

SoNAR AP-3: Wissenschaftliches Konzept für die Visualisierung von und Interaktion mit Graphen & Projektdokumentation

Mark-Jan Bludau, M.A., UCLAB, Fachhochschule Potsdam
Marian Dörk, Prof. Dr. , UCLAB Fachhochschule Potsdam

“Auf der Basis wissenschaftlicher Anforderungen an die Analyse der aufbereiteten Daten werden neuartige Visualisierungs- und Interfacedesignkonzepte entworfen und in einer prototypischen Anwendung getestet. Die Anforderungen variieren nach Fragestellung und können von einer hohen Komplexität sein. In enger Abstimmung mit AP-2 werden visuelle und interaktive Repräsentationen der aufbereiteten Daten für methodisch vielfältige Untersuchungen erstellt. Die entstehenden Konzepte und Techniken werden mit AP-2 und AP-4 in iterativen Designzyklen entwickelt und evaluiert. Das Arbeitspaket umfasst Visualisierung und Interfacedesign als zwei Arbeitsbereiche, die in allen Phasen von Konzept, Prototyp und Evaluierung zusammen zu denken sind: Während Informationsvisualisierung die Überführung komplexer, abstrakter und umfangreicher Daten in graphische Form betrifft, widmet sich das Interfacedesign der Gestaltung einer integrierten Nutzerschnittstelle für die intuitive Interaktion mit den Daten. In einzelnen Projektphasen werden mittels Rapid Prototyping für sehr vielversprechende Visualisierungs- und Interaktionsansätze funktionale Prototypen für praxisnahe Tests entwickelt.”

(SoNAR(IDH) Projektantrag)

1. Einleitung und Hintergrund

Im Projekt SoNAR (IDH) wurde AP-3–Visualisierung und Interfacedesign von der Fachhochschule Potsdam geleitet und bearbeitet. Das Dokument beschreibt aufbauend auf einem iterativen Design-Prozess in enger Abstimmung und Zusammenarbeit mit Projektpartner*innen das entstandene Visualisierungskonzept und dessen Implementierung in einem funktionalen interaktiven Prototypen und greift dabei auf die vier Aufgabenpakete (AP) ein:

- AP-3-1 Graphen im Kontext
- AP-3-2 Visualisierungen
- AP-3-3 Interfacedesign
- AP-3-4 Finalisierung und Demonstration

1.1 Netzwerk-Visualisierung in der Historischen Netzwerkanalyse (AP-3.1)

Die im **SoNAR-Projekt** fokussierte Historische Netzwerkanalyse (HNA) beschreibt Methoden der Netzwerkanalyse welche Formen der (sozialen) Netzwerkanalyse mit Hilfe historischer Quellen für die historische Forschung nutzen (Rollinger *et al.*, 2017). Verbindungen oder Links in HNA werden daher oft aus historischen Dokumenten oder anderen Quellen konstruiert, was häufig zu Herausforderungen im Bezug auf die Arbeit mit geisteswissenschaftlichen Daten führt durch etwa das Vorhandensein von Unsicherheit, Mehrdeutigkeit, Subjektivität und Interpretation (Drucker, 2011).

Für die Interpretation eines ganzen Graphen, aber auch einzelner Verbindungen oder Beziehungen in einer Netzwerkvisualisierung ist es daher nicht nur von Bedeutung, die Existenz von Beziehungen zwischen zwei Entitäten zu visualisieren, sondern für Historiker*innen ist es für ihre Interpretation dieser Beziehungen unumgänglich, historische Quellen und Beziehungstypen und Faktoren, die hinter diesen Beziehungen stehen, offenzulegen (Düring and von Keyserlingk, 2015).

Generell ist die Visualisierung von Graph-Daten ein umfangreiches Feld mit einer Vielzahl an bereits vorhandener Literatur und angewandter Forschung zu in zum Beispiel graphenbezogenen Algorithmen (Gibson *et al.*, 2012; Jacomy *et al.*, 2014; z.B. Behrisch *et al.*, 2016), Task-Taxonomien für

Graphvisualisierungen (Lee *et al.*, 2006; z.B. Ahn *et al.*, 2013; Kerracher *et al.*, 2015), State of the Art Visualisierungstechniken (Van Ham and Perer, 2009; von Landesberger *et al.*, 2011; z.B. Pienta *et al.*, 2015) sowie der Einsatz von visuellen Hilfestellung für die Erstellung von Graph-Datenabfragen (z.B. Pienta *et al.*, 2017). Die **SoNAR-Netzwerke** können als multivariate oder auch facettierte Netzwerke bezeichnet werden. Multivariate Netzwerkvisualisierungen sind in vielen Bereichen relevant, z. B. in der Analyse sozialer Netzwerke, in Netzwerken der Biologie, im Software-Engineering, in Transportnetzwerken, in der Meteorologie und in anderen Bereichen. Hierbei geht es um visuelle Darstellungen von Graphen, deren Knoten und Kanten zusätzliche, potenziell heterogene Datenattribute enthalten (Kerren *et al.*, 2014; Nobre *et al.*, 2019). Wie auch in **SoNAR**, können diese Attribute zum Beispiel mehrere unterschiedliche Eigenschaften wie Labels, Kategorien, Werte oder sogar zeitliche oder räumliche Dimensionen beinhalten (Kerren *et al.*, 2014; Hadlak *et al.*, 2015; Nobre *et al.*, 2019). Facettierte Netzwerke hingegen sind Netzwerke, die auch zusätzliche Dimensionen wie Zeit oder Räumlichkeit vereinen (Hadlak *et al.*, 2015). Und obwohl in Bereichen der Netzwerk-Visualisierung bereits viel geforscht wurde und es auch praktisch eine Vielzahl an Anwendung gibt, beziehen sich die Forschungen und Taxonomien meist auf das breitere Feld der Graphenvisualisierung und oft berücksichtigen Visualisierungen und digitale Praktiken, die für geisteswissenschaftliche Daten verwendet werden, nicht speziell für HNA-Forschung oder geisteswissenschaftliche Forschung relevante Anforderungen (Drucker, 2011). Dies wird unter anderem von Lamqaddam *et al.* bestätigt, die speziell für geisteswissenschaftliche Daten – nach Durchsicht der Literatur im Bereich der digitalen Geisteswissenschaften und nach einem Workshop und Nutzerstudien mit Forschern der digitalen Geisteswissenschaften – z.B. erschlossen haben, dass die Provenienz und die Offenlegung und Verknüpfung mit Quellen zentrale Faktoren für die Glaubwürdigkeit und das Vertrauen in eine Visualisierung sind, welches geisteswissenschaftlichen Forschern ermöglicht, die Visualisierung/Daten auf der Grundlage ihrer eigenen Fachexpertise zu interpretieren (Lamqaddam *et al.*, 2020). Des Weiteren wird in einem Survey Paper zu facettierten Graphenvisualisierung aus dem Jahr 2015 festgestellt, dass die Visualisierung von Dimensionen wie Unsicherheit oder Provenienz noch weiter erforscht werden muss (Hadlak *et al.*, 2015).

Was Anwendungen im Bereich der historischen Netzwerkanalyse bei Webanwendungen betrifft, so wurden im Laufe der Jahre bereits viele Interfaces und Anwendungen entwickelt, die explorative webbasierte Netzwerkvisualisierungs-Tools für die visuell gestützte historische Netzwerkanalyse anbieten. Beispielsweise basiert das Projekt *histoGraph* (Novak *et al.*, 2014) auf Beziehungen, die algorithmisch aus einer Fotosammlung extrahiert werden, oder das Projekt *Six Degrees of Francis Bacon* (Warren *et al.*, 2016) rekonstruiert soziale Beziehungen aus biografischen Dokumenten, wobei beide Projekte Formen des Crowdsourcing zur weiteren Anreicherung oder Validierung nutzen. Insgesamt wird bei der Visualisierung historischer Netzwerke typischerweise zur Darstellung der Daten als Hauptkomponente der Visualisierung auf die Verwendung von Force Algorithmen und Node-Link-Diagramme zurückgegriffen, häufig z.B. in Kombination mit anderen Komponenten wie z.B. Zeitleisten und Filter-Möglichkeiten. Während in vielen anderen Forschungsfeldern zunehmend auch Darstellungsformen wie Adjazenz Matrizen als zusätzliche Visualisierungsform Verbreitung finden, u.a. weil sie gerade im analytischen Bereich im Bezug auf viele “Tasks” Vorteile gegenüber der klassischen Node-Link Diagramm Form bieten (Okoe *et al.*, 2019), so scheinen diese in vielen Tools für historische Netzwerkanalyse bislang noch wenig genutzt zu werden. Dies möglicherweise etwas mit der einfacheren Interpretierbarkeit der Darstellung durch Laien zu tun, aber durchaus womöglich auch mit Stärken bei Pfad-basierten “Tasks” (Okoe *et al.*, 2019).

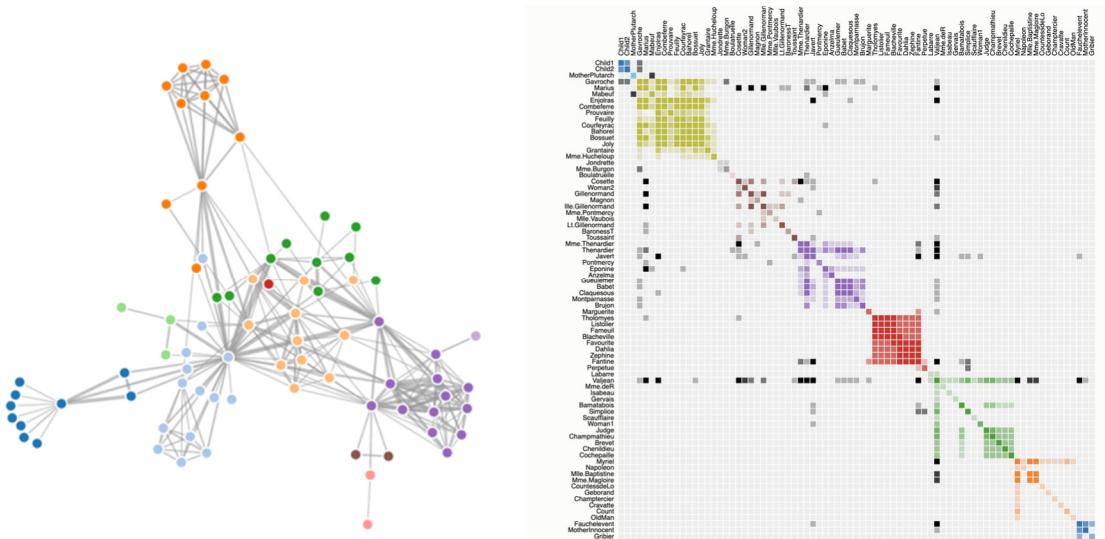


Abb. 1: Traditionelle Node-Link Darstellung (hier basierend auf einem Force-Algorithmus) (links) und Adjazenz Matrix (rechts) vom gleichen Datensatz.

Die bereits erwähnten Projekte *Six degrees of Francis Bacon*¹ (Warren *et al.*, 2016) und *histoGraph*² (Novak *et al.*, 2014) fallen zum Beispiel in dieses Spektrum aber auch Projekte wie Deutsche Biographie³. *histoGraph* zeichnet sich hier z.B., insbesondere durch die Offenlegung der Kanten-Quellen aus, was nicht verwunderlich ist, da es konkret für die historische Netzwerkanalyse entwickelt wurde.

Weiterhin typisch für die Darstellung historischer Netzwerke ist die Mischung unterschiedlicher Ansichten – meist Node-Link-Diagramme, zeitbasierte oder ortsbasierte Ansichten – zum Beispiel bei *Visualizing the Republic of Letters* (Chang *et al.*, 2009), *Kindred Britain*⁴ oder *The Vistorian*⁵ (Bach *et al.*, 2015). Häufig ist insbesondere eine zeitliche Übersicht oder Filterung entweder direkt in die Darstellung integriert, oder es gibt Visualisierungen mit besonderen Fokus auf zeitliche Aspekte. *Tudor Networks*⁶, ein historisches Korrespondenz-Netzwerk, verzichtet zum Beispiel gänzlich auf die gängige force-basierte Node-Link Diagramm Darstellung, bietet aber dafür aber eine zeitbasierte Übersicht, sowie eine zeitbasierte und ortsbasierte Detailansicht. Geographische, kartenbasierte Ansichten werden zudem häufig für eine Verortung von Akteuren und räumlichen Zusammenhängen genutzt, bei der aber unter anderem zwischen einer Einzeichnung von Knoten basierend auf Ortsverknüpfungen wie z.B. Geburtsort oder Sterbeort (z.B. *Deutsche Biographie*⁷) und Karten welche auch Kanten eingezeichnet haben unterschieden werden kann (z.B. *The Vistorian*⁸).

Insbesondere für die Darstellung konkreter Forschungsergebnisse finden auch narrative Ansätze Anwendung. Beim zu den SoNAR-Case Study (AP-2) ähnlichem Projekt *Researcher Connections – Understanding a decade of collaborations in autism science*⁹ werden die Netzwerk-Daten z.B. narrativ in Form von ‘Scrollytelling’ aufbereitet, und erst anschließend zur freien Exploration geöffnet.

¹ <http://www.sixdegreesoffrancisbacon.com/>

² <https://histograph.cvce.eu>

³ <https://www.deutsche-biographie.de>

⁴ <http://kindred.stanford.edu>

⁵ <https://networkcube.github.io/vistorian/web/index.html>

⁶ <http://tudornetworks.net/>

⁷ <https://www.deutsche-biographie.de>

⁸ <https://networkcube.github.io/vistorian/web/index.html>

⁹ <https://connections.spectrumnews.org/>

Im Bereich der webbasierten Netzwerk-Anwendungen gibt es zudem auch deutlich analysefokussierte Ansätze, z.B. GraphVis¹⁰ (Rossi and Ahmed, 2015), welche eine große Vielzahl kombinierbaren Filter-Möglichkeiten, Einstellungen, statistischen Methoden und sogar Machine-Learning bieten können, dafür aber auch deutlich weniger zugänglich für Erstnutzer*innen sind durch eine erhöhte Komplexität und hohe Individualisierbarkeit von Parametern.

1.2 Genutzte Technologien (AP-3.1)

Insbesondere Visualisierungen die über einfache forcebasierte Node-Link-Diagramme hinausgehen verwenden vorwiegend Frameworks wie D3.js und eine SVG Basis oder Canvas Basis, was insbesondere bei SVGs zwar zu Lasten der Performance gehen kann, aber dafür individuellere und einfacher zu implementierende Visualisierungsmöglichkeiten bietet. Visualisierungen die dagegen vornehmlich auf Performanz und die Darstellung von sehr vielen Knoten und Kanten aus sind, verwenden dagegen zunehmend WebGL als Basistechnologie, z.B. über Nutzung von den Visualisierungslibraries wie SigmaJs¹¹ (z.B. histoGraph¹²), nGraph¹³ oder Cytoscape.js¹⁴. Bei Kartenbasierten Darstellungen kommen zudem zum Teil auch Karten-spezifische Libraries wie z.B. Leaflet.js¹⁵ zum Einsatz (z.B. Deutsche Biographie¹⁶). Zusätzlich zu den Visualisierungs-Libraries, welche häufig hauptsächlich auf die Positionierung von Knoten und Kanten konzentriert sind, existieren auch individuelle javascript-basierte Implementierungen einzelner graphmetrischer Methoden, wie z.B. die Louvain community detection¹⁷ oder der Algorithmus von Dijkstra¹⁸ zur Findung des kürzesten Pfades. Zusätzlich bieten aber auch Graph-Datenbanken wie Neo4J integrierte graphmetrische Analysemethoden, welche direkt über die Queries die Nutzung von graphmetrischen Algorithmen oder statistischen Auswertungen ermöglichen¹⁹.

1.3 Fazit und Schlussfolgerung für Visualisierungen in SoNAR (AP-3.1)

Netzwerke scheinen überall zu sein: Von sozialen Medien, Biologie, historischer Forschung, Softwaretechnik bis hin zu Transportnetzwerken (Nobre *et al.*, 2019). Darauf aufbauend gibt es dementsprechend bereits eine Vielzahl vorangegangener Forschung im Bereich der Netzwerk-Visualisierung, häufig aber ohne die gezielte Berücksichtigung von HNA-spezifischen Bedingungen, Herausforderungen und Fragestellungen. Aber auch spezifisch im Bereich der historischen Netzwerkanalyse sind in den vergangenen Jahren Tools entstanden welche die Analyse und Exploration von historischen Netzwerken vereinfachen sollen. Dabei wird deutlich, dass hierbei häufig mehrere Ansichten den Nutzer*innen zur Verfügung gestellt werden.

Während viele dieser Tools aber auf z.B. konkrete Datensätze, Ego-Netzwerke oder Personen fokussiert sind, ist die Besonderheit von SoNAR demgegenüber die große Datenmenge (~52 Mio. Knoten, ~185 Mio. Kanten) und die Nutzung heterogener Datenquellen. Dies macht Visualisierungsmöglichkeiten von SoNAR nur bedingt mit den in Sektion 1 aufgeführten Beispielen vergleichbar und Herausforderungen entstehen insbesondere durch die Notwendigkeit von Skalierbarkeit und Performanz der Visualisierungsansätze sowie durch perzeptive, visuelle und technologische Limitierungen bei der Visualisierung komplexer und großer Daten (Fekete and Plaisant,

¹⁰ <http://networkrepository.com/graphvis.php>

¹¹ <http://sigmajs.org/>

¹² <https://histograph.cvce.eu/>

¹³ <https://github.com/anvaka/ngraph>

¹⁴ <https://js.cytoscape.org/>

¹⁵ <https://leafletjs.com/>

¹⁶ <https://www.deutsche-biographie.de>

¹⁷ <https://github.com/upphiminn/Louvain>

¹⁸ <http://bl.ocks.org/sdjacobs/3900867adc06c7680d48>

¹⁹ <https://neo4j.com/docs/graph-data-science/current/algorithms/>

2002). Was die Nutzung für geisteswissenschaftliche Forschung betrifft, so ist nach Drucker (2011) und Lamqaddam et al. (2020) die Interpretierbarkeit, Offenlegung für Unsicherheiten aber auch die Datentransparenz zudem ein wichtiges Kriterium und Herausforderung zugleich für eine geisteswissenschaftliche Forschung.

Für die Entwicklung von web-basierten Netzwerk-Ansichten wurden bereits eine Reihe von Visualisierungslibraries entwickelt. Für den weiterführenden Prototyping-Prozess im SoNAR-Erprobungsprojekt wurde sich anhand der Flexibilität, der großen Verbreitung in der Nutzung von Visualisierungen, aber auch aufgrund der Vorerfahrungen des Projektmitarbeiters für die Visualisierungslibrary D3.js entschieden, weil die JavaScript basierte Bibliothek im Gegensatz zu performanteren WebGL basierten Libraries insbesonderes im Bezug auf Individualität und Interaktivität eine flexiblere/schnellere/einfachere Implementierung ermöglicht. Zwar hat SoNAR insgesamt viele Knoten und Kanten, aber die jeweiligen einzelnen Netzwerke sind wesentlich kleiner. Performanz-Nachteile können tendenziell auch durch die Verwendung von Canvas anstelle von SVG abgefangen werden. Für eine zukünftige Implementierung der SoNAR Technologie sollte aber dennoch abgewogen werden, ob man auf eine performantere, neuere Technologie wie WebGL ganz oder teilweise umsteigen will, dafür aber speziellere Visualisierungen in der Entwicklung verkompliziert.

2. Methodik

[siehe Publikation “Graph Technologies for the Analysis of Historical Social Networks Using Heterogeneous Data Sources”]

Zur Entwicklung und Erprobung von Visualisierungskonzepten für SoNAR wurde ein iterativer Prototyping-Prozess angewendet. Hier haben sich Prozesse der Entwicklung des Forschungsdesigns, der Datentransformation und -zusammenführung, Entwicklung von Visualisierungen und die Evaluation dieser Arbeitsbereiche gegenseitig beeinflusst und unterstützt. Der iterativer Prozess basierte dabei auf fließenden Wechseln zwischen mehreren Eckpfeilern:

- 1) Co-Design Workshops
- 2) iteratives Rapid-Prototyping in direkter Zusammenarbeit mit HNA Forschern (AP-2) unter Nutzung der SoNAR-Daten (AP-1)
- 3) Erkenntnisse aus Interviews, Nutzerstudien und Feedback (AP-4)
- 4) Feedback über Fachkonferenzen (alle APs)

Wichtig hierbei war das der Prozess und das experimentelle Visualisieren selbst gemäß eines “Sandcasting-Prozesses” (Hinrichs et al., 2019) als Methode zum Erkenntnisgewinn genutzt wurde, der nicht nur die Möglichkeiten der Visualisierung austarieren, sondern auch Herausforderungen mit den Daten und Forschungspotential in Zusammenarbeit mit AP-2 aufdecken sollte.

2.1 Co-Design Workshops (AP-3.1)

Im gesamten Projektverlauf wurden einige kleinere und größere Workshops sowohl nur unter internen Projektbeteiligten, aber auch zusammen mit externen HNA-Forscher*innen durchgeführt. Während zu Projektbeginn auf Präsenz-Formate zurückgegriffen wurde, wurde im Projektverlauf pandemiebedingt zunehmend auf digitale Formate unter Nutzung von Collaboration-Tools wie z.B. Miro und Zoom gesetzt. Ziel der Workshops war es bei der Konzeptentwicklung von Anfang an nah an der HNA-Forschung und den Nutzer*innen zu bleiben und dabei einem “Grounded” Prototyping Prozess zu folgen (Isenberg et al., 2008). Die folgenden Abschnitten beschreiben die unter Leitung von AP-3 (FHP) durchgeföhrten Workshops mit Fokus auf die Visualisierungskonzepte. Neben diesen hier präsentierten Workshops haben jedoch auch weitere Workshops z.B. zu Datenmodellierung (AP-1) oder z.B. ein großer internationaler Workshop zum Thema Forschungsdesign (AP-2) stattgefunden.

2.1.1 Collagen Workshop

Ein erster wichtiger Meilenstein zu Beginn des Projekts war ein Co-Design Workshop zur HNA-Netzwerkvisualisierung. Zu Beginn des Projekts war es wichtig, Diskussionen über das Potenzial von bibliografischen (Meta-)Daten für HNA sowie Anforderungen an die Visualisierung historischer Netzwerke anzuregen. Um zentrale Forschungsfragen, Herausforderungen und Potentiale zu identifizieren, wurde dafür ein Co-Design-Workshop durchgeführt. Dafür haben wir, um neue Einblicke in die historische Netzwerkforschung und -visualisierung zu gewinnen, Fachexpert*innen zu einem Workshop eingeladen, angelehnt an Workshop-Ansätze von Chen et al. (2014) und Henry und Fekete (2006). Insgesamt haben zehn Personen an diesem Co-Design-Workshop teilgenommen, darunter vier historische oder soziale Netzwerkforscher*innen als Domänenexperten (eine Person davon intern von AP-2), zwei projektinterne Informationsvisualisierungs-Designer (AP-3), zwei Personen aus unserem projektinternen Evaluationsteam (AP-4), eine Person aus unserem Team von Data Scientists (AP-1, verantwortlich für die Datentransformation) und eine weitere externe Teilnehmerin mit Designhintergrund und Erfahrung mit dem Co-Design-Format. Ziel der interdisziplinären Zusammensetzung der Workshop-Gruppe war es, die Diskussion zu fördern, indem eine Vielzahl von Perspektiven auf das Thema HNA durch den Blickwinkel von HNA-Experten sowie frische Einblicke durch die Perspektive von Teilnehmern aus anderen (projektrelevanten) Fachbereichen geboten wurden. Der Workshop war für insgesamt drei Stunden angesetzt. Ähnlich dem Ansatz von Henry und Fekete (2006) haben wir den Workshop mit einer kleinen Präsentation eines breiten Spektrums aktueller Möglichkeiten und Entwicklungen der Netzwerkvisualisierung gestartet, einschließlich einiger neuartiger und experimenteller Ansätze. Um den Prozess der Konzeptfindung von Netzwerkvisualisierungen zu erleichtern, folgte eine kurze Visualisierungsaufgabe als eine Art Aufwärmübung, bei der die Teilnehmer*innen gebeten wurden, ein sehr kleines soziales Netzwerk (zehn Knoten) auf der Grundlage einer von uns bereitgestellten Datenmatrix zu visualisieren. Nach diesem Warm-up haben wir eine kurze Einführung über die Ziele unserer Forschung sowie über die Daten in unserem Projekt präsentiert. Anschließend haben wir die Teilnehmer*innen gebeten je eine Collage über Ansätze zur HNA-Forschung mit Blick auf unsere Daten und unser Projekt zu erstellen (Abb. 2).

Für die Collagen haben wir verschiedene Materialien zur Verfügung gestellt; so wurden Karten, Personen-Icons und beispielhafte Datenauszüge als Material bereitgestellt, sowie eine große Anzahl an Bastelmanualien, wie Papier, Kleber, Scheren, Buntstife, Filzstifte, Marker, farbige Aufkleber, farbiges Papier, Faden, etc. Während Chen et al. (2014) visuelles Material aus ihrer Fotosammlung zur Verfügung gestellt haben, sind unsere Daten eher abstrakt und weniger visuell. Um den Prozess des Collagierens mit visuellen Hilfsmitteln weiter zu unterstützen, haben wir daher weiteres visuelles Material zur Nutzung ausgelegt, darunter Ausdrucke einer leeren Karte, gedruckte Icons (z. B. als Darstellungen von Netzwerkknoten) und einige gedruckte Scans aus unseren Volltextdatenquellen. Um den Prozess in Gang zu bringen haben wir zusätzlich einige Fragen in den Raum geworfen, z. B. "Wie würden Sie sich gerne durch die Daten bewegen?" - "Welche Rolle spielen Datendimensionen wie Zeit, Raum oder semantische Beziehungen?", haben die Teilnehmer*innen aber dennoch ermutigt, sich frei zu fühlen, diese Eingangsfragen zu ignorieren. Generell bestand die Aufgabe des Collage-Prozesses nicht darin, Wireframe-ähnliche Skizzen für eine konkrete User-Interface-Lösung zu erstellen, sondern sich allgemeine Ansätze, Funktionalitäten und Einstiegspunkte in die HNA-Forschung und unsere Daten vorzustellen. Nach etwa 30 Minuten wurden alle Collagen einzeln im Plenum diskutiert. Zunächst wurden die Teilnehmer*innen, die nicht an einer Collage beteiligt waren, gebeten, das Gesehene zu interpretieren und darüber zu spekulieren. Daraufhin wurden die Ersteller*innen der Collagen gebeten, Erklärungen abzugeben und ihre Vorgehensweise mit der Gruppe zu diskutieren. In diesem Schritt sollen die typischerweise auftretenden Fehlinterpretationen der Ergebnisse zu weiteren Diskussionen und neuen Ideen anregen. Im letzten und zusammenfassenden Schritt wurde alle Teilnehmer gebeten ein abschließendes Statement zu den wichtigsten Erkenntnissen aus dem Prozess sowie zu den für sie wichtigsten Themen der Diskussion abzugeben.

Für die weitere Analyse und Dokumentation wurden Audioaufnahmen während des gesamten Workshops aufgenommen und der Prozess zudem fotografisch dokumentiert. Die Audioaufnahmen wurden anschließend transkribiert. Ziel des Workshops war es nicht, funktionale Wireframes oder konkrete Interaktionsprinzipien zu gestalten, sondern Diskussionen anzuregen, eine Sensibilität für die HNA und die Daten zu erlangen und wichtige domänenspezifische Forschungsaspekte und Herausforderungen aufzuzeigen. Während die Projektpartner der HU (AP-4) die Transkriptionen zusätzlich kodiert und eigene Auswertungen vorgenommen haben, haben wir für die Visualisierung relevante Themen und Diskussionspunkte extrahiert.

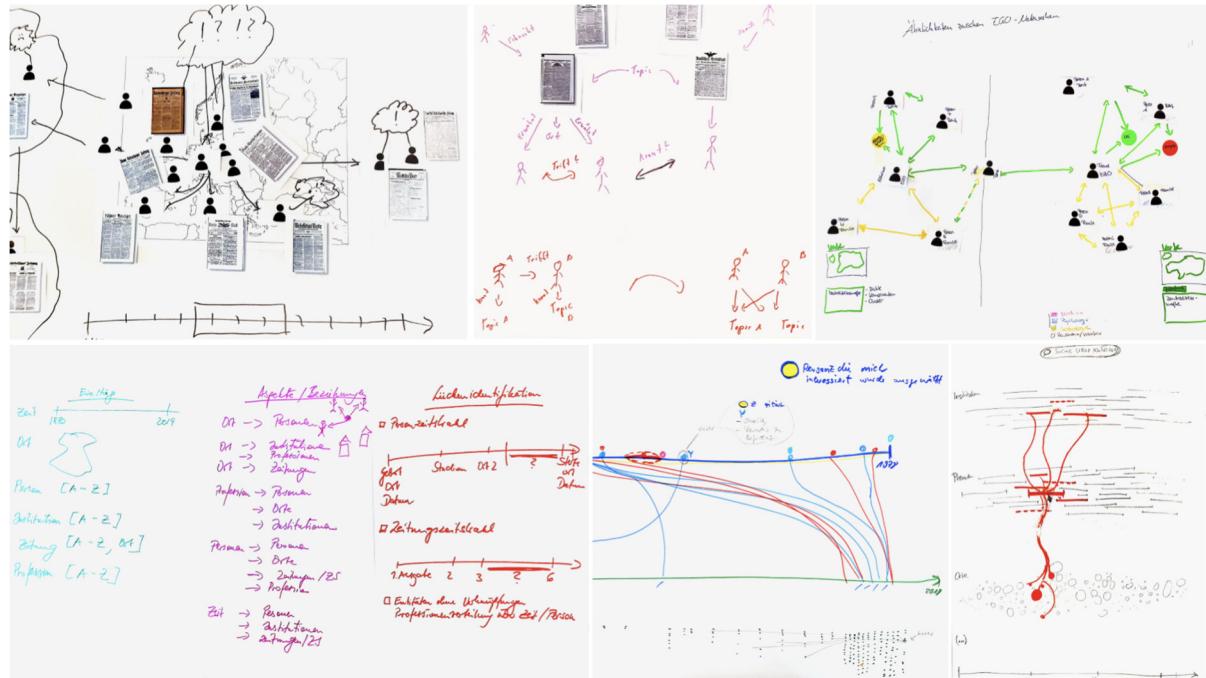


Abb. 2: Im Workshop entstandene Collagen

Als Ergebnisse des Collagen-Workshops sehen wir wie bereits aufgeführt in erster Linie nicht die im Workshop entstandenen Collagen und Visualisierungsideen selbst, sondern vielmehr die Diskussionspunkte und herausstechenden Themenfelder. Auch wenn der Workshop zu Beginn des Projektes bereits stattgefunden hat, so hat er sehr wichtige Erkenntnisse für die weitere Konzeptphase erbracht:

1. Datentransparenz und Sichtbarmachung von Unsicherheiten und Fehlstellen

Datentransparenz war ein entscheidendes Thema im Workshop, was immer wieder Erwähnung fand. Um Forschung und eine sachgerechte Interpretation zu ermöglichen ist es für die Nutzer*innen wichtig zu wissen, was die Daten enthalten, was sie nicht enthalten und wo es Ungenauigkeiten und Fehlstellen geben könnte, aber auch mit Blick auf Offenlegung von Graph-Metriken: *“Was können die Daten? und was können sie eben nicht?” – “Also auf das was fehlt aufmerksam zu werden ist häufig viel entscheidender als das was man da tatsächlich sieht. Und eine ganz andere pragmatische Frage die wir immer häufig sehen ist Fehler in den Daten.”* –

“Und was ich sonst noch interessant fand, wie manche visuell diese Unsicherheit in den Daten dargestellt haben, das fand ich ziemlich cool, weil das glaube ich schon öfter vorkommt und das glaube ich wird auch noch zu wenig gemacht in Informationsvisualisierungen so Unsicherheiten in Daten darzustellen.” – “Ich finde auch, dass gerade diese Daten-ethische Perspektive auf jeden Fall wichtig ist, gerade auch die Frage nach dem geographischen Raum, den würde ich auf jeden Fall auch nochmal transparent machen.”

2. Relevanz der Zeitleiste

Viele Collagen und Diskussionen haben die Wichtigkeit von zeitbasierter Selektion in historischen Netzwerken hervorgehoben: *“Ohne Timelines nutzen mir die Visualisierungen nichts – weder für die Analyse noch für die Vorstellung von Ergebnissen. Das heißt für die Rechercheweg und die Analyse brauche ich schon die Visualisierung in der Form dass ich vor- und zurückfahren kann. Ohne das macht es keinen Sinn. Wenn ich ein Netzwerk über 100 Jahre anzeige, okay, da kann ich vielleicht irgendwas erkennen, aber es ist vollkommen artifiziell.”*

3. Potenzial dynamischer/flexibler Ansichten

Da es für unterschiedliche Forschungsfragen unterschiedliche Schwerpunkte und Ansätze gibt, haben die Teilnehmer*innen häufig eine Wichtigkeit von Flexibilität und von Filtern betont. Zudem werden Visualisierungen auch zur Exploration und Findung neuer Forschungsfragen genutzt, weshalb eine flexible Darstellung umso wichtiger ist: *“Es ist natürlich ein großer Unterschied ob man jetzt z.B. verwandt ist, oder ob man Korrespondenzpartner ist oder ob man sich auf einem Kongress mal bei der Kaffeepause gesehen hat. Das sind alles Beziehungen, die aber natürlich in der Interpretation unterschiedliches Gewicht haben. Das ist z.B. etwas, was wir in der Visualisierung sehr gerne hätten. Dass man eben nicht nur sagt, so wie die hier, die haben Beziehung A B, sondern unterschiedliche Gewichtungen, je nach Forschungsfrage, also was interessiert mich, eben auch nicht statisch.”* – *“Was man dann für Beziehungen in den Daten sucht, das passiert ganz oft erst in dem Moment wo du das erste mal auf den Haufen drauf guckst.”* – *“Das Kriterium wonach man das sortiert und was man da zusammenfindet, das weiß man manchmal gar nicht was wichtig ist oder nicht, das sieht man in dem Moment wenn man ein Bild davon bekommt.”*

4. Reproduzierbarkeit und Nachnutzbarkeit Service-Charakter

Für die Teilnehmer*innen ist es ein wichtiger Aspekt, dass Ergebnisse und Visualisierungen auch festgehalten, zitiert und die Daten erneut genutzt werden können. Wichtig ist hierbei auch, dass es nicht nur um eine einmalige einfache Speicherung von visuellen Ergebnissen geht, sondern vielmehr um eine Zitierbarkeit und Reproduzierbarkeit auch durch andere Nutzer*innen. Zudem ist auch die Nachvollziehbarkeit verbunden mit einer einfachen Verknüpfung zu den Ausgangsdaten ein genannter Wunsch: *“Es ist besonders interessant für den Historiker, wenn er diese Ansicht zitieren kann.”* – *“Uns würde es natürlich sehr viel Spaß machen, wenn das direkt auch einen gewissen Service-Charakter hätte, also wenn ich auf die Knoten gehe, dass ich dann eine Verlinkung habe z.B. zum Kalliope.”*

5. Spannung zwischen klassischen und neuartigeren Visualisierungsformen

Es gab was die Entwicklung neuartiger Visualisierungen im Kontrast zu klassischeren alt-bewährten Formaten eine gewisse Grundskepsis auf Seiten einiger HNA-Forscher*innen. Hier ist es entscheidend, dass Innovation und Ästhetik die Nutzbarkeit nicht einschränken, weshalb die enge Zusammenarbeit mit HNA-Forscher*innen und Nutzer*innen im Projekt umso wichtiger ist: *“Und ganz oft sag ich: schön, aber es ist halt einfach nur ästhetisch. Aber ich habe keinen Erkenntnisgewinn wenn ich da drauf gucke. Da kann ich ne Stunde drauf gucken und ich verstehe es erst wenn ich die Erklärung dazu lese. Und dann ist die Frage, warum brauche ich die Grafik, wenn ich dann doch den Text lese.”*

2.1.3 Andere Workshops

Neben dem ausführlich beschriebenen Co-Design Workshop und genereller iterativer Zusammenarbeit, haben wir im Projektverlauf immer wieder auf ähnliche Methoden zurückgegriffen. Da im Rahmen der Corona-Pandemie ab der zweiten Projekt-Hälfte reale Treffen in größeren Gruppen nicht mehr durchführbar waren, wurde hierzu auf virtuelle Methoden zurückgegriffen. Beispielsweise wurden im Rahmen iterativer Konzeptionsprozesse und für Brainstorming das virtuelle Kollaborations-Tool Miro²⁰ in Kombination mit dem Videotelefonkonferenz-Softwareprogramm Zoom²¹ verwendet. Miro ist eine Art

²⁰ <https://www.miro.com>

²¹ <https://www.zoom.com>

Online-Whiteboard welches virtuelle Kollaborationsmöglichkeiten bietet. Im Rahmen unseres Projektes haben wir Miro unter anderem dazu genutzt Ansichten von einzelnen Prototypen virtuell auszulegen und in einem internen Workshop-Format mit Notizzetteln Kommentare, Fragen oder ähnliches anbringen zu lassen um so gezielter Feedback zu sammeln und Diskussionen anzuregen (siehe Abb 3).

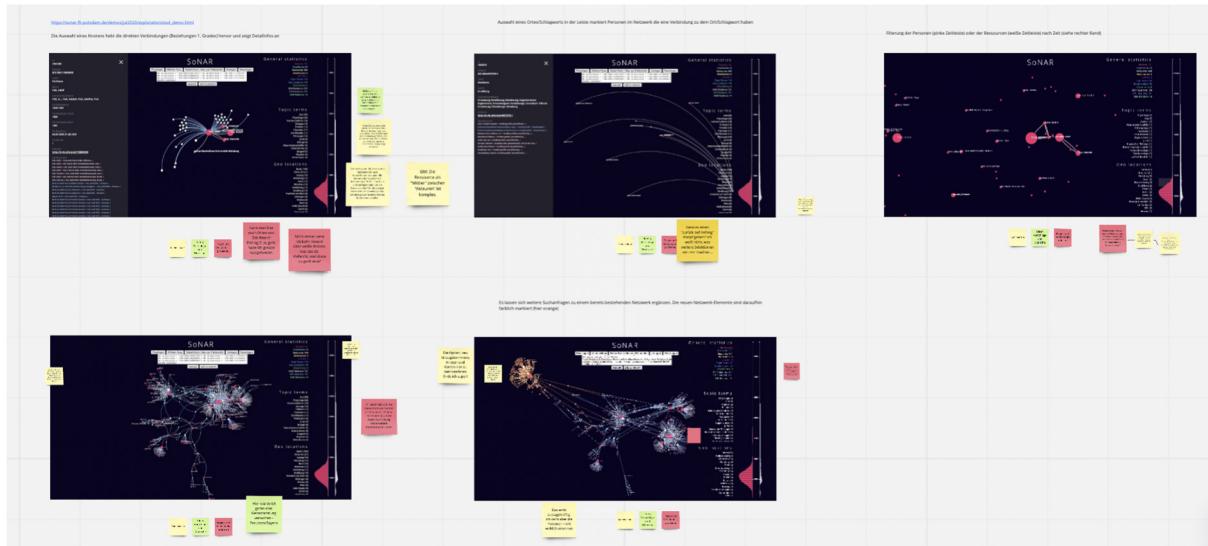


Abb. 3: Miro-Workshop mit Projektmitarbeitenden zu einzelnen Visualisierungskonzepten

Im späteren Projektverlauf wurde Miro in Kombination mit Zoom (zur Kommunikation) ebenfalls genutzt um Anforderung für die Nutzung von Jupyter Notebooks (siehe Sektion 6) zu ergründen und in Zusammenarbeit mit AP-2 Konzepte für Fragestellungen, eine Strukturierung und Schwerpunktsetzung in diesen zu erörtern (Abb. 4).

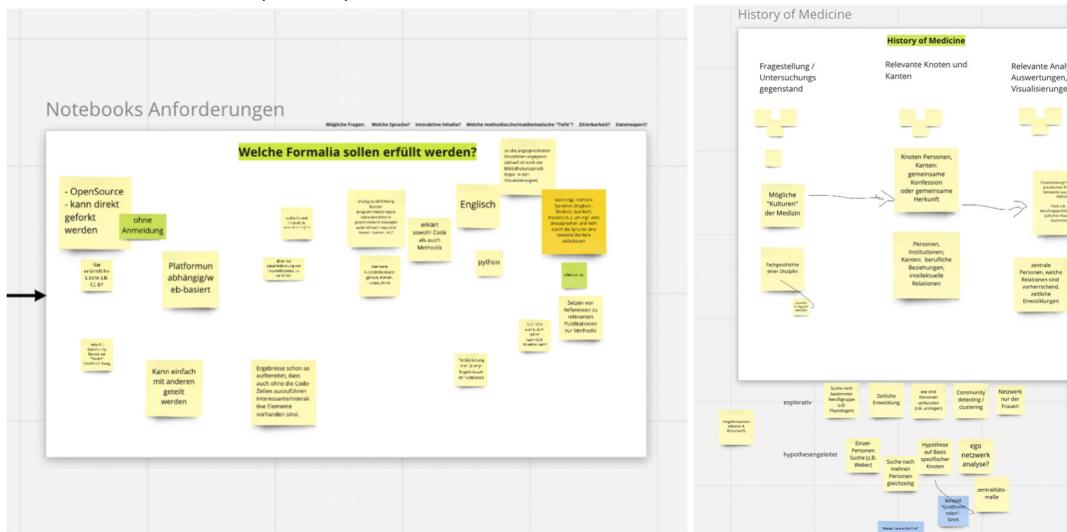


Abb. 4: Miro-Workshop mit Projektmitarbeitenden zum Thema Jupyter Notebooks als Hilfsmittel zur HNA-Forschung

2.2 Iterativer experimenteller Visualisierungsprozess (AP-3.2/3.3)

In einem iterativen Prototypingprozess in Zusammenarbeit des Projektteams, insbesondere mit den HNA-Forscher*Innen (AP-2) haben sich Prozesse der Entwicklung des Forschungsdesigns, der

Datentransformation und -zusammenführung, Entwicklung von Visualisierungen und der Evaluation dieser Arbeitsbereiche gegenseitig beeinflusst und unterstützt.

Obwohl der Co-Design Workshop, unsere Gespräche mit den projekt-internen HNA-Experten, Recherche existierender Tools und Forschungsergebnisse und bestehende Task-Taxonomie (Lee *et al.*, 2006; Ahn *et al.*, 2013; Kerracher *et al.*, 2015) bereits eine Vielzahl potenzieller Aufgaben, Bedürfnisse und Anforderungen an Graph-Visualisierungen offenlegen, sehen wir den Prototyping-Prozess selbst als eine Form von 'Research through Design' (Zimmerman *et al.*, 2007). Zusammen mit dem Datenmodellierungsprozess und den oben beschriebenen Co-Creation-Ansätzen sehen wir unseren Visualisierungsprozess als eine Form des schnellen, experimentellen und iterativen Prototyping-Prozesses und weiterhin als Mittel der Datenexploration. Wir verstehen experimentelle Ansätze und Umwege im Visualisierungsprozess selbst als eine Methodik der Wissensproduktion. Verglichen mit dem potenziell kürzesten Weg zu einem fertigen Visualisierungs-Tool entspricht dieser Ansatz eher einem Neugier getriebenen "Herumprobieren". Hierbei wurden Visualisierungen und kleine Prototypen nicht unbedingt mit dem Ziel erstellt, sie in einen endgültigen Prototyp oder ein Konzept umzusetzen. Vielmehr wurden sie zu einer Methode, um die Daten oder einzelne Facetten der Daten zu erforschen, zu einem Werkzeug, um die grundsätzlichen Herausforderungen von Daten oder deren Kodierung zu untersuchen, oder zu einem visuellen Vermittler, um die interdisziplinäre Kommunikation und die Entwicklung neuer und anregender Ansätze zu fördern (Hinrichs *et al.*, 2019).

Das gesamte Projekt und dessen grundsätzliche Arbeitspakete wurde von Anfang an interdisziplinär und parallel durchgeführt, d.h. Datenverarbeitung, Fallstudienentwicklung, Visualisierung und Auswertung erfolgten parallel. Ganz zu Beginn wurden die Daten weder für die Visualisierung aufbereitet, noch waren sie über eine Art API zugänglich, so dass zunächst mit einem Subset ausgewählter Daten gearbeitet wurde. Dies machte es zwar schwierig, alle Facetten und Herausforderungen, die mit dem Umgang mit dem gesamten Datenbestand verbunden sind, zu antizipieren, aber die frühe Arbeit mit Daten-Subsets gab uns die Möglichkeit, iterativ Einfluss auf die Datenverarbeitung und das Datenmodell zu nehmen. Im weiteren Projekt-Verlauf erfolgte dann schrittweise zunächst eine offline-Anbindung an erste Datentransformationsstufen über Neo4J, welche erstmals ein Zugriff auf die vollen Daten ermöglicht haben und sowie weitere iterative Überarbeitung des Datenmodells aufbauend auf Erkenntnissen aus dem Visualisierungsprozess in Zusammenarbeit mit AP-2 gefördert haben. In dieser Phase wurden Prototypen häufig über einen explorativen Prozess mit Hilfe von Screensharing zusammen mit den HNA-Forscher*innen von AP-2 besprochen und erkundet und so offene Fragen, Wünsche und Anforderungen in regelmäßigen Sitzungen besprochen. In der späteren Projektphase wurden schließlich die Daten auch online über über Neo4J abrufbar, was die iterative Zusammenarbeit mit anderen Projektpartner*innen weiter vereinfacht hat.

Anstatt zu versuchen, alle potenziellen Funktionen und Ideen in einem einzigen, umfassenden Prototyp zu vereinen, konzentriert sich unser Ansatz auf kleine, separate Probleme und Ideen durch eine Vielzahl von kleineren, schnell-entwickelten Prototypen. Viele unserer Designstudien oder Prototypen wurden dabei in enger Zusammenarbeit mit unseren eigenen HNA-Spezialisten entwickelt und/oder stützen sich auf Ergebnisse des Workshops, Erkenntnisse durch AP-4 oder anderer externer Expert*innen, während andere eher experimentellen Charakter haben und oft aus spontanen Impulsen entstanden sind. Zudem wurden basierend auf Visualisierungsbestandteilen Konzepte mit Hilfe von Wireframes und ausgearbeiteten Design-Skizzen erstellt (Abb. 5).

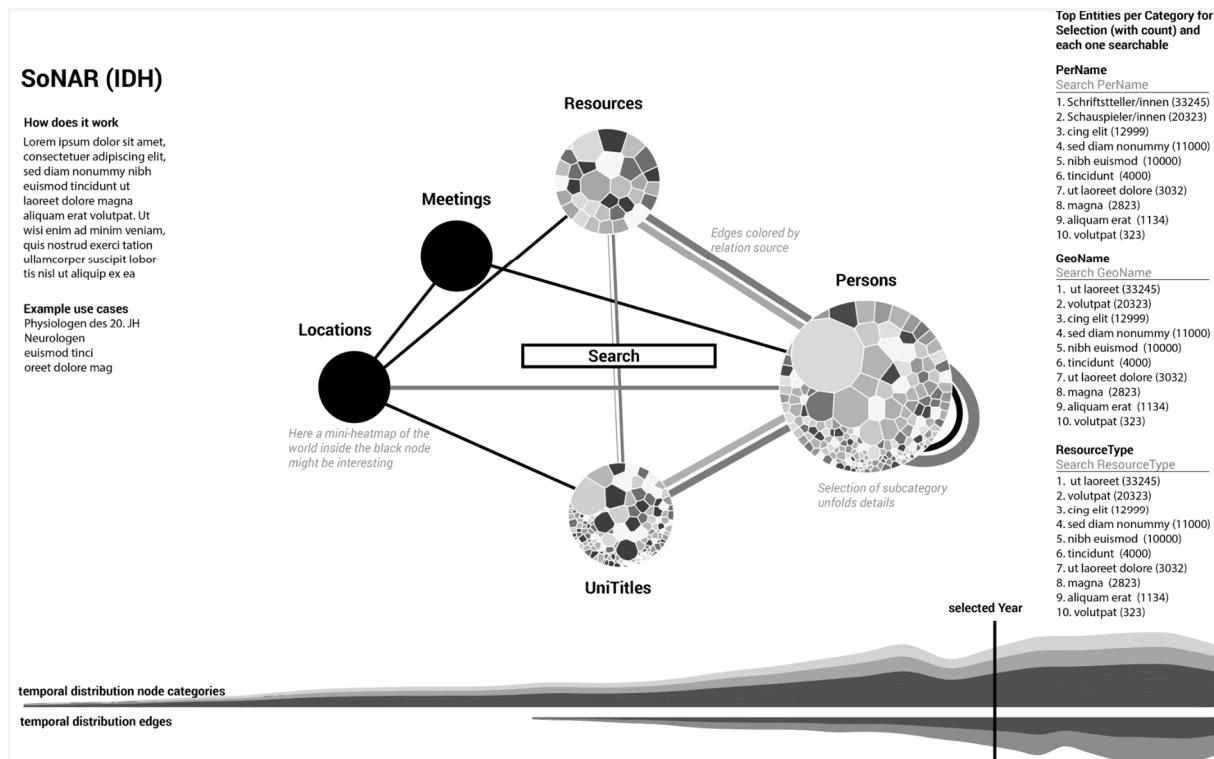


Abb.5: Frühe (nicht komplett datenbasierte) Skizzierung für einen Überblicks Start-Zugang zu den SoNAR Daten

Die folgenden Abschnitte beschreiben eine Auswahl an experimentell Prototypen und Zwischenergebnissen, welche größtenteils mit der Datenvisualisierungsbibliothek D3.js (Bostock et al., 2011) erstellt wurden. Diese sind Schritte auf dem Weg zu einem endgültigen Konzept, die iterativ durch das Feedback unserer Fachexperten und anderer potenzieller zukünftiger Anwender beeinflusst wurden:

2.2.1 Kleine Design-Studien und Visualisierungsexperimente (AP-3.2)

In ersten kleinen Design-Studien und Visualisierungsprototypen (Abb. 6) wurden in einer Art Rapid-Prototyping Prozess viele schnelle kleine Ansätze, Visualisierungs-Libraries oder Technologien ausprobiert um sich verstärkt mit Methoden der Netzwerk-Visualisierung auseinanderzusetzen und die Daten zu explorieren. Dazu gehörten zum Beispiel Ansätze die mehrere Beziehungen zu einer Kante aggregieren und bei Klick auffächern lassen (welche im späteren Projekt-Verlauf weiter verfolgt wurde), aber auch Konzepte zur Visualisierung von Kanten-Unsicherheit, aber auch andere Experimente, die unterschiedliche Parameter zur Anordnung von Knoten auf Basis von kleineren Daten-Subsets erprobt haben, z.B. Anordnung und Darstellung von Knoten oder geographischen Orten nach Häufigkeit, oder die Nutzung von Community-Algorithmen bei größeren Netzwerken.

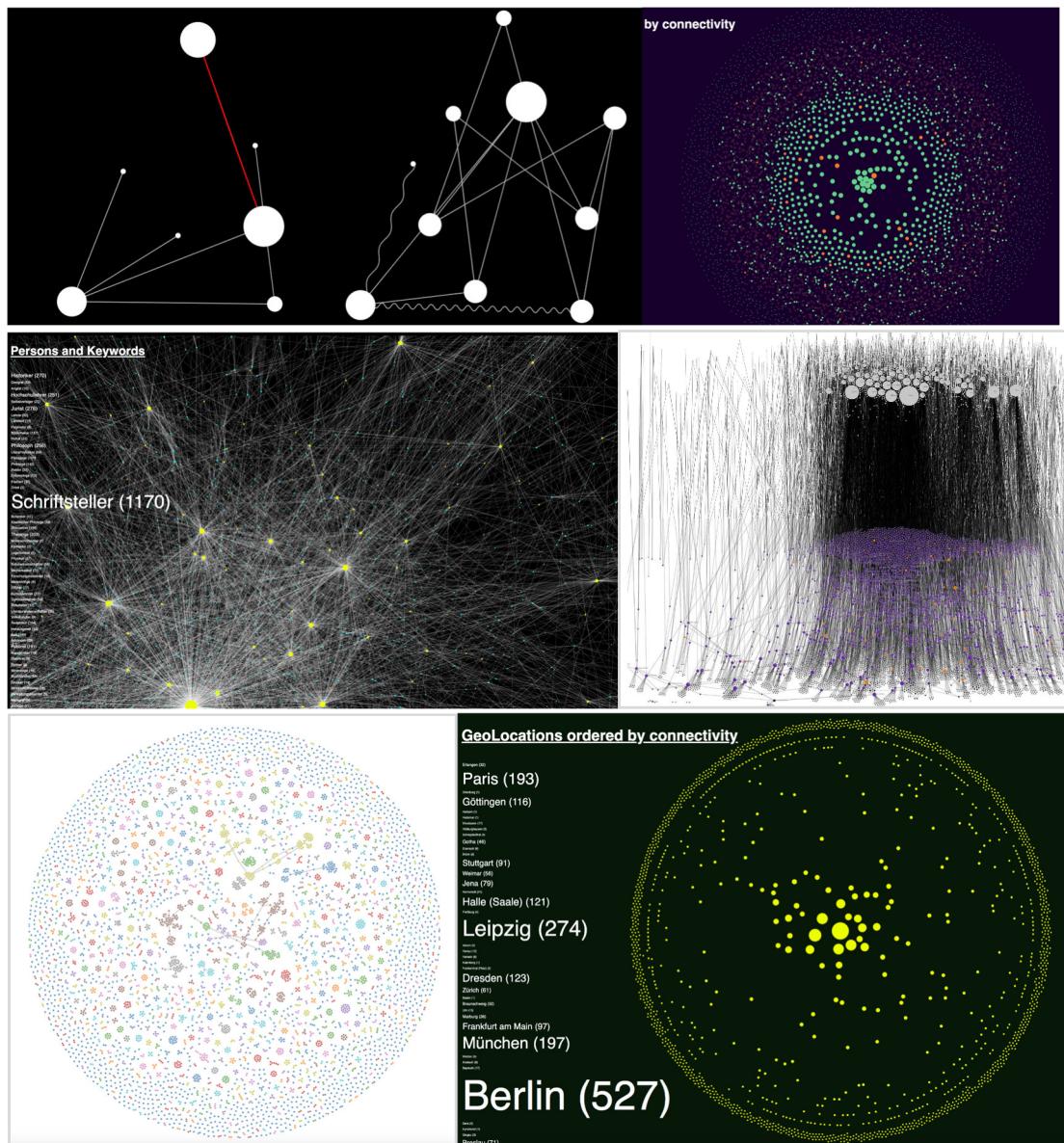


Abb.6: Mehrere erste experimentelle Visualisierungen aus dem ersten Projekt-Halbjahr.

2.2.2 Anordnung nach Schlagwort-Ähnlichkeit und Scrolling durch die Zeit (AP-3.2)

In einem weiteren Schritt wurde mit den verknüpften Schlagworten in einem Daten-Subset von ca. 3000 Personen experimentiert (Abb. 7). Dafür wurde die Dimensionalitätsreduktionstechnik UMAP (McInnes et al., 2018) verwendet. Jede Person im Datensatz kann mit mehreren Schlagworten (häufig Berufe) verknüpft sein. Diese Schlagworte kann man als Daten-Dimension verstehen. Der UMAP-Algorithmus ordnet basierend auf allen diesen Schlagworten die Personen nach Ähnlichkeit an und bildet kleine lokale Cluster an Personen die sich besonders ähnlich sind. Ziel dieses Experiments war herauszufinden ob diese Technik zum einen interessante Ergebnisse bei unseren Daten liefert, zum anderen aber auch wie skalierbar die Technik ist was größere Datenmengen betrifft. Einerseits hat es gezeigt, dass für das Subset die Anwendung durchaus denkbar ist, aber dass insbesondere bei Live-Daten in der Größenordnung unseres Datensatzes es zur erheblichen Performanz-Schwierigkeiten kommen würde und eine Live-Berechnung unmöglich ist. Unabhängig davon führte dieses Visualisierungsexperiment aber auch zu einer entscheidenden Erkenntnis inhaltlicher Natur. In der Abbildung sind männliche Personen gelb gefärbt, während weibliche Personen cyan gefärbt sind. Durch die unterschiedliche Färbung wurde schnell deutlich, dass Schriftstellerinnen und Schriftsteller geschlechtlich getrennt geclustert werden. Dies beruht darauf, dass einige Schlagworte eine

geschlechtliche Unterscheidung besitzen, also z.B. "Schriftsteller"/"Schriftstellerin" obwohl es sich um die gleiche Berufsgruppe handelt. Durch die Visualisierung ist dem Projektteam diese Datenbeschaffenheit der GND-Daten das erste Mal aufgefallen und wurde so für das spätere Datenmodell weiter bedacht.

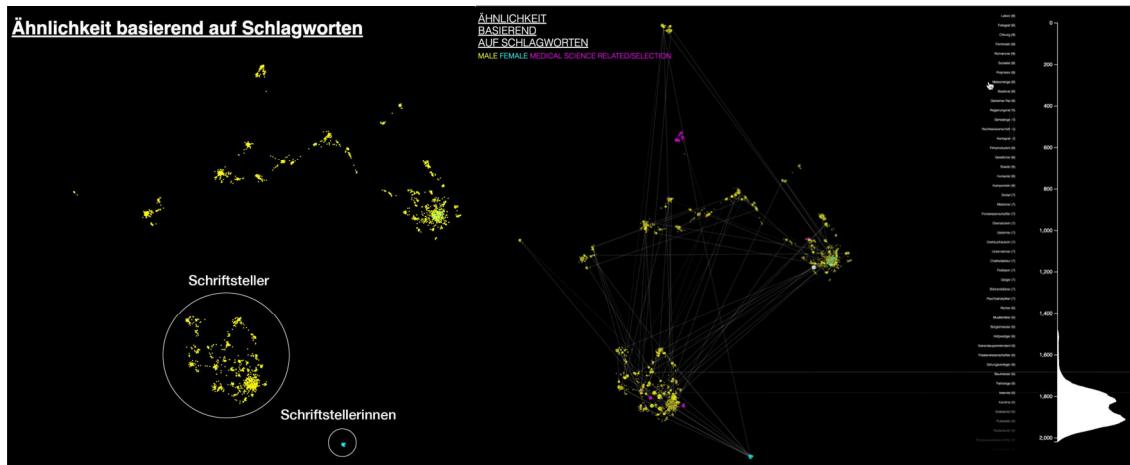


Abb.7: Eine Auswahl an Daten angeordnet nach Schlagwort-Ähnlichkeit unter Nutzung von UMAP. Gelb: männlich, cyan: weiblich, magenta: ausgewählte Filterung

Aufbauend auf der UMAP Anordnung wurde in einer weiteren Iteration zusätzlich die Darstellung der Zeit-Dimension experimentell integriert um Möglichkeiten zur Bewegung durch Zeit zu explorieren (Abb. 8). Die Darstellung und Auswahl von Zeit war bereits in den vorangegangenen Gesprächen und Workshops ein wichtiges Anliegen der HNA-Forscher*innen und viele existierende Tools bieten die Möglichkeit nach Zeit zu filtern oder Animationen durch Zeit abspielen zu lassen. Hier wurde gezielt eine Variante erprobt die eine neuartige Bewegung durch die Zeit, wie durch einen Tunnel erlaubt. Eine Zeitleiste auf der rechten Seite zeigt die allgemeine Verteilung aller Knoten an, während eine Liste daneben alle verbundenen Themenbegriffe, geordnet nach ihrem Vorkommen, enthält. Durch Scrollen können sich die Nutzer durch die zeitliche Dimension des Netzwerks bewegen, wodurch der Eindruck eines Zeittunnels entsteht. Knoten, die zu einem ausgewählten Jahr gehören, werden gelb dargestellt. Zeitlich nahe gelegene Knoten in der Vergangenheit erscheinen weiter vom Betrachter entfernt (sind kleiner) und sind in Rottönen markiert, während solche, die in der Zukunft liegen, in Grün- und Blautönen gefärbt sind und näher zu liegen scheinen (sind größer). Eine Erkenntnis, die wir mit Hilfe dieses Prototyps gewonnen haben, war, dass wir unser Datenmodell und unseren Verarbeitungsansatz noch einmal anpassen müssen um die Daten für die Verwendung in Visualisierungen zugänglicher zu machen, insbesondere im Hinblick auf die zeitliche Filterung. Das Konzept selbst kam visuell und vom Effekt her sehr gut bei Fachkonferenzen und in Sozialen Medien an und es gab Anfragen zur Offenlegung des Codes. Jedoch gab es auch Kritik, insbesondere von unseren eigenen HNA-Forscher*innen, dass der Effekt visuell und von der Interaktion ansprechend ist, jedoch einen wirklichen Erkenntnisgewinn durch die perspektivischen Verzerrungen erschwert. Da es aber in SoNAR aber in erste Linie um Forschungs-Ermöglichung und nicht um Unterhaltung geht, wurde dieser Ansatz nicht weiter verfolgt.

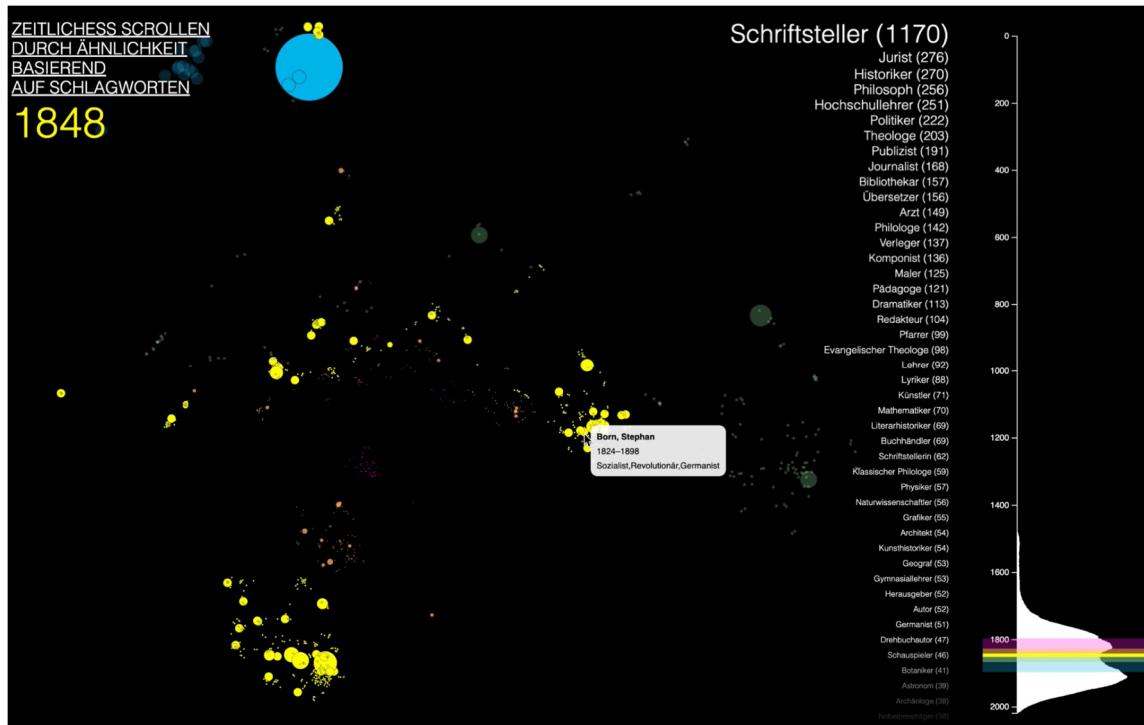


Abb.8: Zeitliche Bewegung durch die Daten unter Nutzung von Scrolling und Anordnung über Schlagwort-Ähnlichkeit

2.2.3 Zeitliche Netzwerk Kaskade nach Communities und Morph von einem Graph-Layout zu einer Zeitleiste (AP-3.2)

Abb. 9 zeigt weitere Ansätze zur Integration der zeitlichen Dimension und zusätzlich die Erprobung von Community Detection-Algorithmen. Dieses Experiment ist aus der Fragestellung entstanden ob Beziehungen in unseren Daten hauptsächlich zwischen Zeitgenossen stattfinden oder auch Zeitübergreifend. In einem ersten Schritt wurden hier zunächst Knoten basierend auf und geordnet nach den Lebensdaten (Geburtstag–Todestag) auf einer Zeitleiste ausgelegt und mit Bögen für Beziehungen verbunden (Abb. 9 links), je eine Zeile pro Person. In einem zweiten Schritt wurden ein Community-Algorithmus auf die Graphdaten angewendet. Daraufhin wurden die Knoten auf der Grundlage der Ergebnisse des Community-Algorithmus geordnet, eingefärbt und auf einer Zeitachse in einer Kombination mit der Anordnung nach Geburts- und Todesdaten platziert (Abb. 9 rechts).

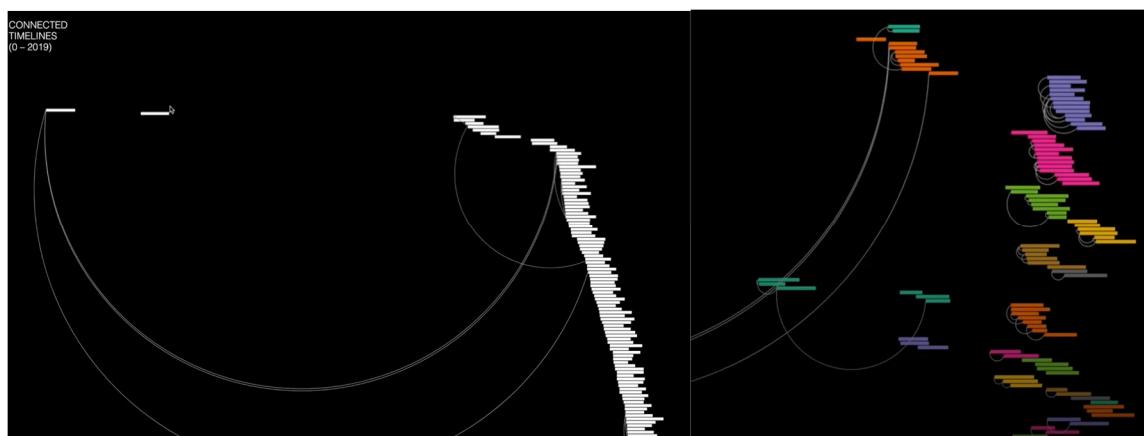


Abb.9: Anordnung von Knoten in einer Zeitleiste zunächst nur nach Geburstag (links) und danach geordnet und gefärbt nach Community-Cluster (rechts).

Da diese Darstellung zunächst positives Feedback hervorgebracht hat und auch gute Lesbarkeit vorweist, wurde der Ansatz iterativ noch weiter getrieben. In einem weiteren Schritt wurde nun versucht die Anordnung der Knoten durch eine Verbindung mit der Anordnung in einem Force-basierten Netzwerk zu erläutern und beide Ansichten miteinander durch sinnige Animationen und Übergänge zu verbinden. Zusätzlich wurden Linien zur besseren Lesbarkeit von Zeiträumen mit Jahreszahlen im Hintergrund ergänzt. Ergebnis ist ein Prototyp, der den fließenden Wechsel zwischen einem Graphen-Layout und einen Zeitleisten-Layout ermöglicht und so eine zusätzliche Perspektive auf Communities und zeitliche Verläufe ermöglicht (Abb. 10). In der Darstellung können z.B. nicht nur Personen mit Lebensdaten integriert werden, sondern beispielsweise auch Briefwechsel zeitlich abgebildet werden.

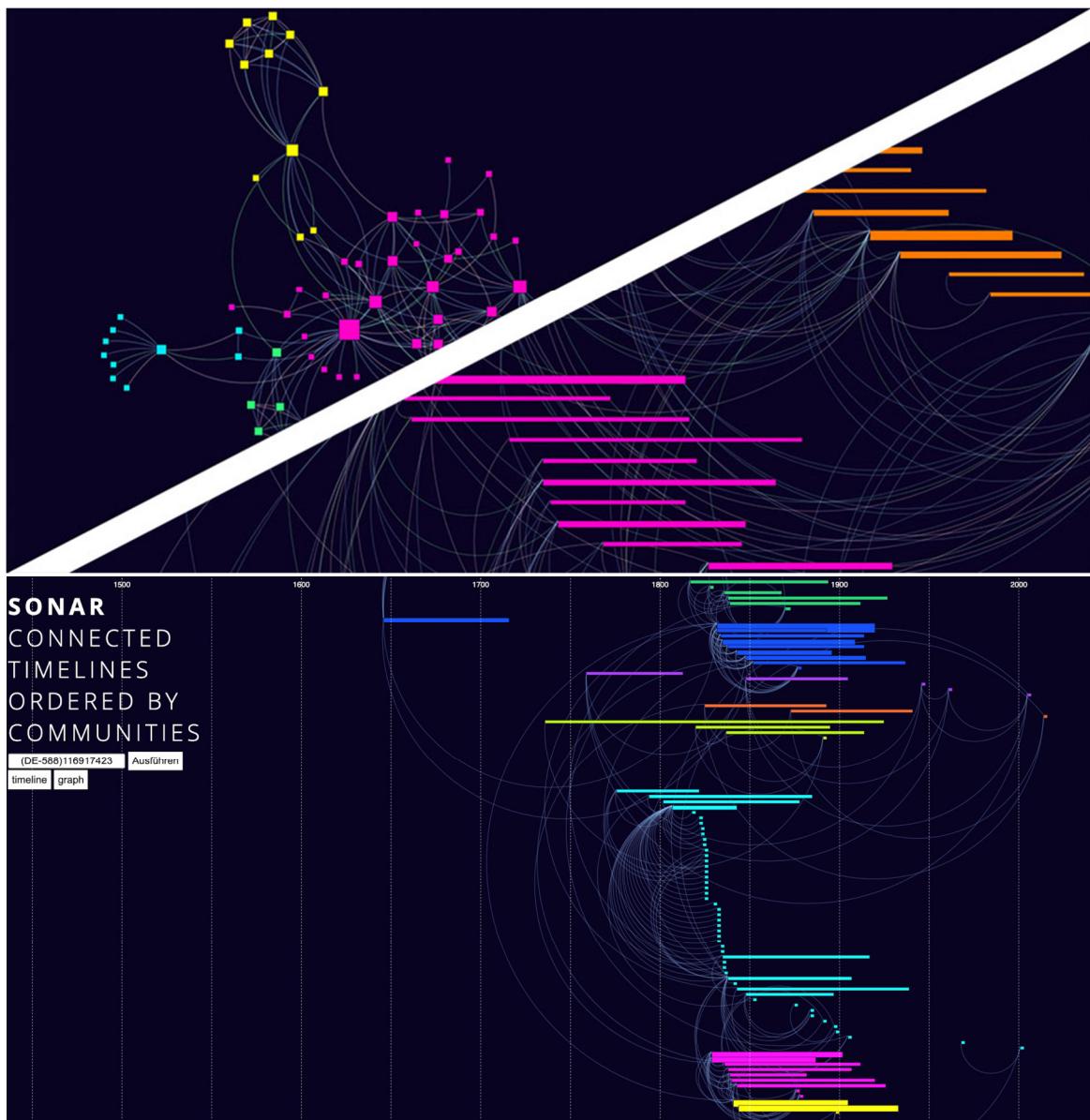


Abb. 10: Die Visualisierung ermöglicht einen animierten Wechsel zwischen einer Graph- und einer Zeitleisten-Visualisierung.

2.2.4 Bar Chart Race zu Personen-Schlagworten und Schlagwort-Voronoi nach Zeit (AP-3.2)

Um spielerisch die in den Daten meist vertretenen Personengruppen zu explorieren wurde anhand der GND Schlagwort-Verknüpfung eine Bar Chart Race Animation mit Hilfe eines dafür bereits bestehenden

Observable Notebooks²² erstellt (Abb. 11). Bar Chart Races sind Bar Charts mit zehn Balken, welche zeitabhängig Top 10 Listen wiedergeben und dabei auch Positionswechsel animiert darstellen (z.B. wertvollste globale Marken). Ziel war es, in einer relativ kurzen Animation schnell und spielerisch einen Überblick über meist vertretenen Personengruppen im Verlauf der Zeit zu bekommen und die Art der Personen verknüpften Schlagworte über die Zeit besser nachzuvollziehen.

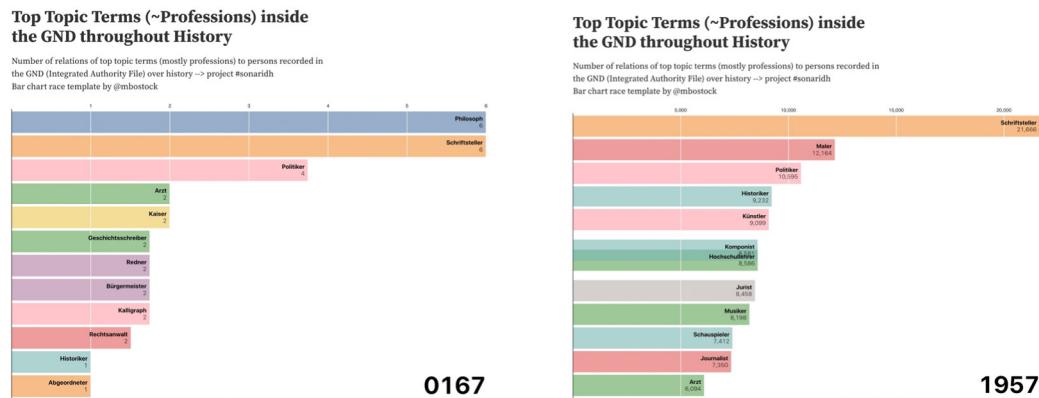
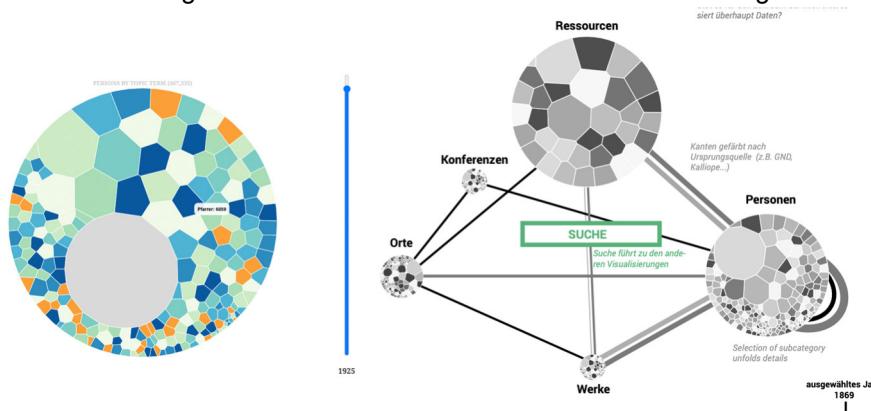


Abb.11: Bar Chart Race zu mit Personen verknüpften Schlagworten in den SoNAR Daten.

In einem nächsten Schritt wurde die Ordnung nach Häufigkeit und Nutzung von Top-Listen aufgegriffen und auf ein Voronoi-Diagramm (Abb. 12 links) übertragen, bei welchem durch einen Zeitschieber die Zeitauswahl der Top 200 Schlagworte (alle weiteren Schlagworte werden unten Sonstige zusammengefasst in einem grauen Feld) nun selbst ausgewählt werden können. Die Kreisgröße spiegelt dabei die Gesamtmenge an enthaltenen Daten wieder für ein Jahr und einzelne Felder stellen den Anteil der meist verknüpften Schlagworte (in der Regel Berufe) dar. Zusätzlich wurden weibliche Berufsbezeichnung orange gefärbt um diesen Aspekt erneut neu zu beleuchten. Beim Hovern über ein Feld lässt sich der Name des Schlagworts und die Anzahl des Vorkommens ablesen. Ziel war hier einen Überblick über einen spezifischen Aspekt zu geben und mehrere Jahre vergleichbar zu machen, so dass Forschende z.B. einfach erkennen können, ob eine bestimmte Berufsgruppe häufig in einem ausgewählten Jahr vorkommt, auf welchen Berufsgruppen der Fokus der Daten liegt und wie viele Daten es überhaupt pro Zeitraum gibt.

Aufbauend auf dieser Visualisierung wurde anschließend ein Konzept entwickelt (Abb. 12 rechts), wie diese Art der Darstellung von Personen-Schlagworten auch auf andere Knotentypen (Geografika, Körperschaften, etc.) übertragen werden könnte und wie durch eine zusätzliche Anzeige von Relationen zwischen diesen eine Übergangsansicht konzipiert werden kann, die Forscher*innen helfen kann den Umfang der Daten und die Relevanz der Daten für die eigene Forschung besser zu erkennen.



²² <https://observablehq.com/@d3/bar-chart-race-explained>

Abb. 12: Die Abbildungen zeigen die Nutzung von Voronoi Charts zur Darstellung von Anteilen.

2.2.5 Geographische Anordnung von Knoten (AP-3.2)

Zunächst wurden die geographischen Koordinaten bei der Datentransformation nicht korrekt transformiert, weshalb die Nutzung örtlichen Datendimensionen zunächst nur über Filterung ermöglicht werden konnte. Zwar ist die geographische Darstellung wichtig für die HNA-Forscher*innen und wird auch ähnlich wie Zeitleisten häufig in bestehenden Tools angeboten, gleichzeitig ist hier aber auch die zeitliche Dimension entscheidend. So ist es bei der geographischen Anordnung auch der Zeitpunkt wichtig: Wann war eine Person wo als welche Beziehung bestanden hat.

Während bei Briefen unter anderem Ort des Absender und Empfängers sowie Zeitpunkt häufig bekannt sind, wodurch sich Orts-Zeit-basierte Netzwerke gut generieren lassen, so gibt es zu einem Großteil der Beziehungen in den SoNAR-Daten (z.B. aus der GND) keine zeitlichen und örtlichen Angaben und örtliche Angaben lassen sich häufig nicht zeitlich zuordnen. Deshalb wurde zumindest in Rahmen dieses Projektverlaufes nur eine generelle örtliche Anordnung mit den Daten erprobt (Abb. 13). Hierzu wurden die geographischen Koordinaten der Ortsangaben zu Personen genutzt um diese auf einer Karte mit Hilfe des Services von DataWrapper²³ auszulegen. Prinzipiell ist aber auch die Nutzung von Visualisierungslibraries wie D3.js oder Leaflet.js problemlos denkbar. Die Tatsache, dass der Großteil der Daten auf den deutschsprachigen Raum fokussiert ist, lässt sich hier gut erkennen.

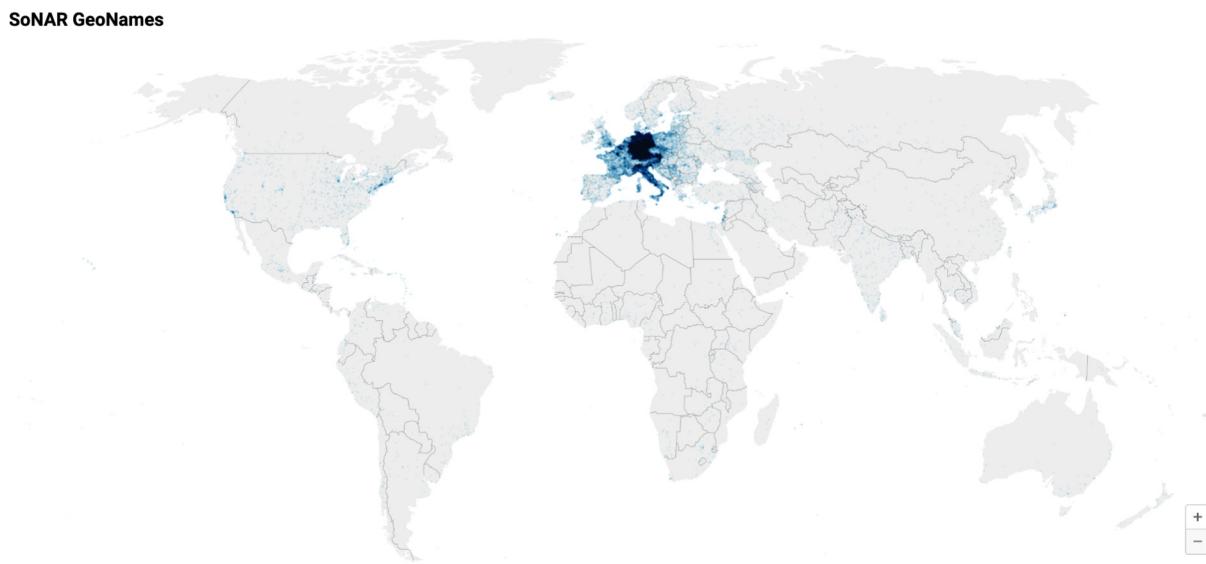


Abb.13: Alle mit Personen verknüpften Orte in den SoNAR-Daten eingezeichnet in eine Weltkarte. Es ist ein klarer Schwerpunkt bei Deutschland zu erkennen.

²³ <https://www.datawrapper.de/>

2.2.7 Entfaltung von Kanten (AP-3.2)

[siehe Publikation “Unfolding Edges for Exploring Multivariate Edge Attributes in Graphs” (Bludau *et al.*, 2021)]

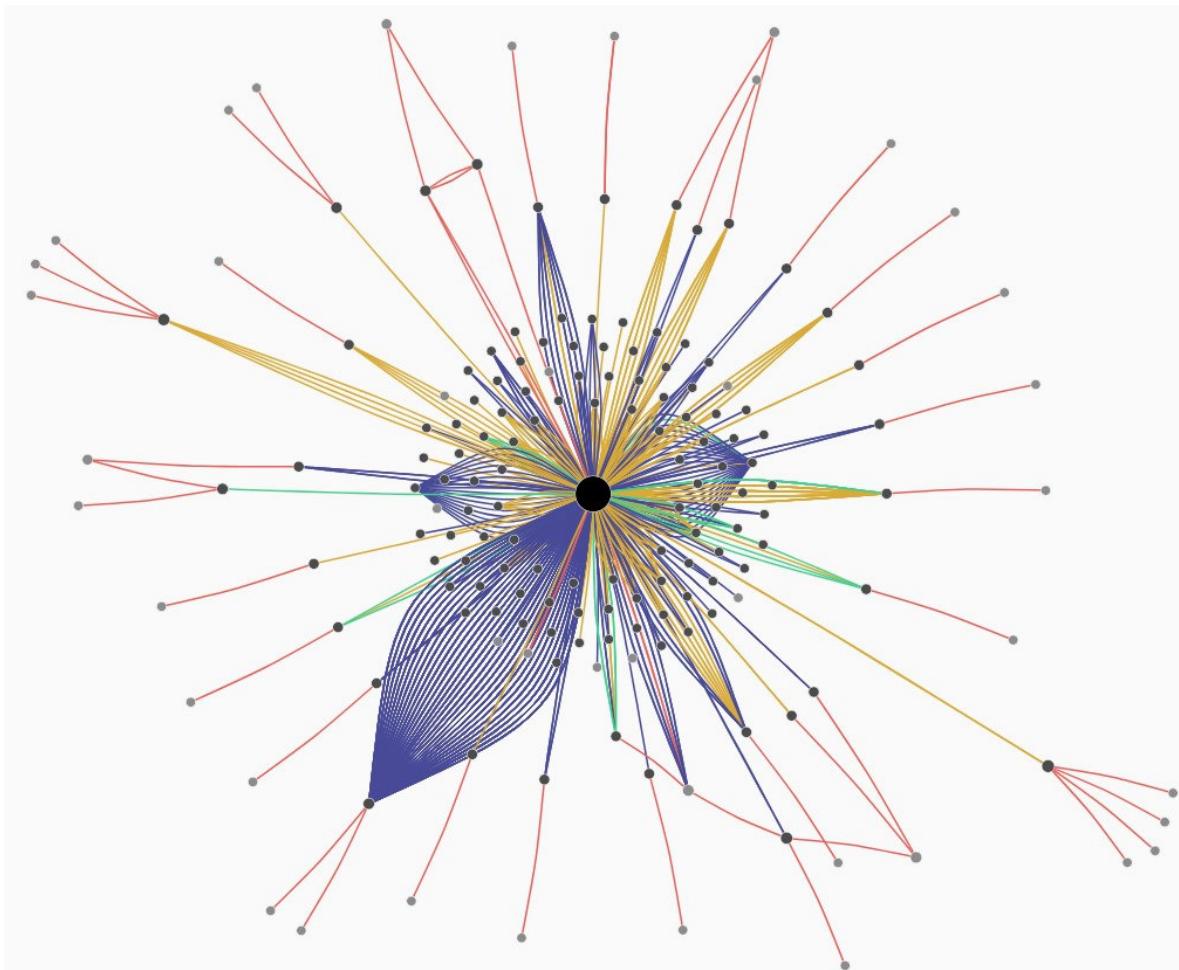


Abb.14: Eine beispielhafte Visualisierung alle Kanten in einem Ego-Netzwerk mit farblicher Kodierung führt zu einer Vielzahl an visuellen Überlagerungen.

Häufig beruhen Kanten in historischen Netzwerken auf Rekonstruktionen von Beziehungen basierend auf Ressourcen wie Briefen, Tagebüchern, Reiseberichten, Protokollen oder Publikationen. Eine Herausforderung besteht in der Visualisierung von Provenienzen, da viele Beziehungen zwischen Akteuren auf mehreren Quellen basieren können, was durch die dargestellten Details zu einer Komplexitätserhöhung führen kann. Ein SoNAR-Netzwerk, in dem alle Beziehungen jeweils in unterschiedlichen Farben je nach Ursprungsquelle dargestellt werden, führt zum Beispiel häufig zu vielen Kantenüberlappungen (Abb. 14). Bei dem Konzept der Kantenentfaltung besteht der Ansatz darin, die Kanten zwischen zwei Akteuren für eine übersichtlichere Ansicht zunächst zu gruppieren (Abb. 15a), und nur bei Bedarf per Auswahl zu entfalten (Abb. 15b) (Brüggemann *et al.*, 2020; Bludau *et al.*, 2021). So lassen sich gezielt Details anzeigen und Quellen verlinken.



Abb.15: Durch die iterative Entfaltung von Kanten werden in einer Grundansicht (a) mehrere Verbindungen zwischen zwei Personen zusammengefasst und die Anzahl dahinterstehender Beziehungen durch Kantenstärke dargestellt. Auswahl einer Kante führt zu bedarfsabhängigen Entfaltungen von Beziehungen (b), welche durch den gezielten Fokus Enkodierungsmöglichkeiten (z.B. für Unsicherheiten oder Kategorien) ermöglichen.

2.2.6 Ressourcen-Fokus im radialen Layout (AP-3.3)

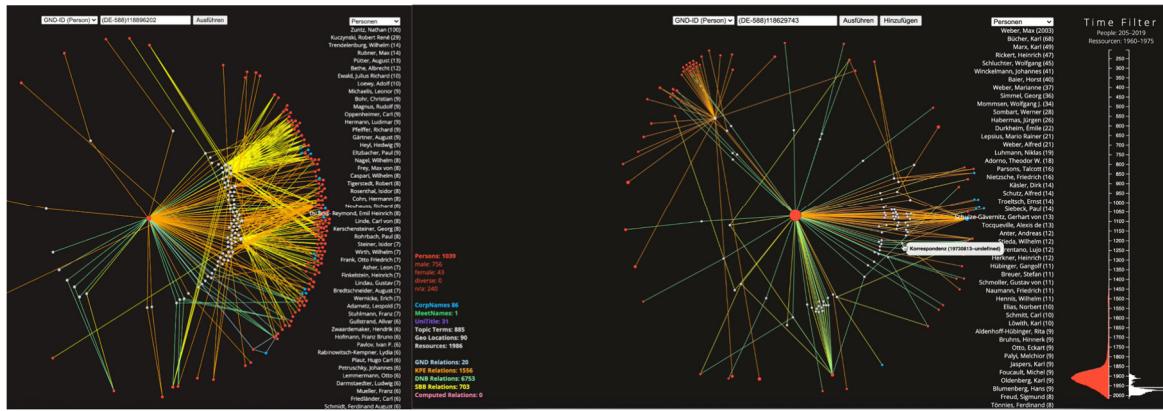


Abb.16: Eine radiale Visualisierung ordnet mit einer ausgewählten Person verknüpfte Ressourcen (z.B. Briefe) in einem inneren Ring und verknüpfte Personen und Körperschaften in einem äußeren Ring nach Zeit filterbar an.

In den SoNAR-Daten sind viele unterschiedliche Knoten-Typen modelliert (z.B. Personen, Geografika, Ressourcen, Schlagworte, Körperschaften, etc.) wodurch sich nicht nur Personen-Netzwerke generieren lassen, sondern auch Netzwerke die andere Verbindungen z.B. zwischen Personen und Ressourcen wie Publikationen zulassen. Viele Fragestellungen, z.B. Fragestellungen zu Zitationsnetzwerken können dabei von Visualisierungen auch dieser anderer Knoten-Typen profitieren. So sind z.B. für manche Fragestellungen nicht nur Personen relevant, sondern auch verbindende Ressourcen (z.B. Briefe, Publikationen). Um Fragestellungen wie diese aufzugreifen wurde eine radiale, zeitlich filterbare Netzwerk-Visualisierung erarbeitet, welche Personen-Ressourcen-Personen-Beziehungen von einer ausgewählten Person aus darstellt (Abb. 16). In dieser egozentrischen Darstellung basierend auf der Auswahl einer Person wird im Zentrum die ausgewählte Person als Knoten angezeigt. Weiterhin werden mit der Person verknüpfte Ressourcen (weiß) in einem inneren Ring und verknüpfte Personen (rot) und Körperschaften (blau) in einem äußeren Ring angezeigt. Beziehungen werden durch Linien dargestellt, farblich abhängig von der Ausgangsquelle (z.B. Kalliope oder DNB). Zusätzlich sind sowohl die Personen über Lebensdaten, als auch die Ressourcen über die Datierung filterbar. Ein Doppelklick auf eine Person generiert ein neues radiales, egozentrisches Netzwerk um die neu-ausgewählte Person. Ziel der Ansicht ist es gewesen herausstechende Ressourcen (z.B. Publikation) sichtbar zu machen, welche z.B. besonders viele andere Personen mit der ausgewählten Person verbinden oder auf Ressourcen basierende Cluster offenzulegen. In einer projektinternen Abstimmung wurde zunächst sich darauf verständigt, soziale Personen-Netzwerke in den Fokus zu stellen, weshalb dieser Ansatz, auch wenn er für zukünftige ergänzende Ansichten vielversprechend scheint, zunächst auf Grund der begrenzten Zeit nicht weiterentwickelt wurde.

2.2.8 Datenexplorations-Interface (AP-3.3)

Einer der wichtigsten Bestandteile des Prototyping-Prozesses war die Entwicklung eines Datenexplorations-Interfaces. Zu Beginn der Projektes war es zunächst für die Beteiligten Projektpartner*Innen unklar, was die Daten genau enthalten und für AP-2 (HHU) war es insbesondere auch unklar welche Möglichkeiten für Case-Studies bestehen. Zudem galt es das Datenmodell zu entwickeln, zu erproben und dabei auf einer Nutzbarkeit in Visualisierungen zu achten und auch mögliche Fehler im Transformations-Prozess ausfindig zu machen. Um die Daten möglichst umfänglich explorieren zu können wurde dafür ein Explorations-Tool angefertigt, dessen Funktionen nach und nach bedarfsabhängig in Absprache mit und basierend auf Wünschen von AP-2 erweitert wurden. Zunächst ging es nur darum Netzwerke über ein Suchfeld mit Hilfe basierend auf der SoNAR-Datenbank generieren zu können und eine zeitliche Verteilung sowie Übersicht über einzelne Kategorien wie verknüpfte Schlagworte oder Orte darzustellen, sowie bei Selektion von Knoten alle verfügbaren Metadaten zu dem Knoten anzuzeigen. Dies hatte zur Folge, dass AP-2 die Daten somit auch mit den

Ausgangsdaten vergleichen konnte und Schlüsse für das Forschungsdesign rausziehen konnte. Zudem wurden einige Auffälligkeiten, Probleme oder Fehlstellen im Datenmodell oder der Transformation ausfindig gemacht, wodurch iterative Anpassungen durchgeführt werden konnten. Weiterhin wurde mit unterschiedlichen Darstellungsformen experimentiert, z.B. die Nutzung von geraden Kanten im Vergleich zu gewölbten Bögen als Kante. Weiterhin wurden Anpassungen gemäß der Entwicklung des Datentransformations-Prozesses gemacht, so waren zu Beginn der Entwicklung des Explorationsinterfaces zum Beispiel in unserem Datenmodell noch keine abgeleiteten Beziehungen vorhanden, weshalb Ressourcen als Zwischenschritt geladen werden mussten, was die Queries komplexer und zeitintensiver gemacht hat und zudem die Visualisierung überladen hat (Abb. 17 oben).

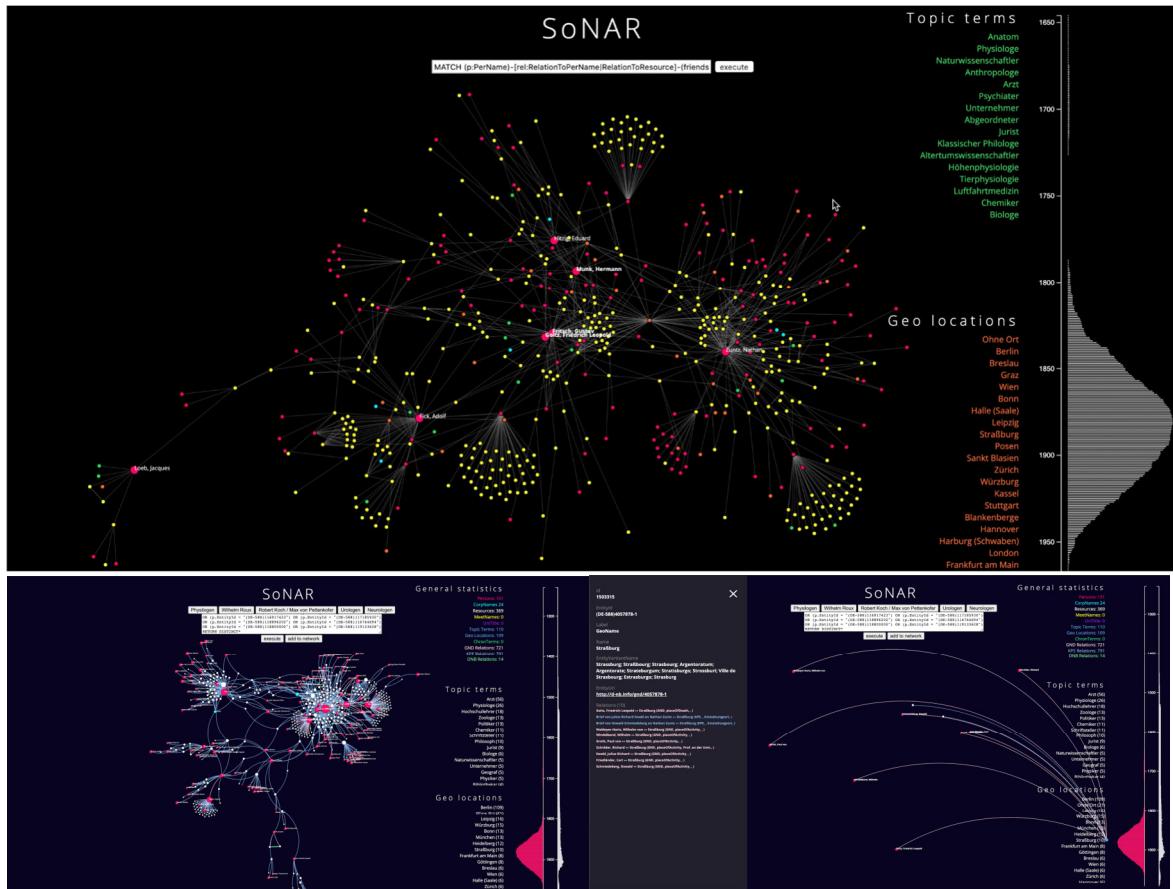


Abb.17: Erste Varianten und Iterationen eines Daten-Explorationsinterfaces welches mit der Datenbank verbunden ist. Zunächst wurden alle Knoten-Typen, d.h. auch verbindende Ressourcen (gelb), dargestellt (oben). In einem späteren Schritt (unten) wurde bei der Datentransformation unter Nutzung eines erarbeiteten Regelsets aus Ressourcen (z.B. Brief zwischen zwei Personen) daraufhin direkte Beziehungen abgeleitet. Anschließend wurde auf direkte Darstellung von Ressourcen in der Visualisierung zur Fokussierung auf Personen-Netzwerke verzichtet.

Dieses Explorationsinterface wurde nach und nach iterative mit Funktionen erweitert (Abb. 17 unten). So wurde eine Zeit-Filterung eingebaut die sowohl Filterung über Lebensdaten von Personen als auch auch eine zeitliche Filterung über Beziehungen ermöglicht, weil die Filterung über Beziehung allein auf Grund der geringen Menge an Relationen die mit einem Zeitattribut versehen sind für die HNA-Forscher*innen nicht befriedigend war. Zusätzlich wurden nach und nach Filter-Möglichkeiten nach Attributen erweitert, z.B. nach Beziehungstyp (z.B. Familie) oder nach verknüpften Körperschaften (z.B. verknüpfte Universitäten). Weiterhin wurde die Generierung von Netzwerken anhand von Schlagwörtern integriert (Abb. 18 oben links), sowie Funktionen um weitere Knoten zu einem Netzwerk über Suche oder die Erweiterung eines Netzwerk über Doppelklick auf einen im Netzwerk enthaltenen Knoten hinzuzufügen (Abb. 18 oben rechts). Außerdem wurde für eine bessere Übersicht auch die Kanten-Entfaltung in die Ansicht integriert (Abb. 18 unten).

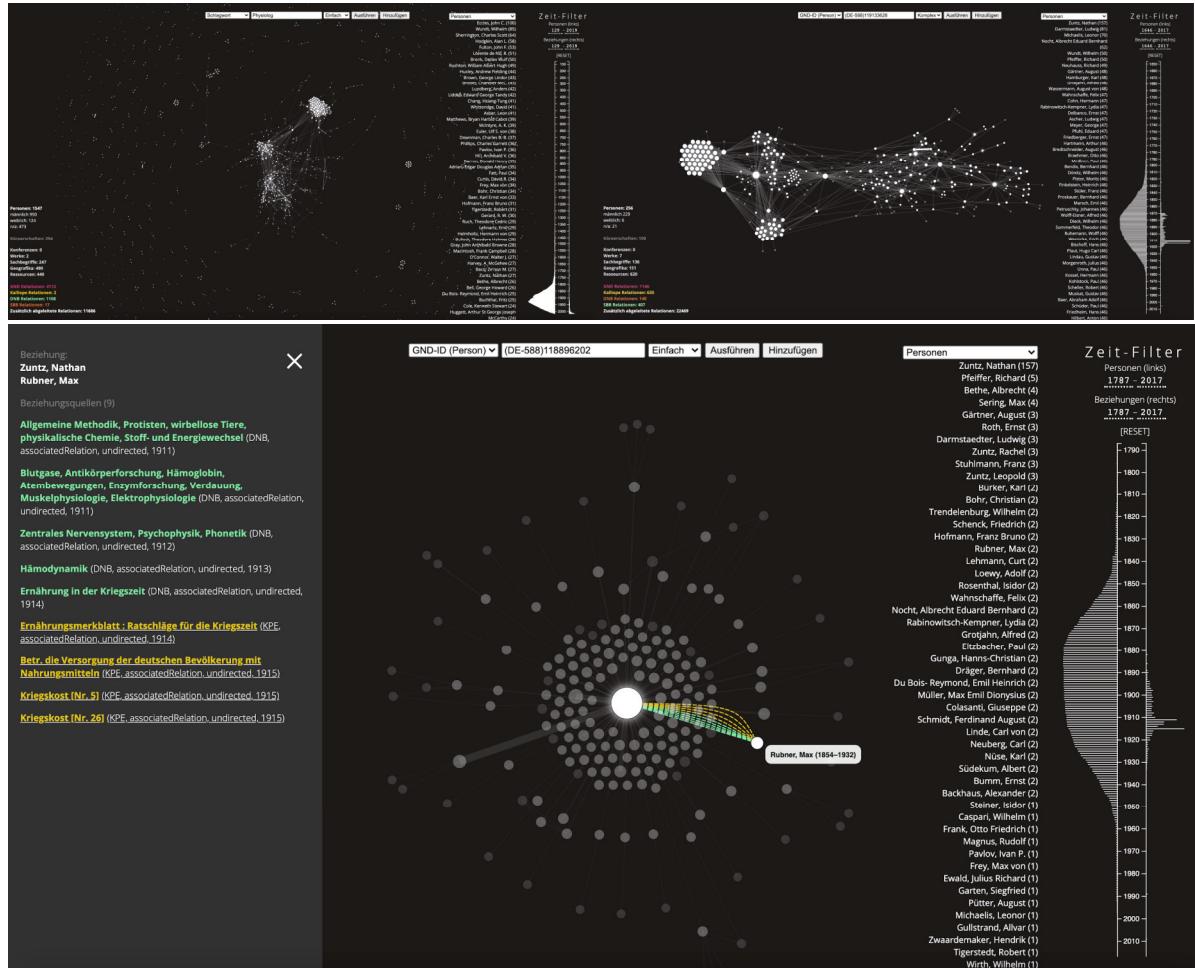


Abb.18: Letzte Iteration des entwickelten Explorationsinterfaces.

3. Schlussfolgerungen aus iterativen Prozess mit AP-2, Workshops und AP-4 Ergebnissen

Der Projektverlauf hat nach und nach wichtige Erkenntnisse hervorgebracht für gute Nutzungsbedingungen für HNA-Forschung bei einer potentiellen SoNAR-Technologie. Eine wichtige Erkenntnis war zuallererst die Notwendigkeit von größtmöglicher Transparenz aller Prozesse der Daten-Nutzungs-Pipeline – von der Datentransformation über die visuelle Aufbereitung der transformierten Daten durch Visualisierungen bis zur möglichen Nachnutzung der Daten und Offenlegung der Prozesse. Hier hat sich Transparenz bezüglich der Daten, Datenprovenienzen, Analysemethoden, Algorithmen und der Visualisierung als grundlegende Anforderungen an ein wissenschaftlich nutzbares Visualisierungswerkzeug herausgestellt und war durch alles Projektprozesse durchweg ein immer wiederkehrender Diskussionspunkt:

“Was können die Daten? und was können sie eben nicht?” (Co-Design Workshop AP-3)

„[...] da würde ich gerne wissen wo diese Informationen herkommen und wie die eigentlich miteinander verknüpft werden und wer auf die Idee gekommen ist das so zu tun.“ (Interview AP-4)

„[...] wenn so eine standardisierte Datenbank aufgebaut wird, werden so viele Entscheidungen getroffen [...] und da muss man wissen, wie kamen die zu dieser Zahl und wie viele alternative Quellen oder wie viele alternative mögliche Jahreszahlen wurden nicht berücksichtigt [...] Es

muss alles genau dokumentiert sein oder es sollte bestenfalls alles genau dokumentiert sein.“
(Interview AP-4)

„[...] wenn ich diese Daten nutze ... muss ich sichergehen können, dass die wissenschaftlich Hand und Fuß haben oder eben auch Rückfragen stellen können - also wo kommen diese Daten her?“ (Interview AP-4)

„Das Wichtigste ist halt immer transparent zu sein, auch wieder, was man gemacht hat, dann könnte auch jemand anderes sagen: ja gut. Mittelwert, Sie haben den Mittelwert genommen, ich denke das ist falsch, ich mach das anders.“ (Nutzer*innenstudie AP-4)

Zusätzlich zu Notwendigkeit von größtmöglicher Transparenz, haben die Use-Cases und der Anforderungskatalog von AP-2 (HHU), Workshops, Nutzungsstudien und Feedback bei Projektpräsentation bei Konferenzen oder dem internationalen SoNAR-Workshop alle ein deutliches Bild abgegeben, dass es im Bereich der HNA-Forschung offensichtlich zwei grobe Gruppierungen gibt, was die Nutzung betrifft (siehe Projektberichte AP-4 HU). Es gibt eine erste Gruppe für die eine freie, interaktive Exploration von Daten über mehrere anpassbare Perspektiven hinweg (z.B. Überblicke, zeitbasiert, egozentrisch) Potential für Forschung zeigt, und die sich vorstellen kann explorativ an Fragestellungen heranzugehen oder sogar neue Fragestellungen daraus zu entwickeln:

“Was man dann für Beziehungen in den Daten sucht, das passiert ganz oft erst in dem Moment wo du das erste mal auf den Haufen drauf guckst.“ (Co-Design Workshop AP-3)

“[...] für sowas finde ich es spannend. um zu gucken, wo gibt es so periphere Akteure, auf die man sonst gar gekommen wäre. Weil die in der normalen Betrachtung immer außen vor gelassen werden.“

(Nutzer*innenstudie AP-4)

Andererseits gibt es auch Forscher*innen, die mit einer generellen Grundskepsis an Visualisierungen herangehen, die selbst hauptsächlich quantitativ mit den Daten arbeiten bzw. arbeiten wollen und deren Interesse insbesondere an einem einfachen Zugang zu den Daten besteht mit dem Ziel selbst quantitativ damit arbeiten zu können. Diese Nutzer*innengruppe hat auch noch mehr als die andere Gruppe ein großer Interesse an größtmöglicher Transparenz was die vorausgegangen Datenverarbeitungsprozesse und genutzte Algorithmen betrifft. Zudem ist für diese Gruppe beim Anbieten einer Visualisierung wichtig, dass sich einzelne Parameter (z.B. der genutzte Force-Algoritmus oder Cluster-Algorithmen) idealerweise selbst einstellen lassen und Netzwerkmaße generell ablesen lassen:

“[...] normalerweise [entwickle ich] zuerst die Fragestellung, bevor ich die Daten sehe. Weil ansonsten ist natürlich ein riesiges Problem, dass ich ein Bias kriege. [...] Netzwerkvisualisierungen sind, je nachdem, wie ich [sie] mache, sehr unterschiedlich aussagekräftig [...]” (Nutzer*innenstudie AP-4)

“[...] also ohne Netzwerkmaße gehts ja nicht. Also ich kann ja keine Visualisierung interpretieren, wenn ich die Maße nicht habe. Also dann, ich meine ich sehe, ob die Relation dicker oder dünner ist, daraus kann ich sicherlich schließen ... da ist eine engere Beziehung und da ist eine lockere, aber ich kann es nicht interpretieren.“ (Nutzer*innenstudie AP-4)

Beide Nutzer*innengruppen gilt es bei einem Konzept zu berücksichtigen. Zudem wurde immer wieder von den Forscher*innen aus AP-2 sowie von externen Personen der Wunsch nach einer größtmöglichen Auswahl- und Kombinationsmöglichkeiten an Filterungen geäußert, so dass

idealerweise alle verfügbaren Parameter filterbar sind um größtmögliche Flexibilität für eine Vielzahl von Forschungsfragen zu ermöglichen:

„Filtermöglichkeiten. Filtermöglichkeiten sind so das Entscheidende, glaub ich.“
(Nutzer*innenstudie AP-4)

“Aber wenn man das dann tatsächlich nochmal filtern könnte und sich tatsächlich nur noch die wissenschaftlichen Arbeiten [...] anzeigen lassen könnten und dann eben sehen könnte, mit welchen anderen Wissenschaftlern in diesem Netzwerk diese wissenschaftlichen Arbeiten in irgendeiner Weise verbunden gewesen sind.” (Nutzer*innenstudie AP-4)

*“Forscher*in möchte Merkmalsausprägungen als Filter für Graphen: z.B. Geschlecht, Alterskohorte, Herausgeberschaft, Beruf, Affiliationen, etc.”* (Anforderungskatalog AP-2)

Insbesondere die Filterung nach Zeit ist für die historische Netzwerkanalyse dabei als entscheidend herausgestochen:

“Ohne Timelines nutzen mir die Visualisierungen nichts – weder für die Analyse noch für die Vorstellung von Ergebnissen. Das heißt für die Rechercheweg und die Analyse brauche ich schon die Visualisierung in der Form das ich vor- und zurückfahren kann. Ohne das macht es keinen Sinn. Wenn ich ein Netzwerk über 100 Jahre anzeige, okay, da kann ich vielleicht irgendwas erkennen, aber es ist vollkommen artifiziell.” (Workshop AP-3)

*“Forscher*in möchte Netzwerkveränderungen in Zeitschnitten visualisieren können.”*
(Anforderungskatalog AP-2)

4. Konzept (AP-3.4)

[siehe Publikation “Was sehe ich? Visualisierungsstrategien für Datentransparenz in der Historischen Netzwerkanalyse”]

Aufbauend auf Literaturrecherche und Recherche von vorangegangen Projekten, sowie den Erkenntnissen auf dem iterativen und kollaborativen Prototyping-Prozess, den Anforderungen von AP-2 und Erkenntnissen durch Interviews und der Nutzer*innenstudie wurde ein Visualisierungskonzept entwickelt welches sämtliche Erkenntnisse synthetisiert und in ersten funktionalen Prototypen zusammenfasst. Auf Grund der zunächst begrenzten Projektlaufzeit, wurde hier insbesondere der Fokus auf die Erprobung neuartigerer Ansätze gelegt, und bei der Implementierung bereits etablierte Mechanismen (z.B. Suche, multi-facettierte Filterung bei Expertensuche) zunächst in den Hintergrund gestellt. Aus der sich in der Projektarbeit ergebenden Forderung nach Datentransparenz haben wir folgende **Designziele (DZ)** für interaktive HNA-Visualisierungen abgeleitet:

DZ1) Aufnahme und Kommunikation von Datenprovenienzen: Um die Datentransparenz auch nach Datentransformation und Zusammenführungen sicherzustellen, müssen Datenprovenienzen über Merkmale bei Knoten und Kanten unbedingt erhalten werden und über URIs auf die Ausgangsdaten verweisen.

DZ2) Dokumentation vorausgegangener Prozesse: Die konkreten Schritte der Datentransformationen und Anwendungen von Algorithmen für Visualisierungen müssen inklusive Code nachvollziehbar und frei verfügbar dokumentiert werden, um Vertrauen zu schaffen und reproduzierbare Ergebnisse sowie kritische Auseinandersetzungen zu ermöglichen.

DZ3) Offenhaltung der Interpretierbarkeit der Daten: Datenunsicherheiten und unterschiedliche Granularitätsstufen müssen für spätere Interpretation in den Daten erhalten bleiben und dürfen nicht durch Normalisierungen entfernt werden. Auch in Visualisierungen müssen Kodierungen verwendet

werden, die fachspezifische Einschätzungen und Interpretationen erlauben. Um unterschiedlichste Forschungsfragen beantworten zu können, muss dazu eine Vielzahl an An- und Übersichten mit bedarfsabhängigen Graden an Fokus und Detail bereitgestellt werden, die das Potential der Daten sowie Fehlstellen und Unsicherheiten offenlegen.

DZ4) Unterstützung von Folgeforschung: Zugriff auf die Datenquellen (z.B. Dokumente, Briefe, Publikationen) müssen direkt in der Visualisierung über URLs verfügbar sein, um weitere Recherchen zu ermöglichen. Visualisierungsergebnisse, spezifische Ansichten und die Daten selbst müssen speicherbar, zu verlinken und reproduzierbar sein.

Interviews mit potentiellen Nutzer*innen und HNA Expert*innen durch AP-4 haben zudem grundsätzlich zwei größere Nutzer*innen-Gruppen identifiziert. Um den vielfältigen Anforderungen an offene Exploration, gezielte Abfragen, spezifischen Datenanalysen und angemessene Datentransparenz dabei zu begegnen, sieht unser Konzept zwei Stränge vor:

1. **eine webbasierte Visualisierung** (siehe Sektion 5) und
2. **ein Jupyter Notebook HNA Curriculum** (siehe Sektion 6)

5. Webbasierte Visualisierung (AP-3.4)

Aufbauend auf dem vorausgegangen Prototyping Prozess wurden verschiedene Ansichten entwickelt, die eine Exploration der Daten aus unterschiedlichen Perspektiven und mit Fokus auf eine Vielzahl von Daten-Dimensionen ermöglichen (Whitelaw, 2015; Dörk *et al.*, 2017). Die webbasierte Visualisierung ist dabei in zwei Hauptbestandteile aufgeteilt:

1. Ein **Datenübersichts-Einstieg**, womit Forschende einen Eindruck bekommen können ob SoNAR für ihr spezifisches Forschungsfeld relevant sein könnte.
2. **Suchbasierte Ansichten**, welche auf gezielten Suchanfragen nach einer Entität basieren, unterschiedliche Perspektiven auf die Daten ermöglichen und detaillierte Filterungsmöglichkeiten anbieten.

Entstanden sind damit Ansichten, die als Zugang eine gesamtdatenbasierte, akkumulierte Überblicks-Ansicht als Einstieg nach dem Prinzip „Overview first [...] then details-on-demand“ (Shneiderman, 1996) mit einzelnen suchbasierten Ansichten nach dem Prinzip „Search, Show Context, Expand on Demand“ (Van Ham and Perer, 2009) kontrastieren. Hier geht man von etwas kleinem (einer Suchanfrage) aus, und kann diese bei Bedarf explorativ erweitern.

Unabhängig von den Visualisierungsdetails der einzelnen Ansichten war es zudem wichtig eine Kohärenz zwischen den Ansichten zu erstellen um eine intuitive Nutzung zu ermöglichen und Frustration durch unerwartete Interaktionen oder Datenkodierungen zu vermeiden. Kohärenz wurde dabei durch eine Vielzahl von generellen Design-Entscheidungen hergestellt:

- Positionierung von Elementen bleibt über alle Ansichten gleich (z.B. Suchschlitz immer oben mitte, Zeitleiste immer unten, Filter immer rechts)
- konstante Interaktionen über Ansichten hinweg (z.B. Pan & Zoom Interaktionen zum Reinzoomen in Details, Klick zur Auswahl und Detailerweiterung)
- konstante Farbvariablen: über die ganze Visualisierung hinweg gibt es eine konstante Farbgebung von Design-Elementen. Die Daten sind sehr divers und bieten eine Vielzahl an möglichen Daten-Attributen und Kategorien die sich über Farbkodierung unterscheiden ließen. Um eine Überladung durch zu viele unterschiedliche Farbkodierungen zu vermeiden wurde

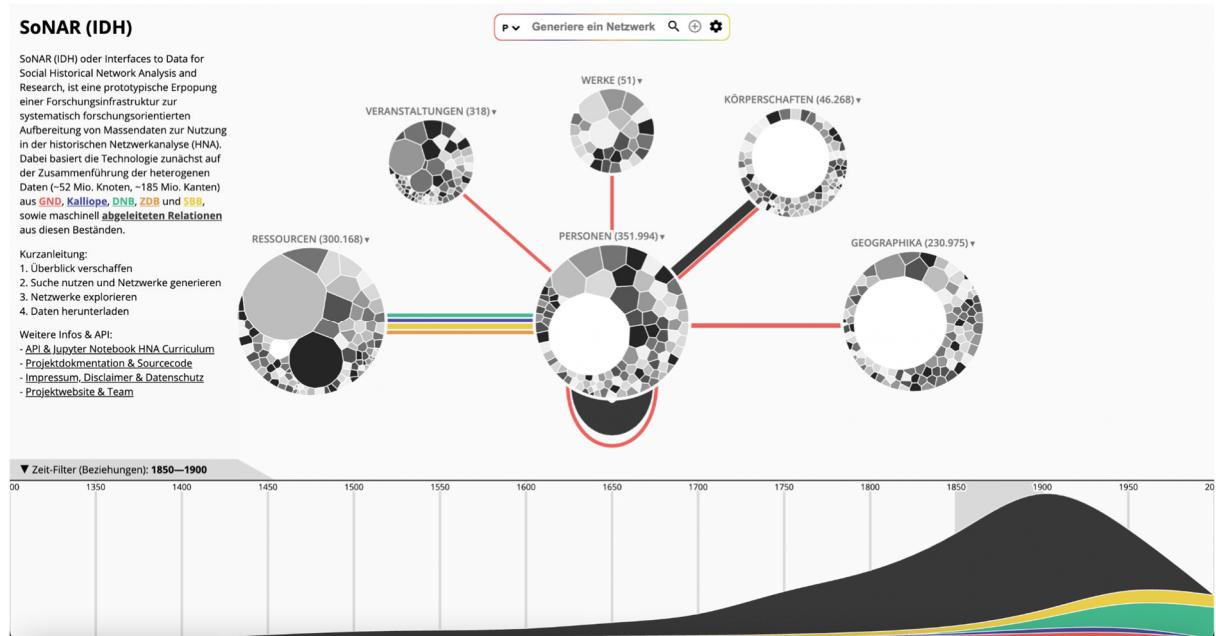
Farbe in erster Linie zur Unterscheidung der Datenquellen verwendet (z.B. rot für GND, blau für Kalliope, grün für DNB, usw.)

- es wurde gezielt auf eine möglichst schlichte Darstellung von Informationen zurückgegriffen, mit einem hellen Hintergrund der sich für Exporte der Visualisierung für Ausdrucke auch komplett weiß färbt

In den folgenden Abschnitten werden nun die einzelnen Visualisierungsansichten beschrieben, erläutert und deren Konzept hergeleitet.

5.1 Übersichts-Zugang

Die Startseite der webbasierten Visualisierung zeigt die übersichtsbasierte Ansicht (Abb. 19) welche dazu dient, die Relevanz für die Bearbeitung von Forschungsfragen ermitteln zu können, und um darzulegen, auf welchen Daten die Visualisierungen basieren. Damit bietet sie eine zeitlich filterbare Meta-Perspektive auf ca. 52 Mio. Knoten und 185 Mio. Kanten ab. Ziel ist es bereits vor der Eingabe von Suchbegriffen einen Eindruck über die Daten vermitteln zu können, Datenprovenienzen offenzulegen und statistische Aussagen über die zeitliche Verteilung zu geben. Durch die Ansicht soll der Ansatz von “*Overview first*” (Shneiderman, 2008) verfolgt werden. Um die vielen Millionen Datenpunkte dabei sinnhaft zu visualisieren wurde auf Aggregation der Daten zurückgegriffen (Shneiderman, 2008). Dafür werden die Daten über Python-Skripte²⁴ vorprozessiert, so dass die Datenverarbeitung und insbesondere der Abruf der Daten nicht live passieren muss, da einige Datenabfragen mit der aktuellen Datenbank bis zu mehreren Stunden dauern können. Die Analyse findet über Python mit vorgefertigten Notebooks statt und sollte bei Update der Daten wiederholt werden. Insgesamt soll die Ansicht Transparenz und Exploration fördern und gemäß des Mantras “*Overview first, zoom and filter, then details-on-demand*” als Einstieg helfen zu vermitteln, welche Daten überhaupt suchbar sein könnten (Whitelaw, 2015). Dabei ist die Ansicht in vier Bestandteile geteilt:



²⁴ siehe Github Dokumentation



Abb. 19: Übersichts-Zugang des entwickelten Prototyps nutzt miteinander verbundene Voronoi-Charts, durchsuchbare Listen und eine Zeitleiste um einen Datenüberblick zu ermöglichen.

Einleitungstext: Am linken Seitenrand befindet sich ein Einleitungstext inklusive Legende, Anleitung und Links zu weiterführenden Informationen, zur Projektdokumentation, zum Jupyter Notebook Curriculum und zum Code Repository (Abb. 19 oben links)

Visualisierung: Zentral befindet sich eine Visualisierung der akkumulierten Gesamtdaten für einen ausgewählten Zeitraum (Abb. 19 oben zentrum). Die Visualisierung zeigt eine Übersicht der Daten um den Knoten-Typ “Personen” zentriert. Die Knoten-Typen Personen, Ressourcen, Geographika, Körperschaften, Werke und Veranstaltungen werden je durch ein kreisförmiges Voronoi-Diagramm dargestellt. Die Gesamt-Kreisfläche basiert dabei auf der Anzahl der enthaltenen Entitäten. Jede Kreisfläche ist zusätzlich in die 100 am häufigsten vorkommenden Untergruppen (Top99 + >Top 99 als “Sonstige” zusammengefasst) gemäß des Anteils an der Gesamtmenge unterteilt. Hover über die Flächen zeigt das Label für eine Fläche und die Anzahl.

Ein Titel über den Kreisflächen zeigt neben der Kategorie die Anzahl der Entitäten an. Ein Klick auf diesen Titel öffnet eine durchsuchbare Liste aller enthaltener Entitäten, geordnet nach Häufigkeit. Hierdurch sollen Forscher*innen sich ein schnelles Bild machen können, ob zu ihrem Forschungsschwerpunkt Daten für einen bestimmten Zeitraum enthalten sein könnten, so kann man z.B. die Anzahl an Physiologen oder die Häufigkeit der Stadt “Dresden” bei Verknüpfungen leicht überprüfen und gewinnt durch die Aufschlüsselung bei den Knotentypen Ressourcen, Veranstaltungen, Werke und Körperschaften ein Gefühl dafür was diese Daten enthalten. Dies greift Feedback aus den Nutzer*innenstudien auf, dass für einige dieser Kategorien es ohne Beschreibung nicht ganz klar ist, was sich dahinter verbirgt. Zukünftig könnte an gleicher Stelle an der man auf die Liste zugreift zum Beispiel auch ein Kartensymbol ergänzt werden, welches für jeden Knotentyp eine geografische Verteilung in Form einer Heatmap anzeigt. Während die Personen, Veranstaltungen, Werke und Ressourcen in die meist verknüpften Schlagworte aufgeteilt sind, werden bei den Körperschaften²⁵ und Geographika lediglich die Entitäten selbst nach Häufigkeit dargestellt.

²⁵ Auch bei den Körperschaften war es vorgesehen, dass die Daten auf Schlagwörter basieren. Die prototypische Datenbank war zum Zeitpunkt der Datenanalyse allerdings nicht performant genug die dazu nötigen abgefragten Query-Ergebnisse innerhalb des bestehenden Timeout-Limits (ca. 2h) zurückzugeben, wodurch für die prototypische Umsetzung auf eine einfachere Alternative zurückgegriffen wurde.

Da die sozialen Beziehungen die zentralste Bedeutung bei SoNAR einnehmen, ist die Fläche für Personen im Zentrum positioniert und alle anderen Kreisflächen darum herum. Zwischen den Flächen wird ausgehend von der Anzahl der Verbindungen zwischen Personen und den anderen Knoten-Kategorien über Verbindungslien die Anzahl der Verbindungen angezeigt und über die Liniensstärke kodiert. Die Farbgebung der Linien bildet dabei die Datenquelle dieser Beziehungen ab, so dass Beziehungen aus GND z.B. rot dargestellt werden und Beziehungen aus Kalliope blau²⁶. Dadurch soll ein Eindruck vermittelt werden wo die Ursprungsdaten herkommen und in welchem Verhältnis diese stehen. So wird zum Beispiel deutlich, dass sämtliche Personen-Geographika-Verbindungen und auch Personen-Veranstaltungs-Verbindungen und Personen-Werke-Verbindungen aus der GND stammen. Für Personen-Personen-Beziehungen gibt es einen eigenen Bogen der von der Personenfläche auf sich selbst zurück zeigt. Zwischenverbindungen zwischen den anderen Knotentypen (z.B. zwischen Ressourcen und Geographika) wurden auf Grund der begrenzten Zeit im Prototyping-Prozess vorerst nicht abgebildet.

Zeitleiste: Am unteren Bildrand befindet sich eine Zeitleiste, welche die zeitliche Datenverteilung visualisiert und welche als Selektionsmittel zur Filterung dient. Die Zeitleiste ist im aktuellen Prototyp in 50er Jahre Schritte unterteilt und umfasst jeweils für jeden 50er Jahre Schritt die aggregierte Menge der enthaltenen Entitäten. In der Standardansicht werden die Anzahl an Personen-Beziehung (inklusive Personen-Beziehungen zu anderen Entitätstypen wie Ressourcen) über den zeitlichen Verlauf kategorisiert nach Datenquelle angezeigt. Ebenso wie die Farbgebung in der Visualisierung steht hier die Farben für die Datenquellen²⁷. Die zeitliche Zuordnung findet konzeptionell über die Lebensdaten statt (hat eine Person innerhalb des ausgewählten Zeitraums gelebt?). In der aktuellen Umsetzung werden allerdings nur die Geburtsdaten verwendet, weshalb in der Darstellung die Kurve von 1900 bis 2000 auch stark abflacht. In späteren Umsetzungen sieht das Konzept vor, dass man unter anderem auch sich nur die Verteilung von Personen oder z.B. Ressourcen anzeigen lassen könnte. Zudem wäre eine freiere Zeitauswahl ohne Auswahl der Zeit in 50er Jahre Schritten wünschenswert. Für die Implementierung in der Erprobungsphase war dies allerdings vorerst nicht umsetzbar, auf Grund der Vorprozessierung, weshalb dafür noch ein Konzept gefunden werden müsste.

Suchfeld: Das Suchfeld ist der eigentliche Zugang zur individuellen Netzwerk-Generierung in SoNAR. Nachdem Forscher*innen einen ersten Eindruck über die Daten haben kann hier über das Suchfeld ein Netzwerk generiert werden. Dies kann z.B. unter Eingabe einer GND-ID (z.B. "(DE-588)115568808") oder eines Schlagworts passieren (z.B. "Physiolog" für ein Netzwerk mit Entitäten die mit einem Schlagwort verknüpft sind das "Physiolog" enthält). Icons werden dabei neben dem Suchfeld angezeigt um konkrete Einstellungen vornehmen zu können. Der aktuelle Prototyp unterstützt derzeit allerdings vorerst nur eine Personen-Suche über die GND-ID (Auswahl von "P"²⁸) sowie eine Schlagwort-Suche (Auswahl "S"). Zusätzlich könnten andere Icons hier auch wesentlich komplexere Einstellungsmöglichkeiten für eine Expertensuche mit detaillierten Filter-Möglichkeiten anbieten.

²⁶ die schwarzen Flächen bestehen aus einer Mischung aller Beziehungen. Im zum Zeitpunkt der Datenanalyse bestehenden Datenmodell konnten diese Anteile auf Grund eines bekannten Transformationsfehlers noch nicht den Datenquellen GND, Kalliope, DNB, ZDB oder SBB zugeordnet werden. In einer Implementierung würden die schwarzen Verbindungen und die schwarzen Flächen in der Zeitleiste nicht existieren.

²⁷ die schwarzen Flächen bestehen aus einer Mischung aller Beziehungen. Im zum Zeitpunkt der Datenanalyse bestehenden Datenmodell konnten diese Anteile auf Grund eines bekannten Transformationsfehlers noch nicht den Datenquellen GND, Kalliope, DNB, ZDB oder SBB zugeordnet werden. In einer Implementierung würden die schwarzen Verbindungen und die schwarzen Flächen in der Zeitleiste nicht existieren.

²⁸ die Buchstaben sollten zukünftig durch klarere Icons z.B. ersetzt werden

5.2 Suchbasierte Ansichten

Die suchbasierten Ansichten sollen unterschiedliche Perspektiven auf Basis gezielter Anfragen liefern, und durch die Priorisierung von unterschiedlichen Dimensionen zusätzliche Details für mehr Interpretationsspielraum bieten. Diese Ansichten basieren dabei im Gegensatz zur Einstiegsansicht auf dem Ansatz *“Search, show context, expand on demand”* (van Ham and Perer, 2009), das heißt gezielte Suchergebnisse lassen sich hier explorativ nach Bedarf erweitern. Gemeinsamkeiten zwischen allen suchbasierten Ansichten (z.B. Abb. 22) sind:

- eine **Zeitleiste** am unteren Bildrand, welche die zeitliche Verteilung der konkreten Daten anzeigt und deren konkrete Inhalte (z.B. Anzahl Beziehungen oder Anzahl Knoten) ausgewählt werden kann. Der angezeigte Zeitraum basiert auf den jeweiligen Daten
- eine **statistische Übersicht** am linken Bildrand, welche auch Netzwerkmetriken enthält
- ein **Suchfeld** am oberen Bildrand, wodurch neue Netzwerke generiert werden oder bestehende Netzwerke gezielt erweitert werden können
- eine **Filter-Leiste** am rechten Bildrand, welche mehrere Auswahlmöglichkeiten für Filterungen enthält. Standardmäßig wird zunächst eine Liste an vorkommenden Personen geordnet nach Beziehungsanzahl gezeigt. Aber auch andere Filterungen (z.B. nach Körperschaft, Ort oder Beziehungstyp) lassen sich hier als durchsuchbare Liste anzeigen
- ein **Detailfenster**, welches sich nur öffnet bei Selektion eines Knotens oder einer Kante (Abb. 23)

Ein Button ermöglicht einen animierten Wechsel zwischen Ansichten (z.B. von der Graph-Ansicht zur Zeitbasierten Ansicht). Im aktuellen Prototyp sind zunächst nur die Graph-Ansicht und eine zeitbasierte Ansicht implementiert. Zukünftig könnte dies z.B. durch eine kartenbasierte geographische Ansicht erweitert werden. Weiterhin lassen sich Netzwerke durch Doppelklick auf Knoten um Verknüpfungen zu dem neu ausgewählten Knoten erweitern. Dies ist genauso möglich über eine erweiternde Suche unter Nutzung des Plus-Icons neben dem Suchfeld. Auf diese Weise lassen sich beliebig große Netzwerke aus mehreren Personen zusammenfügen.

5.2.1 Datenbankabfragen für die suchbasierten Ansichten

Da der Fokus auf sozialen Netzwerken liegt, werden in der Visualisierung selbst nur Personen und Körperschaften direkt im Netzwerk angezeigt. Die anderen Knoten werden aber für die Filterung nach eben diesen und die Ressourcen zur korrekten Kanten-Zuordnung benötigt, da die abgeleiteten Kanten selbst keine konkreten Attribute enthalten und auf eine jeweilige ID zu einer Ressource verweisen. Die Art/Typ der Ressource lässt sich nur anzeigen, wenn der Knoten der Ressource auch geladen wird, was bei größeren Netzwerken zu exponentiellen Steigerungen der zu ladenden Knoten führt.

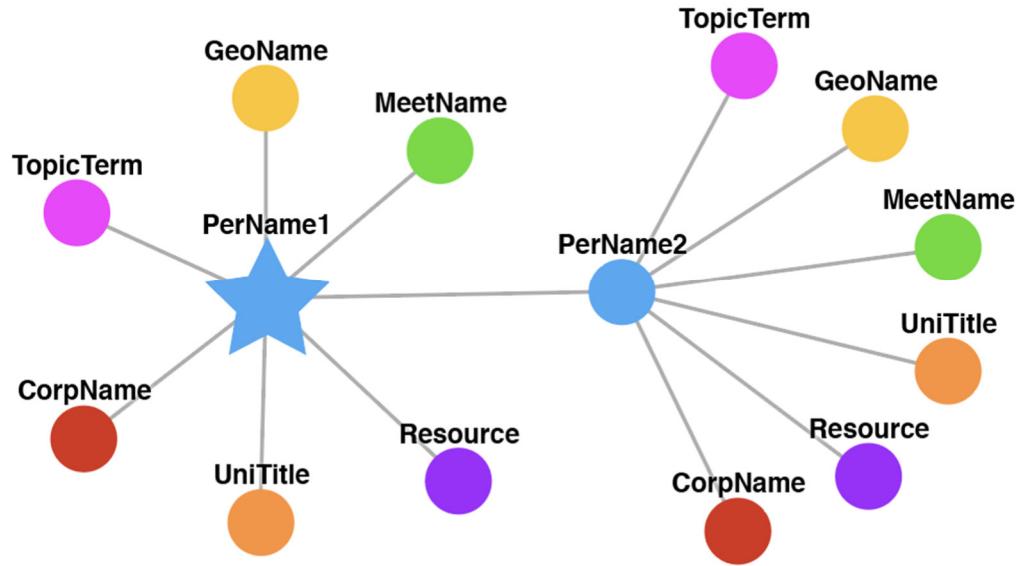


Abb. 20: Query-Struktur für Personen-Suchen (der Stern bzw. PerName1 ist die gesuchte Person).

Um alle Filter in der Visualisierung bedienen zu können und auch im Netzwerk weitere Personen nach TopicTerm oder GeoName z.B. zu filtern, müssen für alle mit PerName verbundene Personen (über SocialRelations und RelationToPerName) ebenfalls sämtliche Verknüpfungen zu TopicTerm, GeoName etc. abgefragt werden. Aktuell werden aus Performanz-Gründen im Prototyp nur direkte Verbindungen zwischen der gesuchten Person und anderen Personen wiedergegeben. Dies würde aber nur Verbindungen von/zu PerName1 anzeigen. Um über ein Ego-Netzwerk hinauszugehen und ein genaueres Bild über das Netzwerk mit den Strukturen und Communities zu erhalten müsste man auch alle Verbindungen zwischen allen Personen die im Netzwerk auftauchen, welche nicht mit PerName1 zu tun haben abfragen. Dies ist allerdings eine aktuell Performanz-Schwache Abfrage welche die Komplexität der Queries zudem erhöht, weshalb eine Optimierung der Queries für zukünftige Implementierung nötig ist.

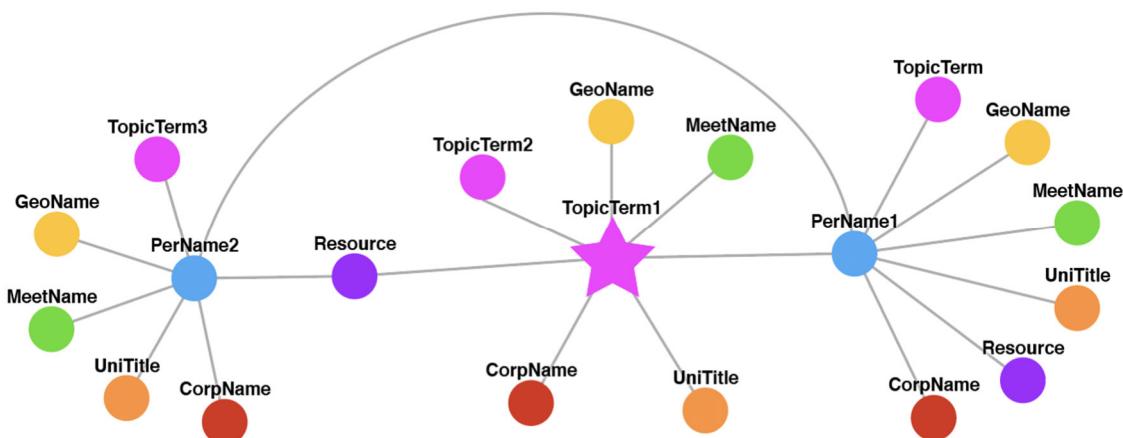


Abb. 21: Query-Struktur für Schlagwort-Suchen (Stern bzw. TopicTerm1 ist das gesuchte Schlagwort).

Neben der Personen-Suche wurde eine Schlagwort-Suche implementiert, welche einige zusätzliche Query-Ergänzungen nötig macht (Abb. 21). Schlagworte (TopicTerms) können zum einen direkt über die GND verbunden sein, häufiger aber noch sind Schlagworte mit Ressourcen (z.B. Publikationen) verbunden. So ist ein Buch zu einem bestimmten Forschungsaspekt häufig mit einem konkreten Schlagwort ausgezeichnet, die Autor*in des Buches möglicherweise aber nicht. Um auf diesen Reichtum an Schlagworten nicht zu verzichten muss also sowohl über direkt mit dem Schlagwort

verbundene Personen, als auch über indirekt über Ressourcen verbundene Personen iteriert werden. Zusätzlich müssen auch hier wie bei der Personen-Suche zu allen Personen alle weiteren Knoten-Typen abgefragt werden um die Personen zum Beispiel filterbar nach Körperschaften oder Orten zu machen. Abschließend ist es insbesondere bei der Schlagwort-Suche wichtig, dass auch sämtliche Verbindungen zwischen den im Netzwerk auftretenden Personen abgerufen werden, damit nicht nur freischwebende Knoten erzeugt werden, sondern sich Personen-Netzwerke bilden. Alles in allem ist diese Abfrage der Daten in der aktuellen Implementierung Zeit-intensiv, insbesondere bei Schlagworten mit sehr vielen Verknüpfungen. Es bleibt zu prüfen ob sich dies optimieren lässt. Weiterhin sind die beschriebenen implementierten Suchabfragen nur als beispielhafte Implementierung zu verstehen. Bei einer Weiterentwicklung des Konzepts sollten auch anderen Knotentypen (z.B. Suche nach Körperschaften) und individualisierbare Suchen (z.B. nur ein bestimmter Zeitraum, Mindestanzahl an Kanten bei angezeigten Knoten) weiter berücksichtigt werden.

5.2.2 Graph-Ansicht

Die Hauptansicht und erste Ansicht nach Betätigung einer Suche ist die Graph-Ansicht. Hier wird das Netzwerk auf traditionelle Weise mit Hilfe des D3.js basierten Force-Algorithmus zentral dargestellt. Abhängig von der konkreten Suche entsteht so entweder ein zunächst egozentrisches Netzwerk um eine Person herum (Abb. 22) oder ein bereits größeres Netzwerk basierend auf einem Schlagwort (Abb. 23). Das entstandene Netzwerk bildet lediglich Personen und Körperschaften ab um einen Fokus auf soziale Beziehungen zu ermöglichen, eine Option auch z.B. Konferenz oder Schlagworte direkt in der Visualisierung oder keine Körperschaften anzuzeigen wäre zukünftig aber denkbar und wurde in der Nutzer*innenstudie gewünscht. Die Knotengröße in dem Netzwerk basiert auf der Anzahl der verknüpften anderen Knoten. Aufbauend auf dem Konzept der Kanten-Entfaltung, basiert Kantendicke auf der Anzahl an akkumulierter Kanten.

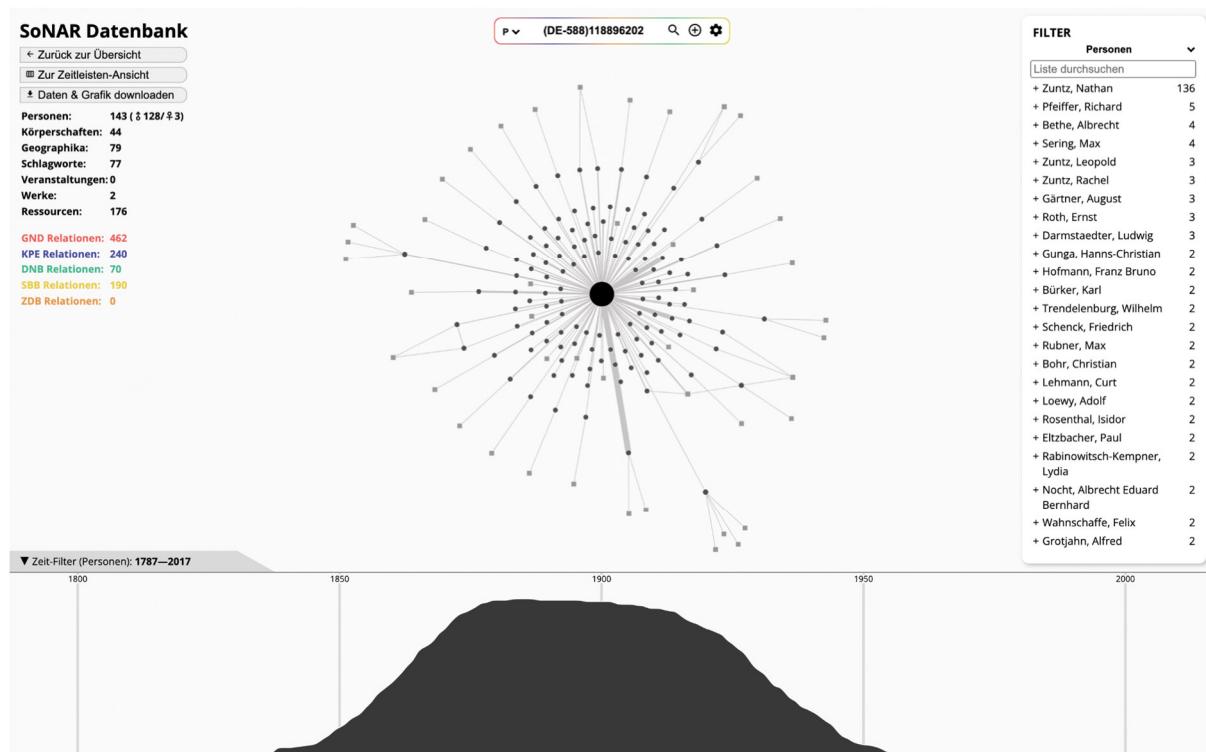


Abb. 22: Die Graph-Ansicht einer Personen-Suche zeigt eine Zeitleiste die als Filterung genutzt werden kann, ein Netzwerk im zentrum, statistische Angaben am linken Bildrand und Filter-Möglichkeiten welche gleichzeitig als Datenübersicht dienen am rechten Bildrand.

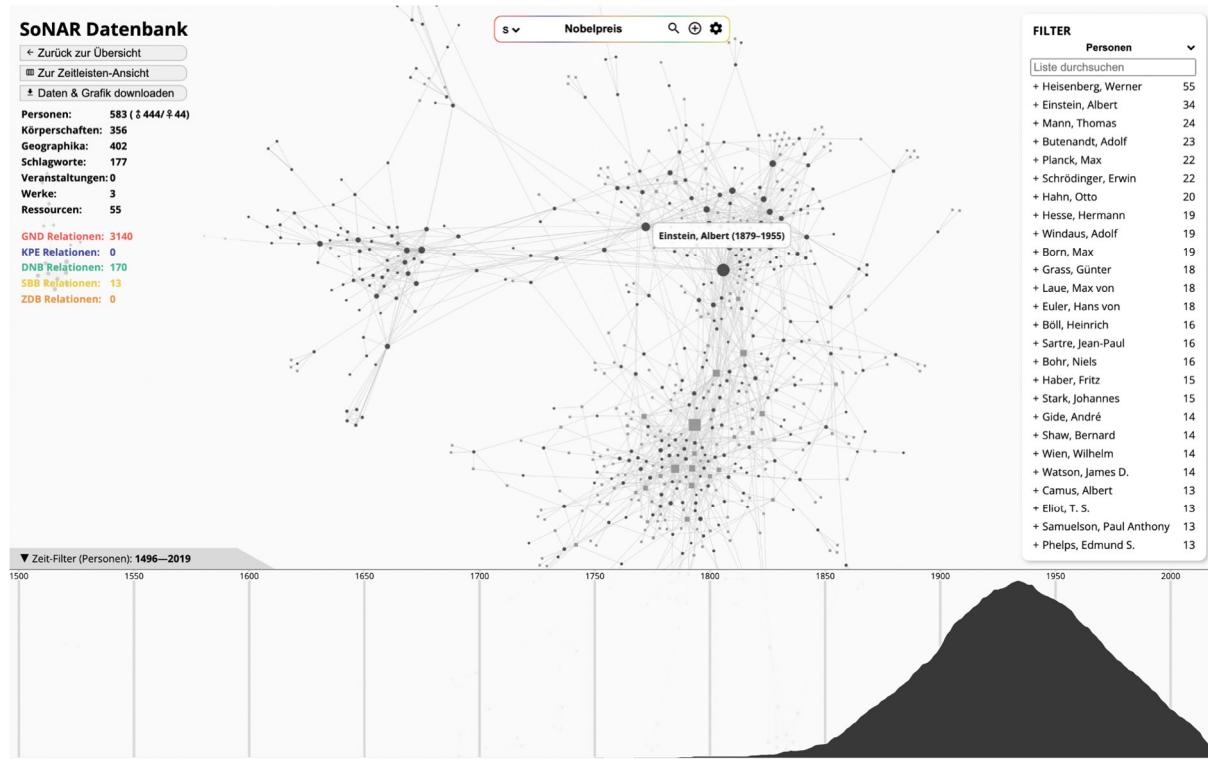


Abb. 23: Die Abbildung zeigt das Ergebnis der Schlagwort-Suche “Nobelpreis” und enthält alle Personen, welche mit einem Schlagwort-Verknüpft sind, dass “Nobelpreis” enthält. Dadurch lassen sich thematische thematische Schwerpunkte mit Hilfe der SoNAR-Daten gezielt abbilden und explorieren.

5.2.3 Zeitbasierte Ansicht

Die zeitbasierte Ansicht führt das Konzept “Morph von einem Graph-Layout zu einer Zeitleiste” aus Sektion 2.2.3 fort und ermöglicht einen Übergang von dem force-basierten Graph-Layout der Graph Ansicht zu zeitbasierten Anordnung der Knoten. Die Knoten werden dabei entsprechend der Lebensdaten zu Balken transformiert und in einer Zeitleiste angeordnet. Die Anordnung basiert auf einer Community Cluster Erkennung über netClustering.js²⁹, welches den Clauset, Newman and Moore Community Erkennungs Algorithmus verwendet. Die Nutzer*innenstudie hat ergeben, dass der Community Algorithmus ohne eigene Anpassungsmöglichkeiten allerdings skeptisch betrachtet wird. In der Zeitleiste wird in erster Linie dennoch darauf zurückgegriffen, weil dies hilft Knoten mit vielen Verbindungen zueinander nahe zueinander anzutragen und somit auch Unübersichtlichkeit in einem gewissen Rahmen vermieden wird. Farblich werden die Cluster nicht mehr unterschieden um mehr Kohärenz zwischen den Ansichten zu schaffen, bei der Farbkodierung keine Konkurrenz zur bereits bestehenden Farbkodierung für die Beziehungsquellen (GND, Kalliope etc.) zu erzeugen und weil die Clusterung über die Anordnung bereits kodiert ist. Weiterhin sollte das Feedback aufgenommen werden klar die genutzten Algorithmen zu kommunizieren und die Anzahl der gefundenen Cluster zu kommunizieren.

Eine weitere Anpassung auf Grund der Nutzer*innenstudie ist eine Verfeinerung der Jahresanzeige. Die Jahreszahlen sind nun immer Bild sichtbar und bei Zoom wird die Granularität der Anzeige feiner und zeigt so konkrete Jahre genauer an. Zudem wird die exakte Jahreszahl über Mouseover über die Timeline kodiert. Auch in dieser Ansicht sollte zukünftig zudem das Konzept der Kanten-Entfaltung implementiert werden.

²⁹ <https://github.com/john-guerra/netClusteringJs>

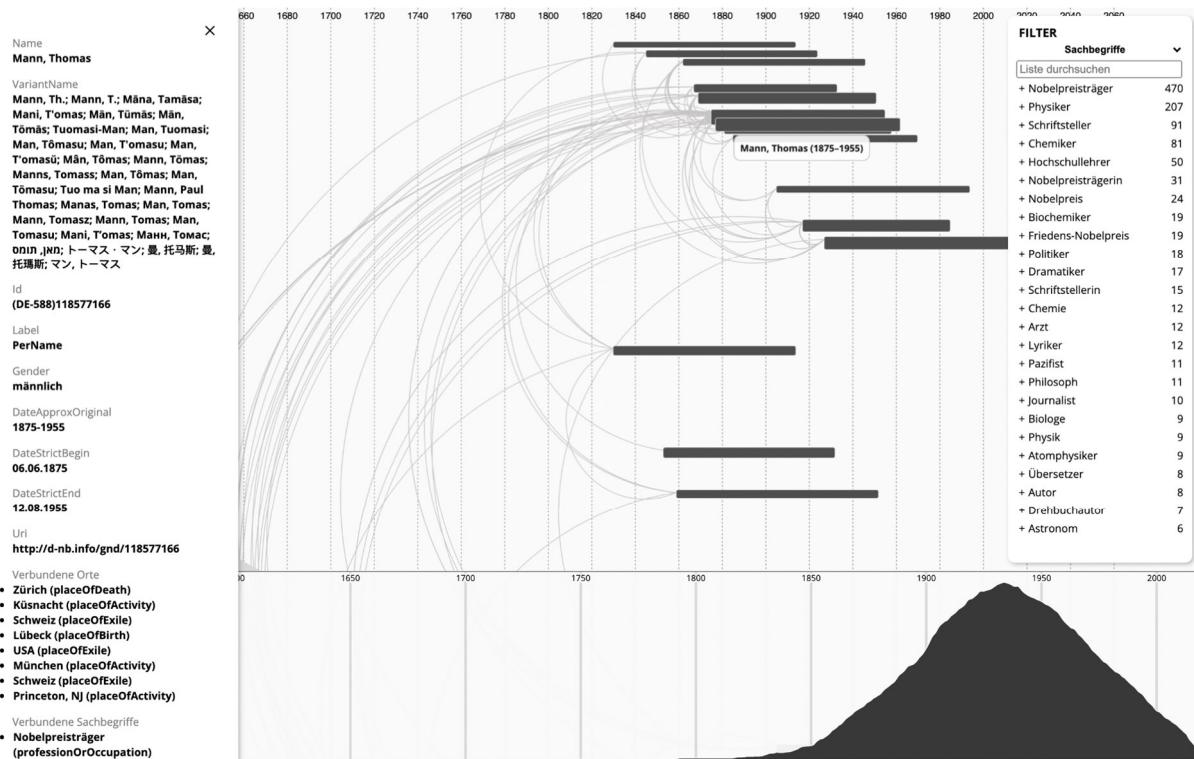


Abb. 24: Ein Button ermöglicht den Wechsel zu einer zeitbasierten Anordnung der Knoten, welche Knoten anhand eines Community-Algorithmus für eine bessere Übersicht gruppiert und zeitlich nach Geburtsdatum ordnet. Hier wird die zeitbasierte Anordnung zum Schlagwort “Nobelpreis” angezeigt mit einer Selektion von “Thomas Mann” innerhalb dieser Ansicht.

5.2.4 Filterung

Alle Suchbasierten Ansichten enthalten die gleichen Selektions- und Filtermöglichkeiten. In der prototypischen Umsetzung werden Filterung zunächst erst nach dem Laden der Daten möglich sein. Denkbar ist aber auch, dass schon bei der Datenabfrage Filterung möglich ist und der Query selbst modifiziert werden kann. Generell kann zwischen mehreren potentiellen Filtermöglichkeiten unterschiedenen werden: **1) Beziehungsbasierte Filterung, 2) Zeitbasierte Filterung, 3) Graphmetrikbasierte Filterung und 4) Attributsbasierte Filterung.**

Die Daten lassen sich über alle suchbasierten Ansichten hinweg über **Beziehungen** und verknüpfte Personen, Orte, Körperschaften und Ressourcen, aber auch über **Beziehungsattribute** (Beziehungstyp) filtern. Hierbei sind Filtermöglichkeiten abhängig von den jeweils geladenen Daten. Es kann nur gefiltert werden, was in den jeweils geladenen Daten auch vorkommt (so gibt es z.B. eine sehr große Vielzahl an möglichen Beziehungstypen wie z.B. familiäre Beziehung, Co-Author, etc.). Gleichzeitig bedeutet das, dass es für viele Attribute keine Standard-Filter gibt (keine Gruppierung von ähnlichen Filtertypen), da diese sich sehr stark unterscheiden können und einen Aufwand an manueller Zuordnung bei der Implementierung bedeuten würden. Filterung kann entweder Knoten oder Kanten betreffen. Falls das zu filternde Attribut sich auf Kanten bezieht, werden die Knoten die mit einer solchen Kante verbunden sind ebenfalls angezeigt. Bei Filterung nach Knoten werden umgekehrt auch die dazugehörigen Kanten der gefilterten Knoten angezeigt.

Weiterhin ist eine **zeitliche Filterung** der Daten möglich. Diese erfolgt durch eine Markierung eines Zeitraums (Brushing: aufziehen eines Kästchens über Klick und Ziehen) in der Zeitleiste. Ein Klick in einen nicht markierten Raum in der Zeitleiste kann dabei Reset auslösen. Ein Klick auf den Zeitleisten-Titel ermöglicht die Auswahl des zeitlichen Filter-Modus.

Generell ist die Zeitfilterung für die Forschenden bisher auf Grund der geringen Anzahl an Kanten mit Zeitattributen in den Daten nicht zufriedenstellend (ein Großteil der Kanten hat keine zeitliche Zuordnung). Ungenauigkeiten bei Zeitangaben mindern zudem die Qualität der Ergebnisse. Insgesamt ist die Filterung zur Vereinfachung der Implementierung nur auf eine Granularität eines Jahres beschränkt, das heißt es lässt sich nach Jahren filtern, nicht aber nach konkreten Monaten oder Tagen. Da nur ein Bruchteil der Beziehungen ein Attribut zur zeitlichen Einordnung enthält, aber wie im Projekt deutlich wurde die zeitliche Filterung ein zentrales Bedürfnis für die Forscher*innen ist, wurden im Projekt zwei Varianten zur Filterung erprobt. Diese Variante wäre bei besserer Datenlage zu den Kanten die ideale Variante, reduziert allerdings bei der aktuellen Datensituation die überhaupt filterbaren Kanten um ein erhebliches Maß und hat in den internen Test mit AP-2 eher zu Frustration geführt.

Um eine zusätzliche zeitliche Einordnung zu erlauben wurde zudem die Filterung über Lebensdaten von Personen im Netzwerk implementiert. Eine zeitliche Auswahl filtert hier über Personen und deren Beziehungen zueinander auf Basis der Lebenszeit einer Person; das heißt es werden nur Personen angezeigt, die in einem in gefiltertem Zeitraum gelebt haben. Hier ist der Nachteil, das zum einen nur Netzwerke möglich sind zwischen Personen aus der gleichen Epoche und somit diese Art der Filterung nicht für z.B. Zitationsnetzwerke brauchbar wäre. Gleichzeitig wird bei dieser Filterung zudem die Zeitlichkeit der individuellen Beziehungen nicht in Betracht gezogen.

Eine dritte Variante der Filterung über Zeit ist zudem die Filterung über Ressourcen (nur in Ansätzen erprobt), das heißt es werden Personen und ihre auf diesen Ressourcen-basierenden Beziehungen zueinander angezeigt, die mit Ressourcen (z.B. Briefen) verbunden sind, welche in einem ausgewählten Zeitraum erstellt wurden.

Insgesamt hat Feedback ergeben, dass alle Varianten ihre Vor- und Nachteile haben, weshalb wir bei einer Implementierung dafür plädieren mehrere Varianten anzubieten. Im Prototypen implementiert ist aktuell nur die Filterung über Lebensdaten, vorherige Prototypen bieten aber auch andere Varianten.

Weiterhin denkbar, gut implementierbar und auch in der Nutzer*innenstudie gewünscht ist eine **metrikbasierte Filterung**. Hier könnte man z.B. über Schieberegler über den Degree (Anzahl Kanten) oder andere Zentralitätsmaße eines Knotens filtern. Auch **attributsbasierte Filterung** zum Beispiel über das Geschlecht oder auch die Datenquelle sind problemlos implementierbar.

Im aktuell implementierten Prototypen sind die Zeitfilterung und beziehungsbasierte Filterung kombinierbar (Abb. 25) und lassen sich zudem durch Selektion von Knoten weiter einschränken.

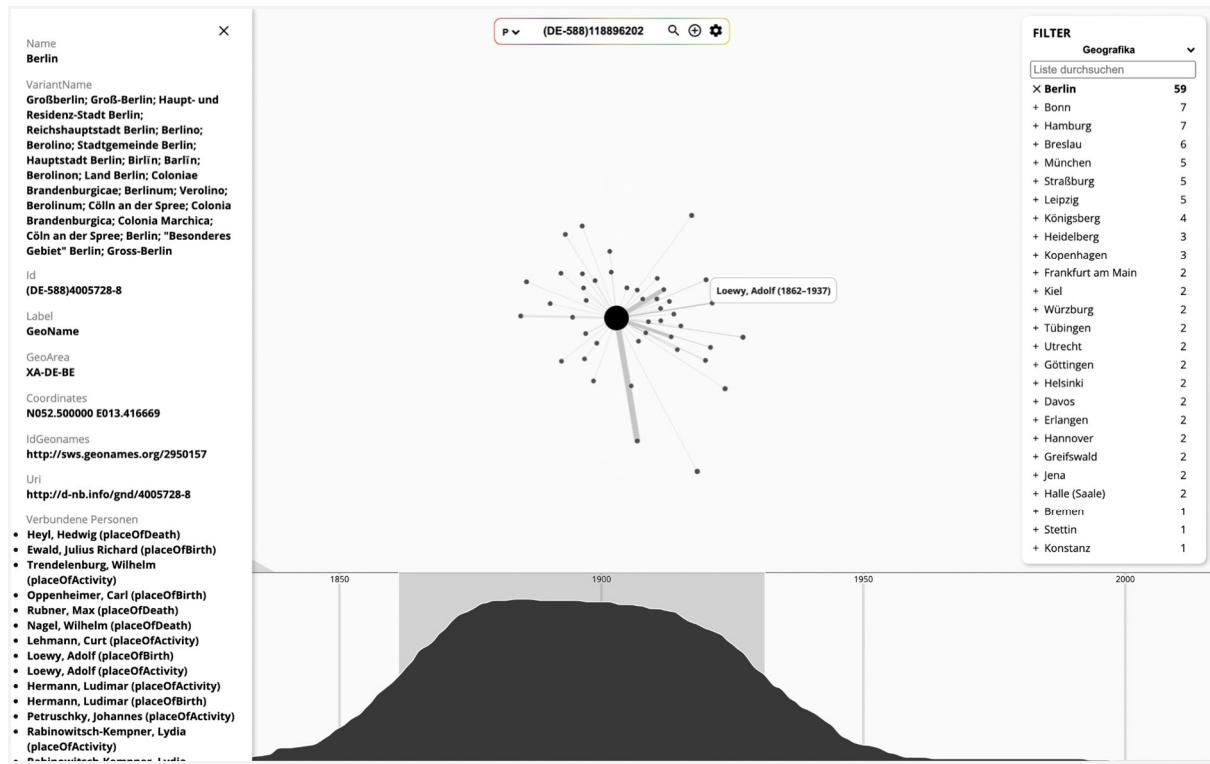


Abb. 25: Filterung nach Geografika ("Berlin") in Kombination mit einer Zeit-Filterung.

5.2.5 Selektion

Neben Filterung lassen sich Details durch Selektion von Knoten (Abb. 24) oder Kanten (siehe Abb. 26) anzeigen. Selektiert werden kann entweder über einen direkten Klick auf einen Knoten oder eine Kante oder über eine Selektion über Suche und Auswahl in den Filter-Listen (z.B. in der Personen-Übersicht). Weiterhin lassen sich in den Detailübersichten einzelner Knoten auch die verbundenen Entitäten über einen Klick auswählen. Eine Knoten-Auswahl blendet alle nicht mit dem Knoten direkt verknüpften Personen aus. Weiterhin werden in einer Detailansicht am linken Bildrand alle mit dem Knopen verknüpften Metadaten angezeigt, sowie eine strukturierte und sortierte Auflistung an verknüpften Entitäten. Bei der Kanten-Auswahl kommt dagegen das bereits beschriebene Konzept der Kanten-Entfaltung zum Tragen. Zusätzlich findet ein Zoom auf die ausgewählte Kante statt und Details zu den enthaltenen Kanten werden mit Verlinkung zur Ausgangsquelle geordnet in einer Detailansicht dargestellt.

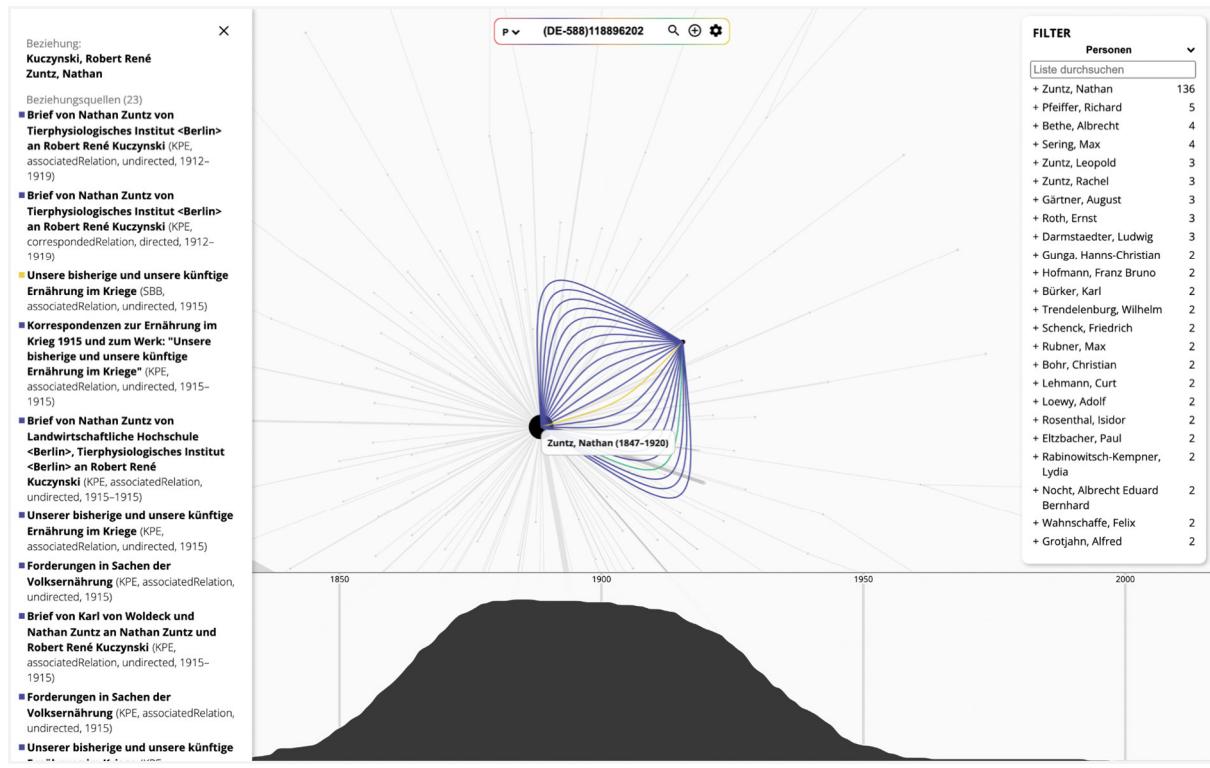


Abb. 26: Bei Klick auf eine Kante werden alle Verbindungen zwischen zwei Knoten aufgefächert und in einer Detailansicht nach Datum sortiert und nach Quelle gefärbt dargestellt.

5.2.6 Download und Export von Daten und Verlinkung zu Ansichten Daten (nicht vollständig implementiert)

Download der Daten und Export der Visualisierung sowie auch Verlinkung von Ansichten oder idealerweise Zitationsmöglichkeiten sind ein wichtiges Kriterium für die wissenschaftliche Nutzung von SoNAR. Zunächst können Suchen, Selektionen, Filterungen und Ansichten in der Visualisierung im URL-Hash abgebildet werden. Dadurch können Ansichten als Favorit gespeichert oder direkt inklusive Filter und Selektionen verlinkt werden und die Browser-History Funktionen (z.B. Zurück-Button) wäre auch funktionsfähig. Durch einen Button könnte so problemlos zusätzlich eine vereinfachte Zitierfunktion angeboten werden. Beispielhaft erprobt wurde dies durch eine direkte Verlinkung von Personen-Suchergebnissen (z.B. [https://sonar.fh-potsdam.de/prototype/\(DE-588\)115568808](https://sonar.fh-potsdam.de/prototype/(DE-588)115568808)).

Weiterhin gibt es unterschiedliche, bisher nicht implementierte Varianten, wie man Nutzer*innen Zugang zu den Daten ermöglichen kann. Neben einem einfachen Zugang über die API von AP-1 sowie eine Bereitstellung von Jupyter Notebooks (siehe Sektion 6), ist es technisch auch unproblematisch die konkret in der Visualisierung geladenen Daten zum Download zumindest als JSON Datei zur Verfügung zu stellen (da es der in der Visualisierung verwendeten Datenstruktur entspricht). Aber auch CSV-Dateien wären mit wenig Aufwand denkbar. Ein verfolgensorwerter Gedanke könnte hier sein, neben einem Download auch einen Wieder-Upload zu ermöglichen. Hierdurch könnte der Stand von Daten gesichert und eine Visualisierung selbst trotz im Laufe der Zeit angepasster Daten rekonstruiert werden ohne aufwändige serverseitige Versionierung. Neuerung im Datensatz könnten dabei anzeigbar sein. Sofern korrekt aufbereitet wäre dabei auch denkbar zusätzliche Daten in die Visualisierung unterscheidbar lokal im Browser hinzuzuladen. Die Visualisierungen werden aktuell über SVGs gerendert und wären auch als solche ohne Qualitätsverlust vor weißem Hintergrund speicherbar oder als PDF über eine Druckfunktion exportierbar. Aber auch ein Rendering auf Canvas ist mit etwas Zusatzaufwand implementierbar, was auch eine Speicherung als JPG ermöglichen würde.

6. Jupyter Notebook HNA Curriculum (AP-3.4)

Die Nutzer*innenstudie und Interviews haben unter anderem gezeigt, dass insbesondere für Anwender*innen mit Fokus auf quantitative Analysen der einfache Zugang zu den Daten selbst wesentlich wichtiger ist als deren Visualisierung, und dass die Dokumentation der Prozesse um das Daten-Retrieval und die Visualisierung eine hohe Priorität haben. Um weitere Analyse-Möglichkeiten bereitzustellen, setzen wir zusätzlich zu Explorations-fokussierten Visualisierung auf einen Zugang zu den Daten über dokumentierende und als Einführung dienende Jupyter-Notebooks. Diese sind durch die Verbindung von auf Python basierenden Code und beschreibendem Text eine in sich geschlossene Dokumentation und sind so in der Lage in der Visualisierung dargestellte Prozesse festzuhalten, sie reproduzierbar zu machen, aber auch darüber hinausgehende Analysen zu ermöglichen die in der webbasierten Visualisierung nicht angeboten werden. Parallel zur Implementierung des webbasierten Prototyps wurde deshalb über die Werkvertragsmittel der FHP die Agentur Limebit³⁰ beauftragt in enger Zusammenarbeit mit AP-2 (HHU) und unter Leitung der FHP (AP-3) ein Jupyter Notebook basiertes HNA Curriculum zu erstellen um eine Nutzung der SoNAR Daten für unterschiedlichste Zielsetzungen zu demonstrieren und erklären. Das Visualisierungs-Interface soll zwar eine Vielzahl an Visualisierungs-Formen der Daten bieten und zum Teil auch Netzwerk-Metriken anzeigen und Zugriff und Download der Daten ermöglichen, aber es ist unmöglich alles anzubieten. Gerade mit Netzwerk-Algorithmen erfahrene Anwender*innen haben den Interviews und der Nutzer*innenstudie häufig mehr Interesse an einem besonders einfachen Zugang zu den Daten für eigene Berechnungen mit der Möglichkeit auch zu sehen, welcher Algorithmus/Code exakt zu dem gezeigten Ergebnissen in der Visualisierung führen. Um diese Nutzer*innengruppe ebenfalls zu berücksichtigen, Transparenz bieten bezüglich der Analyse-Prozesse in der Visualisierung, zusätzliche Analyse-Möglichkeiten zu bieten und die API einzuführen gibt es zusätzlich zu der explorations-fokussierten Visualisierung einen weiteren Interface Zugang über Jupyter Notebooks. Mit Python/Jupyter Notebooks kann man einfach auf die Neo4J API zugreifen und performant gezielte Queries abfragen und individuell anpassen.

Jupyter-Notebooks sind hybride Dokumente, die sowohl Code als auch Markup enthalten. Sie vermischen Code und Dokumentation. Beschreibungen lassen sich durch die Auszeichnungssprache Markdown über Code-Zellen hinzufügen. Die Code-Zellen sind z.B. in der Programmiersprache Python geschrieben. Code ist nacheinander/separat ausführbar. Durch die Vermischung von Beschreibung und Code sind Jupyter Notebooks attraktiv für Dokumentation von Code, aber auch für Schritt für Schritt Tutorials oder nachvollziehbare datenanalytische Aktivitäten. Hier können Notebooks als zusätzliches Outputs dienen, die die Visualisierung ergänzen oder eine vereinfachte Schnittstelle zu den Daten bieten. Man könnte z.B. mehrere Beispiele für netzwerkanalytische Verfahren zeigen und sich den exakten Such-Query einer Auswahl in der Visualisierung zur Extrahierung anbieten, so dass man es unproblematisch in ein Notebook einfügen kann.

Das entwickelte Curriculum (Abb. 27) ist eine Einladung zu lernen, wie man mit interaktiven Jupyter-Notebooks historische Netzwerkanalyse auf einem großen Graphendatensatz durchführt. Zusätzlich geht es in erster Linie darum die Nutzung einer verständliche Einführung in die Nutzung der SoNAR (IDH) Datenbank zu geben und alternative, individuellere Analysemethoden der Daten anzubieten. Insgesamt besteht es aus fünf Tutorials in Form von interaktiven Notebooks, die in die Programmiersprache Python, die Graphdatenbank Neo4j und deren Abfragesprache Cypher einführen. Inhaltlich wurden die Notebooks aufbauend auf den Erfahrungen im Projekt, den geplanten Funktionalitäten im webbasierten Prototypen, mit der Hilfe von zwei gezielten Workshops und engem Austausch mit den HNA-Forscher*innen aus unserem Team erarbeitet. Ergebnis ist ein Curriculum mit den folgenden fünf Kapiteln:

1. Jupyter and Python: Grundlegende Einführung in Python und Jupyter

³⁰ <https://limebit.de/>

2. **Historical Network Analysis:** Grundlegende Einführung in die Graphentheorie und HNA. Beispiel einer Netzwerkanalyse anhand eines Netzwerks von Nobelpreisträgern.
3. **SoNAR (IDH):** Wie man auf SoNAR (IDH) Daten zugreift und sie abfragt. Einführung in Neo4j und Cypher. Beispiele für Datenexploration, deskriptive Analyse und komplexe Abfragen.
4. **History of Physiology:** Fallbeispiel für die Analyse des historischen Netzwerks von Physiologen. Darstellung des explorativen Ansatzes zur HNA.
5. **History of Economy:** Beispiel für die Analyse des historischen Netzwerks des Ökonomen Gustav von Schmoller. Darstellung eines hypothesenbasierten Ansatzes.

Link: <https://github.com/sonar-idh/jupyter-curriculum>

Notebook	Content	Interactive Version
1 - Jupyter and Python	Basic introduction to Python and Jupyter. Can be skipped in case you worked with Python before.	Launch binder
2 - Historical Network Analysis	Basic introduction to graph theory and HNA. Example of network analysis based on a network of Nobel Laureates.	Launch binder
3 - SoNAR (IDH)	How to access and query SoNAR (IDH) data. Introduction to Neo4j and Cypher. Examples of data exploration, descriptive analysis and complex queries.	Launch binder
4 - History of Physiology	Example case of analysing historical network of physiologists. Depiction of exploratory approach to HNA.	Launch binder
5 - History of Economy	Example case of analysing historical network of the economist Gustav von Schmoller. Depiction of a hypothesis based approach.	Launch binder

Abb. 27: Einleitung und Überblick über das Curriculum.

7. Technische Implementierungsdetails

Aufbauend auf den zwei Strängen **1) webbasiertes Explorationsinterface** und **2) Jupyter Notebook basiertes HNA Curriculum** unterscheidet sich die Dokumentation der Implementierung. Generell werden alle Ergebnisse aus AP-3 im SoNAR Github Repository gesammelt und beide Stränge nutzen für die prototypische Umsetzung die Neo4J Datenbank Schnittstelle:

1) bei den webbasierten Visualisierungen werden sowohl die abschließenden übergreifenden Prototypen, als auch zwischen Ergebnisse und ggf. verallgemeinerte Funktionen kommentiert und strukturiert bei GitHub zur Verfügung gestellt:

- die Visualisierungen basieren auf HTML, CSS und JavaScript unter der Nutzung der Visualisierungs-Library D3.js
- während die Übersichtsansicht vor-prozessierte Daten verwendet (Data Analysis kommt auch zu GitHub) werden die Daten für den Explorations-Teil live abgerufen
- aktuell verwenden die prototypischen Visualisierungen keine Frameworks für modulareres DOM-Management. Für ein Folgeprojekt empfehlen wir die Nutzung von Frameworks wie Vue.js oder Svelte um den Instandhaltungsaufwand zu verringern
- aktuell basieren die Visualisierung größtenteils auf SVGs, zukünftig wäre ein Umschwenken zu Canvas oder WebGL aus Performanz-Gründen ratsam
- aktuell werden für die Queries mehrere Queries nacheinander abgeschickt und anschließend im Browser für die Visualisierung kombiniert. Dies führt zu Dupletten und mehrfach abgefragten Knoten, welche die Prozessierung verlangsamen und die Datenmenge erhöhen. Bei anderen konzeptionell notwendigen Queries (z.B. die Notwendigkeit Ressourcen zu laden um die abgeleiteten Kanten Ressourcen zu ordnen zu können) bei größeren Netzwerken kommt die aktuellen Implementierung bereits an ihre Grenzen der Durchführbarkeit. Bei Schlagwort-Suche können so im aktuellen Prototyp die entfalteten Kanten nicht alle korrekt zugeordnet werden und werden so

fälschlicherweise häufig der GND zugeordnet. Mit mehr Know-How bezüglich der Fertigung spezifischer komplexer Queries und mehr Fokus auf die Datenbank und den Server, aber auch eine weitere Optimierung des Datenmodells ließe sich hier zumindest aber vermutlich deutlich die Performanz verbessern.

- Die Einzelnen Prototypen und deren Dokumentation, sowie bekannten Probleme im Code ("Bugs") werden hier gesammelt:
<https://github.com/sonar-idh/visualization-prototypes>

2) die Jupyter Notebooks sind von ihrer Beschaffenheit her und die Verbindung von Code und beschreibenden Text von sich aus eine in sich geschlossene Dokumentation:

- die Notebooks verwenden die Programmiersprache Python unter der Nutzung von Jupyter Notebooks
- Notebooks + Dokumentation/Readme befinden sich hier: <https://github.com/sonar-idh/jupyter-curriculum>

8. Diskussion und Herausforderungen

Im Projektverlauf wurde die zwingende Notwendigkeit deutlich Projekte für Technologien wie SoNAR in einem interdisziplinären Team unter Einbezug von Forschenden und potentiellen Nutzer*innen zu entwickeln, weil ohne diese eine zielgerichtete Implementierung abgestimmt auf die Forschungsanforderungen kaum denkbar ist. Gleichzeitig wurde die in den digitalen Geisteswissenschaften häufig beschriebenen Herausforderungen im Bezug auf geisteswissenschaftliche Daten deutlich: Zum einen ist eine gewisse Normalisierung heterogener Daten für Visualisierungen und Datenanalysen zwingend notwendig, zum anderen ist für eine weiterführende Nutzung und Interpretation der Daten durch Forschende gerade dies ein Hindernis.

Das Konzept und die dargestellten Ansichten enthalten bisher noch so gut wie keine Legenden oder Hifestellungen zur besseren Verständlichkeit der Visualisierung. Dies ist, wie die Nutzer*innenstudien gezeigt haben, in jedem Fall etwas wo in einem zukünftigen Projekt noch deutlich ein Fokus gelegt werden muss um z.B. Knotengrößen, Knotenformen, Farbgebungen, aber auch die Funktionen transparent und leicht verständlich zu kommunizieren.

Herausforderungen im Bezug auf eine zukünftige Implementierung liegen dabei insbesondere in Fragen zur Skalierbarkeit des Gesamtprojektes (z.B. weitere Datensätze, einbezug von externen Metadaten) mit Hinblick auf die Performanz und Komplexität der Visualisierungen und Datenabfragen. Zum einen wünschen sich zum Beispiel die Forschenden möglichst viele Daten, andererseits führen zusätzliche Datenmengen zu wachsenden Problemen, was die Performanz der Datenbank betrifft und die Darstellbarkeit in einer Visualisierung. Falls die Datensätze grundlegend erweitert werden sollten, wäre es z.B. möglicherweise sinnvoll, dass Nutzer*innen bereits vor der Suche auswählen müssen, in welchen Datensätzen überhaupt gesucht werden soll.

Was die Skalierbarkeit betrifft so wünschen sich die Forschenden möglichst viele Daten, andererseits führen zusätzliche Datenmengen zu wachsenden Problemen was die Performanz der Datenbank betrifft und die Darstellbarkeit in einer Visualisierung. Falls die Datensätze grundlegend erweitert werden sollten, wäre es möglicherweise sinnvoll, dass Nutzer*innen bereits vor der Suche auswählen müssen, in welchen Datensätzen überhaupt gesucht werden soll.

Generell gibt es bei den Forscher*innen einen großen Bedarf an zum einen vielen unterschiedlichen Ansichten und zum anderen auch an einer möglichst hohen Individualisierbarkeit über die Nutzung individueller Parameter. Es wäre aus unserer Sicht zwar in jedem Fall sinnvoll zusätzlich zu den implementierten Ansichten z.B. eine kartenbasierte, geographische Ansicht zu ergänzen, weil dies ein

viel gewünschtes Feature ist. Auch Ansichten in denen je nach Bedarf auch andere Knotentypen eingeblendet werden oder einfach weitere Filter ergänzt werden wären sinnvoll. Aktuell werden bei Filterung Knoten auch lediglich ausgeblendet um die Position für eine bessere Übersichtlichkeit zu erhalten. Eine optionale Neuanwendung der Force Algorithmen wäre ein vielversprechender zusätzlicher Ansatz.

Die gewünschte Individualisierbarkeit von eigenen Parametern, Auswahl aus mehreren Algorithmen etc. würden wir aktuell jedoch nicht empfehlen. Einerseits würde dies einen erheblichen Implementierungsaufwand über Jahre hinweg bedeuten, wie in vergleichbaren Projekten wie Gephi sieht. Zudem würden weitere Optionen und Funktionalitäten die Nutzung eines solchen Tools erheblich verkomplizieren. Wir sehen eine Visualisierungstool von SoNAR eher als eine Erstzugang zu den Daten oder als Explorationsmittel, wohingegen eigene Analysen aus unserer Sicht auf Grund der vielen unterschiedlichen Forschungsfragen besser gezielt von den Anwenderinnen selbst außerhalb einer bereitgestellten Technologie erfolgen sollten. Hierfür sehen wir insbesondere Potential von bereits aufbereiteten Jupyter Notebooks, welche Beispielhaft in die Nutzung der API einleiten. Gleichzeitig sind wir der Überzeugung das die Kombination mit den Jupyter Notebooks und einer hohen Transparenz bezüglich der API diese Nutzer*innengruppe welche besser zu erreichen, welche besonderen Bedarf an der Individualisierbarkeit hat.

9. Fazit

Im Projektverlauf wurde die zwingende Notwendigkeit deutlich Projekte für Technologien wie SoNAR in einem interdisziplinären Team unter Einbezug von Forschenden und potentiellen Nutzer*innen zu entwickeln, weil ohne diese eine zielgerichtete Implementierung abgestimmt auf die Forschungsanforderungen kaum denkbar ist. Gleichzeitig wurde die in den digitalen Geisteswissenschaften häufig bereits beschriebenen Herausforderungen im Bezug auf geisteswissenschaftliche Daten deutlich: Zum einen ist eine gewisse Normalisierung heterogener Daten für Visualisierungen und Datenanalysen zwingend notwendig, zum anderen ist für eine weiterführende Nutzung und Interpretation der Daten durch Forschende gerade dies ein Hindernis. Darauf aufbauend ist eine weitere grundlegende Erkenntnis aus der Projektbearbeitung die Wichtigkeit von Transparenz in Bezug auf alle Schritte einer Pipeline einer zukünftigen SoNAR-Technologie: Transparenz ist eine grundlegende Voraussetzung, um Netzwerkdaten und Visualisierungen für die HNA nutzbar machen und in den wissenschaftlichen Diskurs einbringen zu können, weshalb auch bei einer zukünftigen Implementierung unbedingt ein Fokus auf Möglichkeiten zur transparenten Visualisierung und Kommunikation von vorausgegangenen Prozessen gelegt werden sollte.

Referenzen

- Ahn, J., Plaisant, C. and Schneiderman, B.** (2013). A Task Taxonomy for Network Evolution Analysis. *IEEE Transactions on Visualization and Computer Graphics*, **20**(3), pp. 365–76.
- Bach, B. et al.** (2015). NetworkCube: Bringing Dynamic Network Visualizations to Domain Scientists. p. 3.
- Behrisch, M. et al.** (2016). Matrix Reordering Methods for Table and Network Visualization. In *Computer Graphics Forum*. Wiley Online Library, pp. 693–716.
- Bludau, M.-J., Dörk, M. and Tominski, C.** (2021). Unfolding Edges for Exploring Multivariate Edge Attributes in Graphs. In Byška, J.Jänicke, S.and Schmidt, J. (eds), *EuroVis 2021 - Posters*. The Eurographics Association. 10.2312/evp.20211070.
- Bostock, M., Ogievetsky, V. and Heer, J.** (2011). D3: Data-Driven Documents. *TVCG: Transactions on Visualization and Computer Graphics*, **17**(6), pp. 2301–9.
- Brüggemann, V., Bludau, M.-J. and Dörk, M.** (2020). The Fold: Rethinking Interactivity in Data Visualization. *DHQ: Digital Humanities Quarterly*, **14**(3).
- Chang, D. et al.** (2009). Visualizing the Republic of Letters. *Stanford: Stanford University*. Retrieved April, 21, p. 2014.

- Chen, K., Dörk, M. and Dade-Robertson, M.** (2014). Exploring the Promises and Potentials of Visual Archive Interfaces. In *Proceedings of the 2014 ICConference*. ISchools, pp. 735–41.
- Dörk, M., Pietsch, C. and Credico, G.** (2017). One View Is Not Enough. *Information Design Journal*, **23**(1), pp. 39–47.
- Drucker, J.** (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, **5**(1), pp. 1–21.
- Düring, M. and von Keyserlingk, L.** (2015). Netzwerkanalyse in Den Geschichtswissenschaften. Historische Netzwerkanalyse Als Methode Für Die Erforschung von Historischen Prozessen. In *Prozesse*. Springer, pp. 337–50.
- Fekete, J.-D. and Plaisant, C.** (2002). Interactive Information Visualization of a Million Items. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium On*. IEEE, pp. 117–24.
- Gibson, H., Faith, J. and Vickers, P.** (2012). A Survey of Two-Dimensional Graph Layout Techniques for Information Visualisation. *Information Visualization*, **12**(3–4), pp. 324–57.
- Hadlak, S., Schumann, H. and Schulz, H.-J.** (2015). A Survey of Multi-Faceted Graph Visualization. *Eurographics Conference on Visualization (EuroVis) - STARs*, p. 20 pages. 10.2312/EUROVISSTAR.20151109.
- van Ham, F. and Perer, A.** (2009). "Search, Show Context, Expand on Demand": Supporting Large Graph Exploration with Degree-of-Interest. *IEEE Transactions on Visualization and Computer Graphics*, **15**(6), pp. 953–60. 10.1109/TVCG.2009.108.
- Henry, N. and Fekete, J.** (2006). MatrixExplorer: A Dual-Representation System to Explore Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, **12**(5), pp. 677–84. 10.1109/TVCG.2006.160.
- Hinrichs, U., Forlini, S. and Moynihan, B.** (2019). In Defense of Sandcastles: Research Thinking through Visualization in Digital Humanities. *Digital Scholarship in the Humanities*, **34**(Supplement_1), pp. i80–99. 10.1093/llc/fqy051.
- Isenberg, P. et al.** (2008). Grounded Evaluation of Information Visualizations. In *Proceedings of the 2008 Conference on BEyond Time and Errors Novel EvaLuation Methods for Information Visualization - BELIV '08*. the 2008 conference. Florence, Italy: ACM Press, p. 1. 10.1145/1377966.1377974.
- Jacomy, M. et al.** (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PloS One*, **9**(6), p. e98679.
- Kerracher, N., Kennedy, J. and Chalmers, K.** (2015). A Task Taxonomy for Temporal Graph Visualisation. *IEEE Transactions on Visualization and Computer Graphics*, **21**(10), pp. 1160–72. 10.1109/TVCG.2015.2424889.
- Kerren, A., Purchase, H. C. and Ward, M. O.** (2014). Introduction to Multivariate Network Visualization. In Kerren, A. Purchase, H.C. and Ward, M.O. (eds), *Multivariate Network Visualization*. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 1–9. 10.1007/978-3-319-06793-3_1.
- Lamqaddam, H. et al.** (2020). Introducing Layers of Meaning (LoM): A Framework to Reduce Semantic Distance of Visualization In Humanistic Research. p. 11.
- von Landesberger, T. et al.** (2011). Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges. *Computer Graphics Forum*, **30**(6), pp. 1719–49.
- Lee, B. et al.** (2006). Task Taxonomy for Graph Visualization. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors Novel Evaluation Methods for Information Visualization - BELIV '06*. the 2006 AVI workshop. Venice, Italy: ACM Press, p. 1. 10.1145/1168149.1168168.
- McInnes, L., Healy, J. and Melville, J.** (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Journal of Open Source Software*, **3**.<http://arxiv.org/abs/1802.03426> (accessed 3 November 2020).
- Nobre, C. et al.** (2019). The State of the Art in Visualizing Multivariate Networks. *Computer Graphics Forum*, **38**(3), pp. 807–32. 10.1111/cgf.13728.
- Novak, J. et al.** (2014). HistoGraph -- A Visualization Tool for Collaborative Analysis of Networks from Historical Social Multimedia Collections. In *2014 18th International Conference on Information Visualisation*. 2014 18th International Conference on Information Visualisation (IV). Paris, France: IEEE, pp. 241–50. 10.1109/IV.2014.47.
- Okoe, M., Jianu, R. and Kobourov, S.** (2019). Node-Link or Adjacency Matrices: Old Question, New Insights. *IEEE Transactions on Visualization and Computer Graphics*, **25**(10), pp. 2940–52. 10.1109/TVCG.2018.2865940.
- Pienta, R. et al.** (2015). Scalable Graph Exploration and Visualization: Sensemaking Challenges and Opportunities. In *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*. 2015 International Conference on Big Data and Smart Computing (BigComp). Jeju, South Korea: IEEE, pp. 271–8. 10.1109/35021BIGCOMP.2015.7072812.

- Pienta, R. et al.** (2017). Visual Graph Query Construction and Refinement. In *Proceedings of the 2017 ACM International Conference on Management of Data*. pp. 1587–90.
- Rollinger, C. et al.** (2017). Editors' Introduction. *Journal of Historical Network Research*, p. i-vii Pages. 10.25517/JHNR.V1I1.19.
- Rossi, R. A. and Ahmed, N. K.** (2015). The Network Data Repository with Interactive Graph Analytics and Visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.<http://networkrepository.com>.
- Shneiderman, B.** (2008). Extreme Visualization: Squeezing a Billion Records into a Million Pixels. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data - SIGMOD '08*. the 2008 ACM SIGMOD international conference. Vancouver, Canada: ACM Press, pp. 3–12. 10.1145/1376616.1376618.
- Shneiderman, B.** (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium On*. IEEE, pp. 336–43.
- Van Ham, F. and Perer, A.** (2009). 'Search, Show Context, Expand on Demand': Supporting Large Graph Exploration with Degree-of-Interest. *IEEE Transactions on Visualization and Computer Graphics*, **15**(6).
- Warren, C. N. et al.** (2016). Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *DHQ: Digital Humanities Quarterly*, **10**(3).
- Whitelaw, M.** (2015). Generous Interfaces for Digital Cultural Collections. *DHQ: Digital Humanities Quarterly*, **9**(1).
- Zimmerman, J., Forlizzi, J. and Evenson, S.** (2007). Research through Design as a Method for Interaction Design Research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*. the SIGCHI Conference. San Jose, California, USA: ACM Press, pp. 493–502. 10.1145/1240624.1240704.