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 das Ergebnis des Projekts “Interfaces to Data for Historical Social Network Analysis and Research, SoNAR (IDH)”. Es prüfte die Eignung der Daten von Kultureinrichtungen in Verbindung mit Fachanforderungen sowie Ansätzen zur Aufbereitung und Bereitstellung der Daten für Historische Netzwerkanalysen.

Die Staatsbibliothek zu Berlin – Preußischer Kulturbesitz entwickelte, aufbauend auf den Teilergebnissen der Projektpartner, das vorliegende Konzept. Dessen Grundlagen sind:

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

Das Deutsche Forschungszentrum für Künstliche Intelligenz, Abteilung Sprachtechnologie (DFKI), Prof. Dr. Georg Rehm, erarbeitete eine modellhafte Datenprozesskette als Teil praxisorientierter Tests und Evaluationen (Forschungs-, Visualisierungs- und Nutzerstudien). Als Projektionsfläche flossen die Erfahrungen in die Systembeschreibung (Komponenten, Prozesse und Funktionen der Forschungstechnologie), die mit diesem Konzept erarbeitet wurden, ein.

Die Firma effective WEBWORK GmbH begleitete die Konzeptentwicklung. Sie erarbeitete die Systembeschreibung durch Anforderungsanalyse, Ermittlung der Systemprozesse und Abstimmung des Implementierungsansatzes für einen nachhaltigen Betrieb.

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

Berlin, Oktober 2021

Executive Summary

[…]

Inhalt

1. Einführung

SoNAR, Social Network Analysis and related Research, ist eine innovative datenzentrierte, international ausgerichtete Forschungstechnologie. Sie enthält den umfangreichsten und kontinuierlich erweiterten Datenbestand über historische Netzwerke. Ausgangspunkt sind heterogene Datenbestände von Kultureinrichtungen wie Bibliotheken und Archive. Deren Daten integriert SoNAR durch Einsatz etablierter, offener Technologien. Der Datenbestand ist ebenso offen, das heißt ohne geographische, zeitliche oder thematische Beschränkungen für vielfältige Forschungsfelder. Der Datenumfang ist somit nur begrenzt von überlieferten Quellen sowie dem Stand und der Qualität ihrer Erschließung und Digitalisierung. Unterstützende digitale Dienste öffnen Wissenschaftlern den Zugang zu den Daten über moderne Schnittstellen: sowohl für nachnutzende Forschungsumgebungen als auch für eine visuell aufbereitete Erkundung.

So könnte SoNAR mit der Inbetriebnahme beworben werden. Basis für den Aufbau ist das vorliegende Konzept. Es reagiert auf ein Desiderat, einen die Wissenschaft behindernden Faktor: der Zugang zu Daten. Fachwissenschaftliche Test- und Nutzerstudien des Projekts SoNAR (IDH) zeigen überzeugend, dass Kultureinrichtungen umfangreich Daten mit Aussagen über soziale Beziehungen im Zusammenhang mit den aufbewahrten Ressourcen erzeugen, und dass diese zur Analyse historischer Netzwerke bedeutend beitragen können (vgl. HHU 2021, ###; IBI 2021, ###).

Eine Säule des Konzepts ist die Bedarfs- und Umfeldanalyse (Kapitel 2) über die Historische Netzwerkforschung: ihre Genese, Aktivitäten, Anwender, Tools und Projekte. Durch sie wurde SoNAR nun konkretisiert: als Datenangebot für Netzwerkanalysen mit ergänzenden Diensten für

- die Exploration des Datenbestands
- die Entwicklung von Fragestellungen
- die Analyse in Forschungsumgebungen
- das Nachvollziehen von Forschungsergebnissen

Diese Aktivitäten einzelner Phasen wissenschaftlicher Untersuchungen wird SoNAR für die Vorbereitung von Untersuchungen durch standardisierbare Leistungen unterstützen. Die Kernkomponenten, -prozesse und -funktionen, das ist die zweite Säule des Konzepts, sind auf Standards und Offenheit für einen nachhaltigen Betrieb (Kapitel 3) ausgerichtet, um die Pflege und Erweiterung auch nach einer Implementierung zu gewährleisten. Im Kapitel 4 sind Aufbau, Betrieb und Verankerung von SoNAR in Kultur- und Forschungs-Communities erläutert.

Um Grundsätze des wissenschaftlichen Arbeitens zu stärken, werden die FAIR-Prinzipien[1] leitend für die Implementierung sein. Die Daten werden forschungsorientiert mit der Creative Commons Lizenz CC BY 4.0[2] ausgezeichnet. Das Datenmodell für Historische Netzwerke wird Public Domain. Metadaten über Datenprovenienz und -verarbeitung sowie die Möglichkeit, auf alte Datenstände zurückgreifen zu können, tragen zum Vertrauen in Forschungsprozesse und -ergebnisse bei.

SoNAR wird eine neue Brücke zwischen Kultureinrichtungen und quellenorientierter Forschung; sie verkürzt Wege durch standardisierbare Leistungen und eröffnet eine breite kontextsensitive Perspektive auf vergangenen Ereignisse.

[1] FORCE11: The FAIR data principles. https://www.force11.org/group/fairgroup/fairprinciples, 2016 (2021-09-07)
[2] Creative Commons Attribution 4.0 international: https://creativecommons.org/licenses/by/4.0 (2021-09-07)

2. Forschungskontext

2.1 Ausgangssituation

Die Historische Netzwerkanalyse (HNA) 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 HNA ist daher auf die Ressourcen von Bibliotheken, Archiven und Museen angewiesen. Trotzdem, oder gerade deswegen, ist die HNA heute eine akzeptierte Methode in der Geschichtswissenschaft und mit breiten Themen, Fragestellungen und wichtigen Beiträgen zur Methodenentwicklung vertreten (vgl. Rehbein 2020, 256/Ahnert et al. 2020). Die Anwendung der HNA in der Forschung nahm in den letzten zwei Dekaden stetig zu. Dennoch bleibt sie im Vergleich zu klassischen Methoden der historischen Forschung eine Nische (vgl. Rehbein 2020, 259). Dies hat auch forschungspraktische Ursachen: die Datenerhebung. Im Projektantrag für SoNAR (IDH) wurde bereits konstatiert, dass Aufwände zur Sichtung der vielen, dezentral überlieferten Quellen oft prohibitiv sind. Robert Gramsch-Stehfest konstatiert noch 2020: "Zwar gibt es auch 'kleine Formen' der […] Netzwerkforschung, und gerade 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 dennoch aus einem methodischen Ansatz der Geschichtswissenschaft (vgl. Reitmayer/ Marx 2010, Düring/ von Keyserlingk 2015) ein integraler Bestandteil von Digital Humanities, Digital History und historischer Informationswissenschaft (vgl. Rehbein 2020, 277)[3]. Den Beitrag der HNA zur Forschung heben ebenso Wissenschaftler hervor, die im Projekt SoNAR (IDH) interviewt wurden (vgl. Balck 2021, ###).

Augenfällig ist die Diversität; die HNA 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 wie die Historical Network Research Conference[4]. Dies wird bspw. drittmittelfinanziert an europäischen Universitäten ausgerichtet (2013 Hamburg, 2014 Ghent, 2015 Lissabon, 2017 Turku, Brno 2018). Ihre Schwerpunkte sind Methoden, Fragestellungen und Quellen. Das Forschungsfeld gewinnt durch Interdisziplinarität, die sich in der breiten Teilnahme von Vertretern diverser Disziplinen auf Konferenzen (Abbildung 1[5]) und in Forschungsprojekten zeigt. Auch die Teilnehmer von HNA-Workshop[6] und Interviews im Projekt SoNAR (IDH) hatten verschiedene fachliche Hintergründe (vgl. HHU 2021, ###, Balck 2021, ###), sodass die jeweiligen Perspektiven der verschiedenen Disziplinen direkt in die Entwicklung dieses Implementierungs- und Betriebskonzept einfließen konnten.

[3] vgl. auch das Missionstatement der AG Graphen & Netzwerke des Verbands Digital Humanities zu Graphen und Netzwerke, https://graphentechnologien.hypotheses.org/ueber-das-blog (2021-09-06)
[4] https://historicalnetworkresearch.org/hnr-events/hnr-conferences/ (2021-09-06)
[5] Von 42 Referenten der HNR+ResHis Conference 2021 wurden die erstgenannten in der Verteilung berücksichtigt.
[6] Vertreten waren die Fächer: Geschichtswissenschaft, Archiv- und Bibliothekswissenschaft, Sozialwissenschaft, Informatik, Literaturwissenschaft.

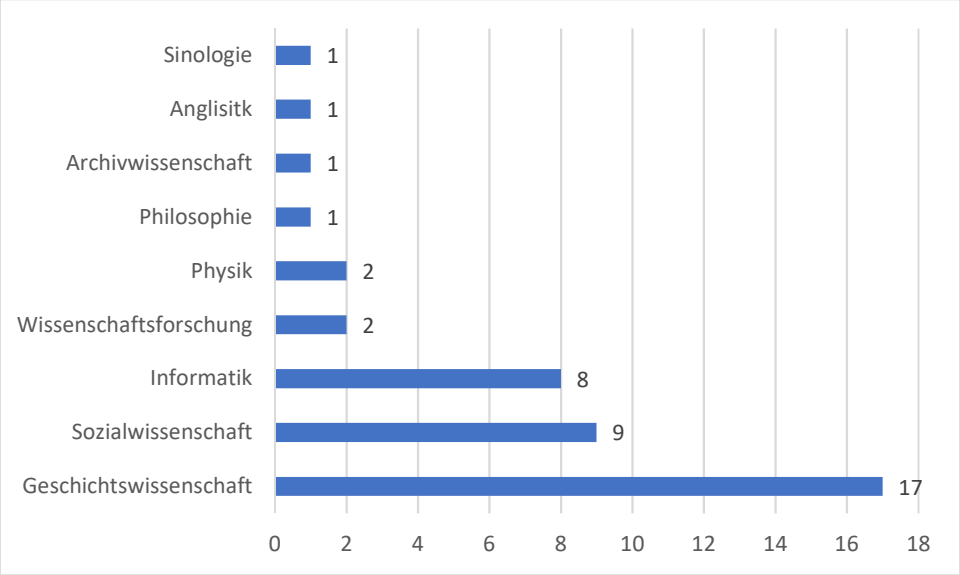

Abbildung 1: Fachlicher Hintergrund Vortragender auf der HNR+ResHist Conference, Juni 2021

Zur Verbreitung der HNA trägt dessen Open-Science-Kultur bei. Zu den Aktivitäten zählen die Gründung von Open Access Journals wie „Journal of Historical Network Research“, die kollaborative Pflege einer HNA-Bibliografie[7], (Micro)-Blogs[8] und Mailinglisten[9] sowie neue Modelle dynamischen Publizierens wie kollaboratives Schreiben (vgl. Rehbein 2020, 264). Ein regelmäßiger Programmteil von HNA-Konferenzen ist die Vorstellung und Bewertung von Forschungstools. Eine systematische Desktop-Recherche hat gezeigt, dass aktuell mindestens 28 Tools der HNA-Forschung zur Verfügung stehen (Anhang 1, Tab. 1). Sie können unterteilt werden nach Schwerpunkt: 1) Analyse und Visualisierung, 2) nur Analyse oder 3) nur Visualisierung. Viele sind kostenfrei und plattformunabhängig (Anhang 1, Tab. 2, 3, 4). Sie unterstützen den Import und Export in gängigen Formaten, z.B. CVS, JSON. Mit Ausnahme von SplitsTree können Daten als Graphen visualisiert und als Grafiken lokal gespeichert werden. Analysetools bieten vielfältige Funktionen für statistische Berechnungen. Workshops und Tutorials, bspw. auf der Website „The Programming Historian“[10], tragen zur ihrer verbreiteten Anwendung und Qualifizierung bei.

Um ein Bild über die Verbreitung von Tools zu gewinnen, wurden alle 26 Beiträge des „Journal of Historical Network Research“ (2018-2020) ausgewertet. In 23 Beiträgen wird der Einsatz von Tools erwähnt, aber diese nur in 18 Beiträgen explizit genannt. Gephi (8) und Visone (4) dominieren die Verteilung. Genannt wurden auch: Vennmaker (2), Node XL (1), Nodegoat (1), Pajek (1), Palladio (1), Cytoscape (1) (Anhang 1, Tab. 5). Die Auswertung von Monografien und Beiträgen in Sammelbänden, basierend auf der HNR-Bibliography Vol 7. (2018-2021), bestätigt den Trend: Gephi gilt als das „bekannteste und vielseitigste“ Tool (Düring et al. 2016, 175). Das „Handbuch Historische Netzwerkforschung“ empfiehlt für Datenanalysen ebenfalls: Nodegoat, NodeXL, Palladio und VennMaker sowie Pajek und UCInet (vgl. ebd., 177).

Die Auswertung von 36 Projekten (Anhang 1, Tab. 6) hat gezeigt, dass sie analog zu den Tools nach ihrem Schwerpunkt – Analyse (Anhang 1, Tab. 7) oder Visualisierung sozialer Netzwerke –

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

gruppiert werden können (Anhang 1, Tab. 8). Beiden ist jedoch die Datengrundlage gemein: Archivalien, z.B. Briefe aus Nachlässen, oder Texte, z.B. Zeitungen und Zeitschriften, Biografien etc.: Das Projekt „Kindred Britain“ nutzt Biografien, „Tudor Networks“ greift auf Briefe zurück und „Mapping Notes and Nodes“ integriert heterogene (Meta-) Datenreihen (vgl. Alvarez/ van der Heuvel, 2014). Einzelne Projekte wie „Hidden Perspectives“ extrahieren Ereignisse, Personen und Körperschaften aus digitalen Archiven. Die Mehrzahl der ausgewerteten Projekte bieten ein webbasiertes User Interface an, um die zusammengetragenen Daten zu explorieren, doch nur sieben bieten sie zum Download an (Anhang 1, Tab. 7, Tab. 8). Einen intelligenten Ansatz 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 ergänzte lediglich fehlende Daten durch eigene Datenerhebungen, ebenfalls direkt in Wikidata.

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, Humboldt-Universität zu Berlin, zusammengebracht werden:

> „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 aggregierten, 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 et al. 2021, 23)

Gerade die zweite Gruppe greift eher auf generische Forschungsumgebungen wie R oder Jupyter Notebooks zurück (vgl. HHU 2021, ###, HU 2021, ###; FHP 2021, ###). Die Nutzung von Tools ist aber nicht nur von Funktionen, sondern von Usability, Support und Performance beeinflusst (vgl. Balck et al. 2021, 18). Es konnten zudem Aussagen vernommen werden, dass in der jüngeren Generation die Bereitschaft besteht, Data Science Methoden und respektive Tools für explorativ geprägte Untersuchungen zu vertiefen (###).

Neben der Verfügbarkeit von Daten ist die Dokumentation von Datenquellen, Arbeitsschritten und Tools ein weiteres Desiderat der Forschungspraxis (vgl. Balck et al. 2021, 23 ff.). Die Transparenz von Datenanalysen ist jedoch die zwingende Voraussetzung, um Akzeptanz für die HNA und ihre Ergebnisse im breiteren Wissenschaftsdiskurs zu gewinnen (vgl. HHU 2021, ###).

2.2 Zielsetzung

Die Ausgangssituation zeichnet das Bild einer dynamischen, expandierenden Entwicklung der Historischen Netzwerkanalyse. Sie bereichert fachwissenschaftliche Untersuchungen und gilt fachübergreifend als fundierte Methode mit dem Potenzial, bedeutend zur Theoriebildung beizutragen. Die Ausgangssituation hat aber ein kritisches infrastrukturelles Defizit gezeigt: den Datenzugang. Das Defizit hat drei Dimensionen: Datenerhebung, Nachnutzung erhobener Daten, Dokumentation von Datenprovenienz und -verarbeitung.

Das Projekt SoNAR (IDH) konnte zunächst drei wesentliche Sachverhalte bestätigen (vgl. ###):

1) Kultureinrichtungen erschließen intellektuell oder zunehmend auch maschinell Ressourcen. Die Daten enthalten explizite und implizite Aussagen über Akteure.
2) Diese teils heterogenen Datenbestände können integriert und die Aussagen über Akteure auf ein HNA-Datenmodell abgebildet werden.
3) Die Verfügbarkeit von Daten für ein einzelnes Forschungsthema ist abhängig von überlieferten Ressourcen sowie dem Stand von Erschließung und Digitalisierung.

Die Erkenntnisse aus dem Projekt SoNAR (IDH) sind, von diesen Bedingungen ausgehend, in drei Ziele geflossen, die das Profil von SoNAR und so die Systembeschreibung (Kapitel 3) bestimmen:

Ziel 1: Historische Netzwerkdaten extrahieren

SoNAR wird strukturierte Daten von Kultureinrichtungen, die diese auf Basis ihrer Ressourcen erzeugen, maschinell aggregieren, aufbereiten und auf ein HNA-Datenmodell abbilden, um eindeutige Daten über Akteure online anzubieten. Mit der fortschreitenden Digitalisierung und Erschließung von Ressourcen in Kultureinrichtungen, wird SoNAR regelmäßig den Datenbestand und, bedarfsabhängig, auch das Datenmodell aktualisieren.

Ziel 2: Replikations- und Sekundärstudien fördern

SoNAR wird Eingangs- und Ausgangsdaten versioniert archivieren[11] sowie Änderungen von Transformations- und Inferenzregeln mit jeder Aktualisierung sichern und technische Metadaten zu den Ausgangsdaten mit Zeitpunkt des Datenbezugs (Download) anbieten, um transparent die Datenprozesskette zu dokumentieren und z.B. so Replikationsstudien zu unterstützen. Dieselben Daten einer Studie können so für Sekundärstudien (veränderte Fragestellung) genutzt werden.

Ziel 3: Historische Netzwerkdaten erkunden

SoNAR wird eine webbasierte Nutzerschnittstelle anbieten, die Recherche und Exploration in den Ausgangsdaten ermöglicht, um das Vorbereiten von Studien mit niedrigschwelligem Aufwand zu optimieren, z.B. Prüfen der Verfügbarkeit von Daten oder Formulieren von Fragestellungen. Recherche- und Explorationsoptionen sowie Schnittstellen und Datenformate für nachnutzende Systeme wie Forschungsumgebungen werden dokumentiert.

Zwei Themen, die im Projektverlauf betrachtet wurden, wurden als Ziele, die externe Systeme bedienen, identifiziert. Sie flossen in Schnittstellenanforderungen der Systembeschreibung ein:

- Die maschinelle Erkennung und Verlinkung von Entitäten in (OCR-erzeugten historischen Zeitungs-) Texten ist fortgeschritten (vgl. Menzel et al. 2021). Digitale Repositorien von Kultureinrichtungen können strukturierte Daten, die intellektuell oder maschinell aus Texten extrahiert werden, auch an SoNAR übertragen. SoNAR wird den Fokus auf die Aufbereitung strukturierter Daten für einen HNA-Datenbestand legen.
- Die Forschung auch im Zusammenhang mit historischen Netzwerken unterliegt einem dynamischen fachwissenschaftlichen Diskurs. Anforderungen an die Datenanalysen mit statistischen oder visuellen Methoden werden von Forschungsgemeinschaften aktiv aufgenommen und fließen in generische oder spezialisierte Tools. SoNAR wird den Fokus auf Schnittstellen mit innovativen Diensten für Recherche und Exploration legen.

Das generelle Ziel hinter SoNAR ist die automatisierte, transparente Aufbereitung strukturierter Daten von Kultureinrichtungen, um Hürden für Historische Netzwerkanalysen zu reduzieren.

[11] Eingangsdaten sind original aggregierte, Ausgangsdaten die nach der Datenprozesskette bereitgestellten Daten.

2.3 Innovation

SoNAR knüpft an Dateninfrastrukturen von Kultureinrichtungen wie Verbünde und digitale Repositorien an. Der Ansatz von SoNAR ist die duale Nutzung von Daten: ihr primärer Nutzen für den Zugang zu Ressourcen sowie, sekundär, als Forschungsdaten für die HNA. Dass die Daten der GND und von Archivbeständen etwa des Kalliope-Verbunds besonders geeignet sind, belegen die Forschungstests des Projekts SoNAR (IDH) (vgl. HHU 2021, ###). Dies gilt, etwas allgemeiner, für alle strukturierten Daten, die Kultureinrichtungen durch formale sowie inhaltliche Erschließung, durch maschinelle Methoden oder Kooperation z.B. mit Editionsprojekten erzeugen. Die Grenze liegt weniger in den Potenzialen, sondern vielmehr im Umfang und der Tiefe der Erschließung.

Der innovative Charakter von SoNAR ist die duale Nutzung der Daten von Kultureinrichtungen, um entlang wissenschaftlicher Anforderungen ein Desiderat zu adressieren. Alle Maßnahmen orientieren sich an den FAIR-Prinzipien. Sie sind das Leitbild für:

Datenaufbereitung

- Der Datenbestand wird maschinell regelmäßig aktualisiert und die Daten sind persistent adressierbar.
- Aussagen über Akteure - ihre Merkmale und sozialen Beziehungen - können stets persistent auf Eingangsdaten und so auf ursprüngliche Quellen zurückgeführt werden.
- Heterogenität (Identnummern von Entitäten diverser Normdateien, Formate) wird in explizite, einfach nachzuvollziehende Aussagen über Akteure transformiert.

Datenmodellierung

- Das HNA-Datenmodell ist neben offenen Formaten ein Teil der Strategie zur Interoperabilität. Es adaptiert etablierte Ontologien (Anhang 3), wird unter der Public Domain Lizenz veröffentlicht und kooperativ gepflegt (Community-Building).
- Transformations- und Inferenzregeln bilden die domainspezifische Fachkenntnis über die Eingangsdaten ab; so werden erhebliche Analyseaufwände für Dritte reduziert.
- Es werden ausschließlich frei zugängliche Standardvokabularien für Eingangsdaten und die Aufbereitung der Eingangsdaten eingesetzt.

Datenschnittstellen

- Die Übertragung von Daten durch anbietende Systeme und für nachnutzende Systeme erfolgt über Standardprotokolle und -formate.
- Für die Nutzerschnittstelle mit visuellen Methoden zur Recherche und Erkundung des Datenbestands werden ausschließlich offene Webstandards verwendet.

Datenreproduktion

- Eingangs- und Ausgangsdaten werden in Archivdateien gesichert und online, z.B. für die Überprüfung von Forschungsergebnissen, frei zur Verfügung gestellt.
- Transformations- und Inferenzregeln werden versioniert und gesichert, die technischen Metadaten enthalten Provenienz- und Verarbeitungsinformationen.

Durch Format- und Schnittstellenstandards, einer forschungsorientierten Datenlizensierung (CC-BY 4.0) und Open-Source-Lösungen wird SoNAR die nötige Transparenz für Forschungsprozesse unterstützen. Durch RDF für Integration und Bereitstellung der Daten werden die Erweiterbarkeit und Pflege des Datenmodells und des Datenbestands langfristig nachhaltig gewährleistet.

3. Implementierung

3.1 Kernkomponenten

Die Implementierung von SoNAR wird drei Prinzipien, die aus der Bedarfs- und Umfeldanalyse hervorgehen, beachten:

- Dokumentation der Datenprozesskette
- Offenheit der Formate und Schnittstellen
- Ausrichtung auf Open-Source-Lösungen

Die SoNAR-Architektur beruht auf dem Microservice-Konzept, sodass Entwicklung, Pflege und Austausch der Komponenten bedarfsorientiert erfolgen kann. Durch die Entscheidung für RDF wird die Integration von Daten auch nach der Implementierung von SoNAR gewährleistet sein. Durch diese Vorfestlegungen wird ein nachhaltiger Betrieb und Anschlussfähigkeit an anbietende und nachnutzende Systeme optimal gesichert.

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

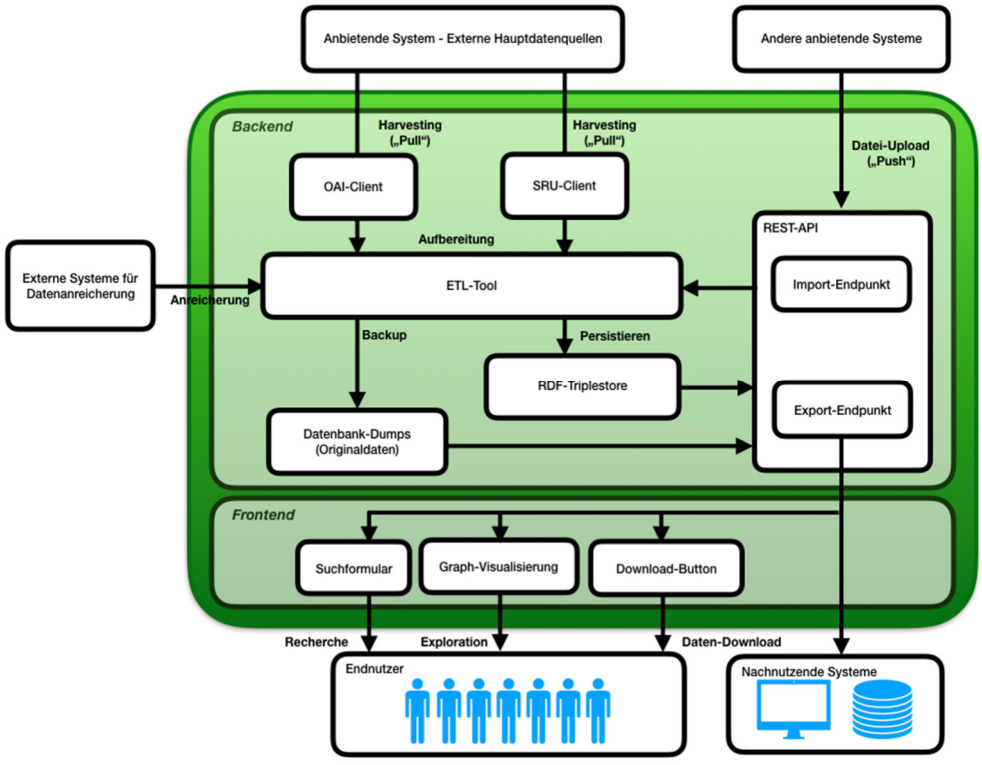

Abbildung 2: Schematische Darstellung der Kernkomponenten und -prozesse

Das Backend stellt vor allem eine Datenbank zur Verfügung, in der die gesammelten Daten gesichert, transformiert und für den Zweck der HNA aufbereitet werden, während das Frontend eine Website mit einem User Interface für die Interaktion der Forschenden 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[12]. Die Kommunikation erfolgt zwischen Datenlieferanten (Data Provider) und Dienstanbieter (Service Provider) automatisiert über einen OAI-Harvester. 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[13]. Die mit der Abfrage identifizierte Teildatenmenge kann in die eigene Umgebung übernommen werden. SoNAR als Service Provider nutzt SRU als zweite optionale Methode neben OAI, um Daten von Datenlieferanten zu übernehmen.
- **Datenbank-Dumps / Datenspeicher:** … ist der Ablageort für Eingangsdaten inklusive technischer Metadaten. Sie werden nach Liefer- bzw. Harvesting-Datum als komprimierte Archivdateien aufbewahrt. Die Erstellung von Archivdateien ist Teil des ETL-Prozesses (s.u.) und Voraussetzung für die Reproduktion eines ETL-Prozesses zur Unterstützung von Forschungsprozessen.
- **ETL-Tool:** Für die prototypische Demonstration erfolgte die Aufbereitung XML-basierter Eingangsdaten mit Python-Skripten[14]. Für die produktive Implementierung werden für den Aufbereitungsprozess etablierte Tools für „Extract, Transfer, Load"-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.
- **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 der Benutzungsschnittstelle. Hier laufen die Interaktionen von Forschenden zur Selektion von Daten zusammen. Sie kann zur gezielten Suche nach Werten genutzt werden. Es unterstützt die einfache Suche, die Suche mit und in Facetten sowie eine Expertensuche (Retrievalsprache: SPARQL)
- **Visualisierung:** … ist eine weitere zentrale Komponente der Benutzungsschnittstelle. 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, Entitätenklassen). Knoten und Kanten sind mit den Ausgangsdatensätzen verlinkt. Zusätzlich werden Häufigkeitsverteilungen von Merkmalsausprägungen von

[12] Open Archives Initiative: http://www.openarchives.org/ (2021-09-23)
[13] Search/Retrieval via URL: https://www.loc.gov/standards/sru (2021-09-23)
[14] https://github.com/sonar-idh/Transformer

Knoten und Kanten angezeigt. Es können Vermittler/Hubs identifiziert werden. Gemeinsame Merkmale von Akteuren eines Netzwerks können hervorgehoben werden, z.B. Themen, Orte und Affiliationen zu Körperschaften. Die Reduktion auf egozentrische Netzwerke in einer Visualisierung ist möglich. Visualisierungen können in den Formaten SVG und PNG gespeichert werden.

- **Download-Button:** … ist die Komponente des Frontends, um den Datenexport über die Benutzungsschnittstelle auszulösen. Er löst eine mehrstufige Interaktion aus, um den Export in Bezug auf Umfang und struktureller Form zu parametrisieren.
- **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.

3.2 Kernprozesse und Funktionen

Die in 3.1 beschriebenen Kernkomponenten sind Voraussetzung zur Durchführung der zentralen Systemprozesse (Abb. 2). Sie beruhen auf den identifizierten Anwendungsfällen (Anhang 2). Die Prozesse betreffen: (1) Datenaggregation und -aufbereitung (Backend) sowie (2) Bereitstellung der Daten (Frontend).

(1) Datenaggregation und -aufbereitung (Backend)

Die für die HNA benötigten Daten, speziell die Meta- und Normdaten, werden aus sehr diversen Quellen gewonnen. Sie müssen für die SoNAR-Dienste integriert werden. Der erste Schritt ist das Aggregieren der Daten über einschlägige Schnittstellen (OAI, SRU). Für die Hauptdatenquellen[15] geschieht dies automatisch im monatlichen Rhythmus. Weitere, von einem Metadatenmanager zertifizierte anbietende Systeme stoßen die Übertragung selbständig an.

Die eingesammelten Eingangsdaten werden im Datenspeicher abgelegt, um unabhängig von Datendumps der Anbieter stets auf eine aktuelle, stabile Version der Datenanbieter zugreifen zu können. Für jede Datenquelle wird immer eine neue Archivdatei angelegt, die den kompletten, aktuellen Datenbestand eines Zeitpunkts enthält. Die Archivdateien sind versioniert und online verfügbar. Zudem wird auch eine Version der aufbereiteten SoNAR-Daten (Ausgangsdaten) nach jeder monatlichen Aktualisierung als Archivdatei online bereitgestellt. Die Versionsbeschreibung der Archivdatei der Ausgangsdaten enthält die URLs der in die jeweilige Version eingegangenen Archivdateien der Eingangsdaten und die URL (z.B. des GitHub-Repositoriums) der Konfiguration der ETL-Komponente. Die URL der Versionsbeschreibung kann zur Zitation genutzt werden[16].

Die Notwendigkeit zur Archivierung und Versionierung resultiert aus der in Kapitel 2 formulierten Anforderung an Transparenz und Reproduzierbarkeit von Forschungsergebnissen. Diese Lösung ermöglicht es, nicht nur auf die Eingangs- und Ausgangsdaten von SoNAR zu einem Zeitpunkt X zuzugreifen, sondern vielmehr die gesamte Umgebung eines Zeitpunkts wiederherzustellen. Die

[15] Hauptdatenquellen sind größere Datenverbünde und -portale, speziell in einer ersten Implementierung: Gemeinsame Normdatei (GND), Kalliope-Verbund (KPE), Zeitschriftendatenbank (ZDB), Gemeinsamer Bibliotheksverbund (GBV), ggf. zur Ergänzung der GND: Social Network and Archival Context (SNAC).
[16] Alternativ kann bei einer Implementierung für jede Versionsbeschreibung eine DOI erzeugt werden.

Versionsbeschreibung dokumentiert so die Genese der Ausgangsdaten eines Zeitpunkts. So wird es möglich, potenzielle Probleme der Datentransformation ex-post zu erkennen[17].

Regelmäßig werden die aktuellen Daten in mehreren Schritten durch das ETL-Tool aufbereitet. In einem ersten Schritt werden Validierungs- und Konsistenzprüfungen durchgeführt, um die Datenqualität zu sichern, und, wenn sie in einer XML-Struktur vorliegen, in RDF transformiert. In einem zweiten Schritt werden Daten maschinell ergänzt. Hierzu zählen:

1) Gleiche Entitäten werden zusammengeführt, z.B. Hannah Arendt, 1906-1975: Sie kann in Eingangsdaten durch verschiedene Identnummern diverser Normdateien (VIAF, SNAC, LoC-NACO, Wikidata, ISNI) repräsentiert sein. Mittels ID-Konkordanz, z.B. Lobid.Org, VIAF und Wikidata, werden diese Entitäten unter einer ID in SoNAR zusammengeführt.
2) Für Entitäten, die z.B. durch Wikidata oder ISNI identifiziert sind und keine GND-Daten ermittelt werden können, werden Beschreibungen maschinell von den referenzierten Normdateien abgefragt und integriert.
3) Zusätzlich werden Beschreibungen der GND um Daten für ausgewählte Merkmale aus der Wikidata ergänzt.

In einem dritten, letzten, Schritt erfolgt die Abbildung der Eingangsdaten auf das Datenmodell der HNA (Anhang 4). Die Transformation der Eingangsdaten erfolgt automatisiert durch Skripte des ETL-Tools, die nicht durch Code, sondern einer domänenspezifischen Sprache konfiguriert werden. Die erzeugten Daten werden in den RDF-Triplestore überführt. Sie beinhalten die Daten über historische Netzwerke sowie die Provenienz und Verarbeitung der Daten.

Die Transformations- und Anreicherungsprozesse der Aufbereitung der Eingangsdaten bedürfen der bibliothekarischen Betreuung. Hierfür wurde die Rolle Metadatenmanager/-in definiert. Die Aufgabe ist die Konfiguration des ETL-Tools, wo die Regeln für die Transformation festgelegt und angepasst werden. Zu den Tätigkeiten 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. Es macht das Angebot im Web auffindbar und ist die Schnittstelle für Recherche, Exploration und Download. Das Web-Frontend wird responsiv realisiert. SoNAR soll auf klassischen stationären und mobilen Rechnern als auch auf Tablets und Smartphones genutzt werden können. Für Recherche, Exploration und Download wird ein Dashboard mit drei Arbeitsflächen bereitgestellt: (1) Datenauswahl (Suche, Facetten), (2) Graph-Visualisierung von ausgewählten Datensegmenten und (3) Dokumentation. Abhängig von der Bildschirmgröße des Endgeräts können die Bereiche unmittelbar nebeneinander dargestellt oder ausgewählt werden.

Der Bereich **Datenauswahl**[18] enthält Funktionen zur Suche (s. Kernkomponente *Suchformular*). Diese umfasst die einfache Suche, Suche mit und in Facetten sowie die Expertensuche (SPARQL). Facetten (Anhang 4) sind ein Ansatz für das explorierende Browsen. Werte einer Facette können alphabetisch oder nach Häufigkeit sortiert oder nach der Datenprovenienz (anbietendes System)

[17] 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.
[18] Beispielkonzept für Datenauswahl mit explorativen Methoden: https://github.com/sonar-idh/visualization-prototypes/blob/main/img/prototype01.jpg

gruppiert werden. In jeder Facette kann analog zu einer erweiterten Suche (feldbezogen) gesucht werden. Es können ein oder mehrere Werte einer Facette oder eine Kombination von Werten mehrerer Facetten zur Bildung von Teildatenmengen markiert oder ausgeschlossen werden. Eine begleitende Filterung der Suchergebnisse nach zeitlichen oder räumlichen Kriterien sowie Entitätenklassen ist für beide Suchen (Suchformular, Visualisierung) zwingend angedacht.

Der **Visualisierungsbereich** (s. Kernkomponente *Visualisierung*) bildet die Graphen visuell ab. Sie reagiert direkt auf Selektionen im Bereich Datenauswahl. Die Grpah-Visualisierung unterstützt die Exploration des Datenbestands mit dem Ziel, einerseits Fragestellungen und Hypothesen zu entwickeln und andererseits Daten für nachnutzende Systeme (Forschungsumgebungen) zu selektieren. Hierfür unterstützen interaktive Schaltflächen das Hinzufügen und Entfernen von Knoten und Kanten. Die Bereiche Datenauswahl und Visualisierung sind so zwei Seiten einer Medaille: für die Recherche und Exploration der SoNAR-Daten.

Gegenüber diesen beiden Bereichen unterstützt der dritte Bereich, **Dokumentation**, nicht die aktive Selektion von Daten. Er informiert über die ausgewählte Datenmenge. Hierzu zählen:

- Liste der Ausgangsdatensätze des Graphens mit Link zum anbietenden System[19]
- Werte und Häufigkeiten der Merkmale der Akteure und Relationen im Graphen[20]

Die Liste der Ausgangsdatensätze enthält stets auch den Zeitstempel der Datenintegration und den Hinweis auf das anbietende System. Die Liste der Ausgangsdaten kann nach den anbietenden Systemen gruppiert werden. Ausgangsdaten, die markiert werden, führen zur Hervorhebung der Kanten und Knoten im visualisierten Graphen. Dasselbe gilt für Werte der Merkmale, die in der Dokumentation angeklickt werden, sodass diese zügig visuell im Grafen entdeckt werden.

Für die Visualisierung der Netzwerkgraphen sind drei Konzepte von Bedeutung. Sie wurden von den Visualisierungsstudien der Fachhochschule Potsdam adaptiert:

Fächer für multimodale Beziehungen zwischen zwei Akteuren[21]

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 4, 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 einzelne Stab eines Fächers repräsentiert einen Beziehungstyp, der wiederum angeklickt werden kann, sodass in der Dokumentation die Ausgangsdaten aufgelistet werden, die den Beziehungstyp belegen. Die Ausgangsdaten in der Dokumentation sind mit den Repräsentationen der anbietenden Systeme durch einen persistenten Link direkt verbunden.

Hervorhebung von Akteuren mit gemeinsamen Merkmalen

Durch die Auswahl eines oder mehrerer Werte oder Ausgangsdatensätze in der Dokumentation werden Akteure und Kanten im Graphen hervorgehoben, die diese Werte enthalten, sodass mit dieser Funktion Zusammenhänge zwischen Akteuren eines Graphens sichtbar werden. Dies sind

[19] Beispiel für Liste der Ausgangsdaten des Graphens mit Link zum anbietenden System: https://github.com/sonar-idh/visualization-prototypes/blob/main/img/prototype04.jpg
[20] Beispiel für einzelne Statistiken (Geschlecht): https://github.com/sonar-idh/visualization-prototypes/blob/main/img/prototype03.jpg
[21] Beispiel für Fächer: https://github.com/sonar-idh/visualization-prototypes/blob/main/img/17.jpg

bspw. ego-zentrierte Netzwerke, familiäre Beziehungen, das Wirken an einem Ort oder für eine Körperschaft oder persönliche Merkmale wie Geschlecht, Sprache, Religion oder Herkunft.

Bildung von merkmalsbezogenen Clustern

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 nicht beeinflusst, bei Clustern werden sie dagegen im Graphen nach den Gemeinsamkeiten gruppiert.

Die Hervorhebung und das Clustern von Akteuren nach Merkmalen unterstützt die Selektion von Daten. Die Rechercheergebnisse können zur Analyse in externen Anwendungen in verschiedenen Formaten zusammen mit technischen Metadaten, d.h. Daten zu Provenienz und Verarbeitung, in den Formaten RDF (JSON-LD, XML) oder CSV heruntergeladen werden. (s. Kernkomponente *Download-Button*). Die Ausgangsdaten sind persistent für eine Nachprüfung im Datenspeicher abrufbar; auf entfernte Ausgangsdaten durch Aktualisierung weist das SoNAR-System 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. Aus den Erfahrungen können folgende Empfehlungen für die einzelnen SoNAR-Komponenten und ihr Zusammenspiel abgeleitet werden (Abb. 3). Die Anforderungen und Zielsetzungen legen eine bevorzugte Verwendung von Open Source-Software nahe:

Catmandu: … ist ein CLI-Werkzeug (Command Line Interface). Es ermöglicht die Konfiguration und Durchführung von ETL-Prozessen (Extract, Transform, Load). Es ist eine etablierte Open-Source-Anwendung. Die SBB verfügt über Erfahrungen mit der Anwendung; es wird u.a. bei der Zeitschriftendatenbank (ZDB) eingesetzt. Da Catmandu von Haus aus einen OAI-PMH- und einen SRU-Client zur Aggregation von Daten mitbringt und gängige Bibliotheksmetadaten wie MARC21 und MODS transformieren kann, erfüllt es bereits viele Anforderungen an ein ETL-Tool für eine Forschungstechnologie SoNAR und ist damit eine solide Option für die ETL-Systemkomponente.

Metafacture: … ist ein alternatives, in Java geschriebenes Open-Source-Tool für ETL-Prozesse. Es kann als eigenständiges CLI-Werkzeug eingesetzt oder als Java-Bibliothek in Projekten wie SoNAR eingebunden werden. Metafacture ist modular aufgebaut 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: … ist als RDF-Triplestore des SoNAR Systems besonders geeignet. Hierbei handelt es sich um eine in Java geschriebene Open-Source-Applikation, die speziell auf Performance und Skalierbarkeit ausgelegt ist. Blazegraph unterstützt die RDF-Spezifikationen für die Beschreibung,

Verarbeitung und Repräsentation von Graph-Daten. Sie kann direkt in der jeweiligen Anwendung eingebettet oder als eigenständiger Datenbankserver betrieben werden, der über die eingebaute REST-Schnittstelle abgefragt werden kann.

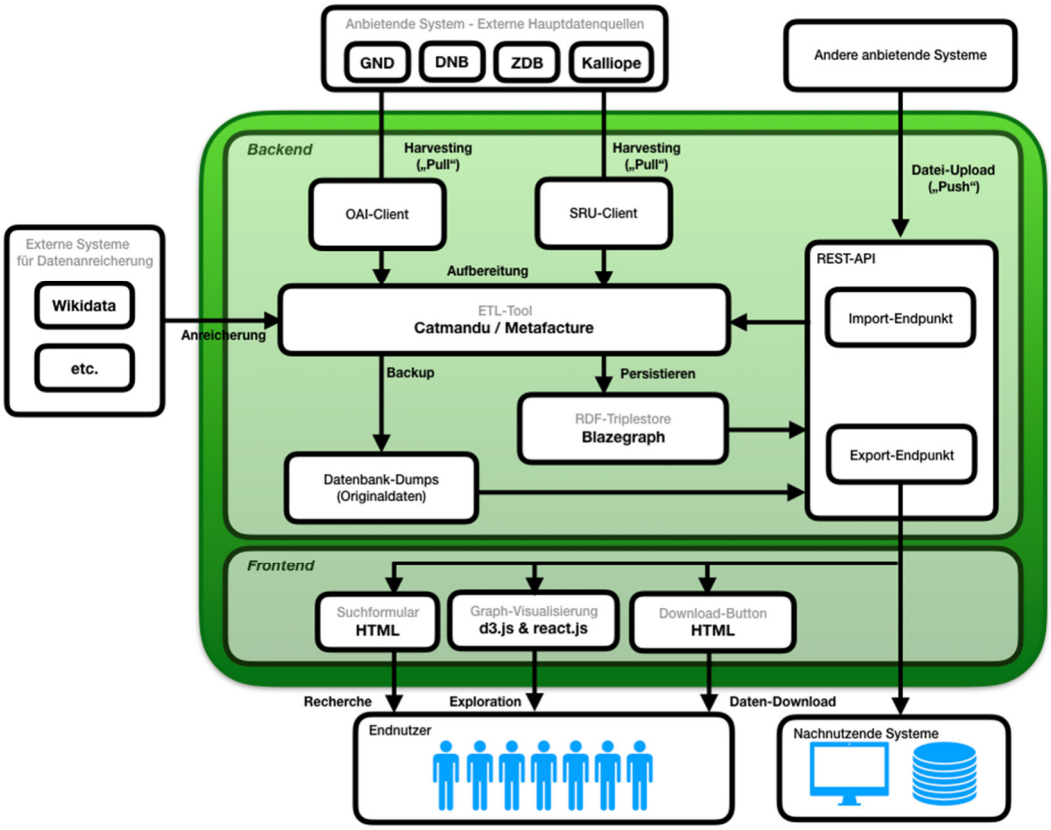

Abbildung 3: Implementierungsempfehlung für Kernkomponenten

D3.js & react.js: … ist eine JavaScript-Bibliothek zur Erzeugung dynamischer Visualisierungen. Sie unterstützt Vektorgraphiken und eignet sich daher für die Darstellung und auch Speicherung von Graph-Daten im Web-Frontend von SoNAR.

Das JavaScript-Framework react.js ist konkret für die komplexe Benutzerschnittstellen von Web-Anwendungen gemacht. In Verbindung mit modernen Web-Technologien wie HTML5 und CSS3 kann es in SoNAR für die Entwicklung eines modernen User Interfaces (UI) eingesetzt werden.

Sowohl D3.js als auch react.js sind Open-Source-Software.

4. Ausblick

Eine Forschungstechnologie SoNAR ist ein Datenangebot. Sie schließt eine Lücke der Forschung, indem umfangreich Datenbestände integriert, Daten über historische Netzwerke 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 geführt werden kann. Der technologische Linked-Data-Ansatz mit einer HNR-Datenmodellierung ist entscheidend, um auch nach einer Implementierung von SoNAR den HNR-Datenbestand in Verbindung mit offenen und dokumentierten Schnittstellen kontinuierlich erweitern zu können: in Bezug auf die anbietenden Systeme und in Bezug auf das Set möglicher Aussagen über Akteure und ihre Beziehungen. Damit

wird SoNAR zu einem Informationsangebot, dass sukzessive das bekannte, dezentrale Wissen von Kultureinrichtungen über Akteure zusammenführt. Dieser Prozess wird von Verbund- und Portalstrukturen als Hauptdatenquellen für einen HNR-Kerndatenbestand unterstützt. Weiterhin werden digitale Repositorien einzelner Einrichtungen und Zusammenschlüssen berücksichtigt, um auch sehr spezifische Aussagen über Akteure, die quellenbasiert etwa in themenzentrierten Projekten wie Matrikelportalen von Hochschulen[22] integrieren zu können. SoNAR digitalisiert so Datenerhebungsmethoden für wissenschaftliche Untersuchungen, in denen etwa Interviews und Beobachtungen nicht zur Verfügung stehen können.

Die Entscheidung, die FAIR-Prinzipien als Leitbild (und -planken) bei der Spezifizierung von SoNAR zu nehmen, hat zu einem konsequent offenen Design des Implementierungskonzepts geführt:

- Offenheit des Datenmodells
- Offenheit der Datenprovenienz
- Offenheit der Datenverarbeitung
- Offenheit der Datenbereitstellung
- Offenheit gegenüber Softwarelösungen

Für Forschungsprozesse bedeutet dies, mit einfachen Mitteln Transparenz über die Herkunft von Daten in Forschungsprojekten herzustellen. Zugleich können durch einen einfachen und zugleich innovativen Ansatz Datenstände etwa für Reproduktionen wiederhergestellt werden.

Neben der Forschung wird SoNAR auch für die Erschließung und Digitalisierung von Ressourcen in Kultureinrichtungen als eine Datenquelle in Informationssystemen eingesetzt werden können, z.B. in Archivanwendungen. SoNAR bietet einen umfassenderen Informationsstand im Vergleich zu einer einzelnen Normdatei. Dabei wird SoNAR nicht an die Stelle von Normdateien treten; es setzt vielmehr die Arbeit mit Normdateien etwa für die eindeutige Identifikation von Entitäten durch eine Normdatei voraus. Es setzt ebenso voraus, dass Identifikationsnummern von diversen Normdateien wie GND, ISNI, SNAC, Wikidata etc. durch Konkordanzen aufeinander abgebildet sind, um in SoNAR gleiche Entitäten trotz unterschiedlicher Normdatensatznummern erkennen zu können. Normdateien bieten zudem die notwendigen Schnittstellen, um Daten etwa über die Akteure zu erfassen. Indem SoNAR diverse Datenbestände jedoch wieder zusammenführt, kann ein erweiterter Datenbestand bzw. eine umfassendere Menge von Aussagen über Akteure aus den Daten für Historische Netzwerkanalysen und weitere Nutzungsszenarien extrahiert werden. Wie das SoNAR-Projekt zeigte, beinhalten sowohl Normdaten als auch Erschließungsdaten und bibliographische Metadaten vielfältige Aussagen über Akteure. SoNAR vereinigt diese in einem akteurszentrierten HNR-Datenbestand, der so für Forschung einen einmaligen Wert gewinnt (vgl. HHU 2021, ###).

Das vorgelegte Konzept ist das Ergebnis des Projekts SoNAR (IDH). Es ist vorgesehen, innerhalb des Förderprogramms e-Research-Technologien die Implementierung zu beantragen. Die Pflege des SoNAR Systems wird die Staatsbibliothek zu Berlin mit den Abteilungen IIE, Überregionale Bibliographische Dienste, und IDM, Informations- und Datenmanagement, übernehmen. Die Weiterentwicklung durch funktionale Erweiterung oder von Visualisierungskonzepten wird dann

[22] Beispielhaft genannt seien hier die Professoren- und Matrikelportale von Universitäten, z.B. Rostock (http://matrikel.uni-rostock.de/) oder Hamburg (https://www.hpk.uni-hamburg.de/)

projektorientiert in kooperativer Arbeit mit Forschungsvorhaben geplant und, die Finanzierung vorausgesetzt, umgesetzt.

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. Juli 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/ von Keyserlingk, Linda (Hg.): Handbuch Historische Netzwerkforschung. Grundlagen und Anwendungen. Berlin, 2016

Düring, Marten/ von 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: RiC-O projects and tools. 2021 https://ica-egad.github.io/RiC-O/projects-and-tools.html

HHU 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/ von 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/ von Keyserlingk-Rehbein, LInda/ Stark, Martin/ Düring, Marten (Hg.): The Power of Networks. Prospects of Historical Network Research. New York, 2020. 253-279

Schnaitter, Hannes, Evaluierung IV

Anhang

A1 Bedarfs- und Umfeldanalyse

https://github.com/sonar-idh/reports/blob/main/A01-SBB-Umfeldanalyse.pdf

A2 Systembeschreibung

https://github.com/sonar-idh/reports/blob/main/A02-SBB-Systembeschreibung.xlsx

A3 Datenmodellierung

https://github.com/sonar-idh/reports/blob/main/A03-SBB-Datenmodellierung.pdf

A4 Aufwandsabschätzung

s. Tabelle SoNAR-2021-A3-Aufwandsabschätzung.xlsx