists. Effective Modeling in RDFS and OWL. Amsterdam u.a., 2011

Alvarez Francés, Leonor/ van der Heuvel, Charles: Mapping Notes and Nodes in Networks. Exploring potential relationships in biographical data and cultural networks in the creative industry in Amsterdam and Rome in the early modern period. External research report (2014), http://mnn.nodegoat.net

Balck, Sandra/ Menzel, Sina/ Petras, Vivien: SoNAR (IDH). AP4-4 Evaluierung III: Analyse des Forschungsprozesses von HNA-Expert:innen und sich daraus ergebende Bedürfnisse an eine Infrastrukturlösung. Humboldt-Universität zu Berlin. Version 2.0. https://github.com/sonar-idh/reports/blob/main/AP4-HU-4-4-2-Evaluierung-III.pdf, 2021

Bludau, Mark-Jan/Dörk, Marian: Wissenschaftliches Konzept für die Visualisierung von und Interaktion mit Graphen und Projektdokumentation. https://github.com/sonar-idh/reports/blob/main/AP3-FHP-Projektdokumentation.pdf, 2021

Carius, Hendrikje: Europäische Gelehrtennetzwerke digital rekonstruieren. Vernetzung von Briefmetadaten mit Early Modern Letters Online (EMLO). In: Bibliotheksdienst. 55 (2021), 1. 29-41. https://doi.org/10.1515/bd-2021-0008

Düring, Marten/ Eumann, Ulrich/ Stark, Martin/ Keyserlingk, Linda (Hg.): Handbuch Historische Netzwerkforschung. Grundlagen und Anwendungen. Berlin, 2016

Düring, Marten/ Keyserlingk, Linda: Netzwerkanalyse in den Geschichtswissenschaften. Historische Netzwerkanalyse als Methode für die Erforschung historischer Prozesse. In: Jordan, Stefan/ Schützeichel, Rainer (Hg.): Prozesse. Formen, Dynamiken, Erklärungen, Wiesbaden, 2015. 337-350

EGAD (Expert Group Archival Description) / ICA: Records in Contexts. Conceptual Model. Consultation Draft 0.2 https://www.ica.org/en/records-in-contexts-conceptual-model, 2021

EGAD (Expert Group Archival Description) / ICA: RiC-O projects and tools. 2021 https://ica-egad.github.io/RiC-O/projects-and-tools.html

Fangerau, Heiner et al.: SoNAR AP2. https://github.com/sonar-idh/reports/blob/main/AP2-UDK-Projektdokumentation.pdf, 2021

Gramsch-Stehfest, Robert: Von der Metapher zur Methode. Netzwerkanalyse als Instrument zur Erforschung vormoderner Gesellschaften. In: Zeitschrift für Historische Forschung. 47 (2020), 1-39

Kerschbaumer, Florian/ Keyserlingk-Rehbein, Linda/ Stark, Martin/ Düring, Marten (Hg.): The Power of Networks. Prospects of Historical Network Research. New York, 2020

Lemercier, Claire: Formale Methoden der Netzwerkanalyse in den Geschichtswissenschaften: Warum und Wie? In: Österreichische Zeitschrift für Geschichtswissenschaft. 23 (2012), 1. 16-41

Menzel, Sina et al.: Named Entity Linking mit Wikidata und GND. Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten. In: Franke-Maier, Michael et al. (Hg.): Qualität in der Inhaltserschließung. Bibliotheks- und Informationspraxis. 70 (2021). 229 – 257

Rehbein, Malte: Historical Network Research, Digital History and Digital Humanities. In: Kerschbaumer, Florian/ Keyserlingk-Rehbein, Linda/ Stark, Martin/ Düring, Marten (Hg.): The Power of Networks. Prospects of Historical Network Research. New York, 2020. 253-279

Schnaitter, Hannes et al.: SoNAR (IDH). AP4-5 Evaluierung IV. Nutzer:inenstudie. Version 15.11.2021. https://github.com/sonar-idh/reports/blob/main/AP4-HU-4-5-3-Evaluierung-IV.pdf, 2021

Anhang

A1 Bedarfs- und Umfeldanalyse

s. Dokument SoNAR-2021-A1-Bedarf_Umfeld.docx

A2 Systembeschreibung

s. Tabelle SoNAR-2021-A2-Systembeschreibung.xlsx

A3 Datenmodellskizze

s. Dokument SoNAR-2021-A4-Datenmodellierung.docx