








Graph Technologies for the Analysis of Historical Social Networks Using Heterogeneous Data Sources

Sina Menzel^{*†}  Mark-Jan Bludau^{*‡}  Elena Leitner^{*§} 
Marian Dörk[‡]  Julián Moreno-Schneider[§]  Vivien Petras[†] 
Georg Rehm[§] 

^{*}The authors contributed equally to this work as first authors.

[†] Humboldt-Universität zu Berlin

[‡] FH Potsdam – University of Applied Sciences

[§] DFKI – Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Abstract

Over the last decades, cultural heritage institutions have provided extensive machine-readable data, such as bibliographic and archival metadata, full text collections, and authority records containing multitudes of implicit and explicit statements about social relations between various types of entities. In this paper, we introduce approaches to examine and evaluate viable ways to build and operate an advanced research infrastructure based on heterogeneous data sources from cultural heritage institutions to support Historical Network Analysis. We describe challenges and strategies from our interdisciplinary research, focusing on the data processing, the human-centered approach in form of a preliminary co-design workshop, as well as an iterative approach to data visualization creation.

Copyright © by the paper's authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In: TODO

1 Introduction

The study of historical events is relevant to many disciplines in the Digital Humanities. The analysis of relationships between agents often is the key to understanding and explaining social phenomena. At the same time, historical research topics often depend on data sources from their time period. Therefore, the joint combination of available historical sources is crucial for the reconstruction of historical networks. This is where the method of Historical Network Analysis (HNA) comes into play. As derivation of Social Network Analysis, HNA includes the dependency on numerous historical sources that ideally support each other (Jansen and Wald, 2007).

One limiting factor in HNA can be a lack of awareness with regard to the availability of suitable research data. At the same time, cultural heritage institutions have produced very large amounts of machine-readable, typically standardized and also well-organized data over the last decades: bibliographic and archival metadata, full text collections, and sets of authority or reference records. These data sets contain a plethora of implicit and explicit statements about social relations, which can be exploited for HNA research. However, systematically combining multiple data sources, and extracting as well as visualizing the complex resulting networks currently requires extensive knowledge in graph theory as well as time-consuming manual work carried out by the individual researcher. This is not least due to the heterogeneity of the data sources provided by cultural heritage institutions, e. g., in terms of data formats.

The research project *SoNAR (IDH)*¹, *Interfaces to Data for Historical Social Network Analysis and Research*, addresses this issue. We examine and evaluate approaches to develop and operate a research infrastructure supporting HNA based on heterogeneous cultural heritage data. In this paper, we present first insights from the project and describe challenges in the process of modeling and transforming heterogeneous data sources as well as designing user-centered visualization for historical social networks. By sharing our approach and challenges alongside the process, we aim to contribute to the ongoing research on the suitability of bibliographic big data for HNA and the process of developing corresponding research technologies.

2 Related work

The following section gives an overview of prior research as well as projects related to graph modeling and visualization approaches within the Digital Humanities from the perspective of historical network analysis.

¹<https://sonar.fh-potsdam.de>

2.1 Related projects

Open knowledge graphs have been discussed frequently over the past years as an alternative to a document-based approach (Auer and Mann, 2019). Several large initiatives such as *EOS*², *Europeana*³ and *CLARIN*⁴, provide access to cultural data for researchers in the Digital Humanities. At the same time, the issue of decentralized and heterogeneous bibliographic data sources is being addressed by projects such as *Culturegraph* (Vorndran, 2018), and *DARIAH-DE*⁵ in the Digital Humanities, *Lynx*⁶ in the legal domain, and, to a certain extent, *ELG*⁷ in language technology (Rehm et al., 2020). Most of these initiatives connect to infrastructures of cultural heritage institutions, often hosted by libraries or archives.

Even though these initiatives often enable, i. a., the derivation of new, previously unidentifiable or implicit information, they do not primarily focus on the extraction of network data. Therefore, HNA researchers are often left to create their individual graphs after gathering data that is suitable to address their research question. This is mostly done in software tools such as *Gephi*⁸, *Palladio*⁹ or *VennMaker*¹⁰, to name only a few open source solutions.

Along with the establishment of network analysis as a method in historical research, we observe an increase of joint research projects on the extraction of historical networks for researchers within the Social Sciences and Humanities.

For example, the project *Six degrees of Francis Bacon*¹¹ uses statistical methods on the base data to infer relations to reconstruct and visualize an early modern Britain historical social networks. The project allows for the expansion and curation of the data through collaborative annotation by the users (Warren et al., 2016). The *histoGraph*¹² initiative also follows a collaborative approach by offering users the possibility for exploration and collaborative research of historical social networks in multimedia collections, with special focus on the computational construction and crowd-sourcing of re-

²European Open Science Cloud, <https://www.eosc-portal.eu>

³<https://www.europeana.eu>

⁴Common Language Resources and Technology Infrastructure, <https://www.clarin.eu>

⁵<https://de.dariah.eu>

⁶<http://www.lynx-project.eu>

⁷<https://www.european-language-grid.eu>

⁸<https://gephi.org>

⁹<https://hdlab.stanford.edu/palladio/>

¹⁰<https://www.vennmaker.com>

¹¹<http://www.sixdegreesoffrancisbacon.com>

¹²<http://histograph.eu>

lations from photo collections (Novak et al., 2014). In a project between several European research institutions, *Issues with Europe – A Network analysis of the German-speaking Alpine Conservation Movement (1975-2005)*¹³ currently examines the disputes over European alpine transit policy. Moreover, the Austrian project *APIS – Mapping historical networks* has been working on the extraction and visualization of networks from more than 18,000 records in the Austrian Biographical Encyclopaedia¹⁴. The German project *Gesellschaftliche Wissensproduktion in der Aufklärung – Text- und netzwerkanalytische Diskursrekonstruktion* considers full texts of more than 300 periodicals published in Halle, Germany, between 1688 and 1815, and combines the methods of topic modeling with historical network analysis in order to systematically analyze public discourse during the age of enlightenment (Purschwitz, 2018).

These are only a few examples of the ongoing efforts to provide users with direct access to networks in existing data collections. In our project, we are working with data sources that have not been modeled for HNA before. Our generic data approach is closely connected to associated projects like the US-American cooperative *SNAC – Social Networks in Archival Context*¹⁵ and the French project *PIAAF*¹⁶, which both have a strong focus on archival metadata and full texts.

2.2 Network visualization

Regarding the visualization of data for HNA, many interfaces have been developed over the years that offer explorative web-based network visualization tools for the visually aided historical network analysis. To name a few, the already mentioned *Six degrees of Francis Bacon* (Warren et al., 2016) and *hstoGraph* (Novak et al., 2014) or, in addition, *Visualizing the Republic of Letters* (Chang et al., 2009), *Kindred Britain*¹⁷ or *Deutsche Biographie*¹⁸.

Furthermore, graph visualization is an extensive field in itself, offering a wide range of literature regarding graph-related algorithms (e. g., Gibson et al., 2012; Jacomy et al., 2014; Behrisch et al., 2016), task taxonomies for graph visualizations (e. g., Lee et al., 2006; Ahn et al., 2013; Kerracher et al., 2015), state of the art visualization interaction techniques and developments (e. g., van Ham and Perer, 2009; von Landesberger et al., 2011; Pienta et al.,

¹³<https://www.uibk.ac.at/projects/issues-with-europe/index.html.en>

¹⁴Österreichisches Biographisches Lexikon, see <https://apis.acdh.oeaw.ac.at>

¹⁵<https://snaccooperative.org>

¹⁶Pilote d’interopérabilité pour les autorités archivistiques françaises, <https://piaaf.demo.logilab.fr>

logilab.fr

¹⁷<http://kindred.stanford.edu>

¹⁸<https://www.deutsche-biographie.de>

2015) as well as the use of visual facilitators for the construction of graph queries (e. g., Pienta et al., 2017). Nevertheless, these research and taxonomies mostly address the wider field of graph visualization and, often, visualizations and digital practices that are used for humanistic data are not specifically considering HNA research or humanistic data practices, such as uncertainty, subjectivity or observer-dependence (Drucker, 2011).

2.3 Human-centered design

A key element in the examination and development of a new research infrastructure designed for human-computer interaction is the centrality of the people it is intended to assist. This human-centered approach can be taken on the basis of *Grounded Theory*, which generates inductive results by means of sociological methods (Glaser and Strauss, 1967).

Isenberg et al. (2008) adapted *Grounded Theory* for the evaluation of information visualizations. They suggest iterative evaluation throughout the process of system development using several points of qualitative inquiry to ensure the focus of a system's intended use. This includes field research to examine potential contexts of human interaction with the system. Following this argument for grounded evaluation, the neuralgic points for evaluation in our project are based on Munzner's nested model for visualization design and validation (Munzner, 2009). This allows for iterative improvement of the prototypes. The stages of evaluation include the assessment of possible use cases. On the top level, the problems and data of a particular user domain are investigated. For this, the inclusion of domain-experts in the creation process is becoming a common method in Digital Humanities projects. Chen et al. (2014) present an approach of co-creation through a workshop where participants are asked to create collages to make sense of a photo archive with the aim of creating collection-sensitive interfaces. Henry and Fekete (2006) used a similar participatory approach in the development of a tool for the exploration of social networks, by letting social sciences researchers create paper-prototypes, enabling them to create a list of domain-requirements for their tool, resulting in a prototype with novel features. A thorough evaluation of such co-creation methods, conducted in a co-design process with social science researchers, found that domain experts in general appreciate their additional empowerment in the process and the domain-customized results based on their specific needs. Nevertheless, regarding their personal involvement and necessary time commitment, some participants do not perceive their personal involvement as beneficial for the facilitation of their own research (Molina León and Breiter, 2020). Besides the use of co-design techniques, there also is a shift from merely perceiving the process of visualiz-

ing and visualizations as service tools for humanistic research, towards the acknowledgment of visualization and visualization processes as methodology and facilitator of cross-disciplinary research itself (Hinrichs et al., 2019). While we noticed increased attention to the method of HNA, to the best of our knowledge, there has been little investigation of the modelling and visualization of (bibliographic) big data for this purpose.

3 Networks within heterogeneous data

The interdisciplinary project SoNAR (IDH), which studies the potential of large heterogeneous data collections for HNA, has a runtime of 24 months and includes partners from historiography, information visualization, artificial intelligence and computer sciences as well as information science. This variety of disciplines opens different perspectives on the requirements and challenges connected to the use of heterogeneous (meta)data for HNA. The distinctive aspect of our approach is the synchronous operation of all components of the project, i. e., the design of the data technology, the development of a model research design for HNA and the design and testing of innovative visualization and interface approaches with the involvement of HNA experts are intertwined and influence each other.

The project is based on heterogeneous source data from authority files, bibliographic records and full texts. The data is available in various XML-based formats such as MARC21 (Kruk et al., 2005), EAD (Allison-Bunnell, 2016), and METS/ALTO (Cantara, 2005; Layout, 2016):

- *The Integrated Authority File (GND)*¹⁹ represents and describes 8,295,047 entities, i. e., people, corporations, conferences, geographical areas, technical terms and works;
- *The German National Library (DNB)*²⁰ contains descriptions of bibliographic resources. The data set has 19,926,573 records of books, magazines, newspapers, cards, music, standards, music recordings or audio books;
- *The German Union Catalogue of Serials (ZDB)*²¹ describes newspapers, magazines, serial titles, yearbooks etc. and consists of 1,908,334 records;
- *The Kalliope Union Catalog (KPE)*²² is a collection of personal papers, manuscripts and publishers' archives, which consists of 26,752 records;

¹⁹https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

²⁰https://www.dnb.de/EN/Home/home_node.html

²¹<https://zdb-katalog.de/index.xhtml>

²²<https://kalliope-verbund.info/en/index.html>

- *The Newspaper Information System (ZeFYS)*²³ represents 2,596,641 digitized pages of historical newspapers and full texts;
- *The Exile Press*²⁴ represents German-language exile journals between 1933 and 1945 and consists of 5,336 digitized pages.

Since the source data – describing entities (authority files) and resources (bibliographic files) – is encoded in various formats, in a first step these formats are analyzed to enable the design an appropriate data model and transformed into a uniform, generic format. Full texts are also prepared for automatic enrichment (i. e., named entity recognition and linking) and converted to a corresponding format.

4 Data processing

In this Section, we will give an overview of the data transformation and graph modeling process as well as the challenges we encountered.

4.1 Data model

The technical goal of the project is the integration of the various source data sets into a common research infrastructure. We currently use the graph database *Neo4j*²⁵, which is well suited for the efficient storage and high-performance analysis of large amounts of highly networked information (Efer, 2016; Matschinegg and Nicka, 2018; Wintergrün, 2019). Entities are modeled as nodes and relations as edges with absolute and relational features.

There are a total of 9 entity types²⁶ extracted from the source data:

1. Person `PerName`;
2. Corporate body `CorpName`;
3. Place or geographic name `GeoName`;
4. Conference or event `MeetName`;
5. Subject heading `TopicTerm`;
6. Work `UniTitle`;
7. Temporal information `ChronTerm`;
8. Information about ISIL²⁷ `IsilTerm`;
9. Resource `Resource`.

Six entity types (i. e., person `PerName`, corporate body `CorpName`, place or geographic name `GeoName`, conference or event `MeetName`, subject heading `TopicTerm`,

²³<http://zefys.staatsbibliothek-berlin.de/index.php?id=start&L=1>

²⁴https://www.dnb.de/EN/Sammlungen/DEA/Exilpresse/exilpresse_node.html

²⁵<https://neo4j.com>

²⁶Names of the entity types for *Neo4j* are preliminary and can change.

²⁷International Standard Identifier for Libraries and Related Organizations

and work UniTitle) are taken from the corresponding classes of the authority files. Bibliographic entities are represented as Resource. We added two types: ChronTerm describes temporal information encoded in entity types from authority files; and IsilTerm is used to identify the libraries related to other entity types. Each entity has general features, such as unique source identifier, URI, name, link, etc., and specific features, such as age, gender, coordinates, etc. Furthermore, there are also nine relation types that correspond to entity types, such as RelationToPerName, RelationToCorpName, RelationToGeoName. Relations between entities have information about the relation source, relation source type, information about temporal validity, and additional information.

While the relations between entities are explicitly described in authority files, in bibliographic files relations between actors such as person or corporate body that are identified or defined in the resource are only implicitly encoded. Our aim is to derive these implicit connections based on rules and to make them available as explicitly encoded data. In order to derive corresponding relation types, the role of actors regarding a specific resource (e. g., is author, editor, addressee) and the resource type (bibliographic files of primary sources of *the Kalliope Union Catalog* and of secondary sources of *the German National Library* and *the German Union Catalogue of Serials*) are taken into account. Based on this, we infer additional relations, for instance between co-authors, co-publishers, and authors/addressees etc.

In order to exploit full texts for scientific analyses, named entities are automatically recognized, disambiguated, and linked to their associated authority files (e. g., *the Integrated Authority File* or *Wikidata*²⁸). Then, also automatically, relations between detected entities are recognized. These are added to the graph database and connected with their respective full texts, represented as nodes.

At the same time, we are experimenting with linked data as an alternative approach for our tasks and needs. Here, the source data is modeled in the form of subject–predicate–object expressions and stored in *GraphDB*²⁹. This approach simplifies the integration of linked open data datasets (*Wikidata*, *DBpedia*³⁰, *GeoNames*³¹ etc.), and provides more sophisticated inference possibilities. In the preliminary comparison of the two approaches, *GraphDB* also shows better performance. The only disadvantage is that the source data has to be remodeled, for example, to display relation features such as relation type, relation source type, temporal validity.

²⁸<https://www.wikidata.org>

²⁹<http://graphdb.ontotext.com>

³⁰<https://wiki.dbpedia.org>

³¹<https://www.geonames.org>

We have modeled and transformed data for the graph database in such a way that identifiers are used as coordinates for relations between entities. In the *Integrated Authority File*, entities with old identifiers were found, so that an appropriate connection of two entities was not possible. The first chal-

lenge was to detect old identifiers and replace them with valid ones to enable a representation that is free of errors. All replacements were written in a log file. However, during a consistency check we also found relations to entities within the source data, which were without identifiers. Since such entities could not be clearly assigned to existing entities with identifiers, ambiguous relations of this kind had to be ignored.

Information that was encrypted in internal codes in the *Integrated Authority File*, the *German National Library*, and the *German Union Catalogue of Serials* (in format MARC21) was also checked, i. e., for codes of general and specific entity types, codes of relation types between an agent and a resource, and country codes. Further examinations were performed on the consistency of entity names, resource titles, and identifiers. All errors or inconsistencies were written in a log file.

Building on the conclusions we drew from testing *Neo4j*, we decided to adapt the data model to our needs. In order to simplify searching and filtering according to the temporal dimension, time information from the source data must be adjusted. First of all, while retaining the source data, we will additionally separate time intervals, noted as “begin” and “end”. Secondly, we will add a feature to resource descriptions that reflects the year of publication (in addition to the publication date). Thirdly, time expressions that differ in form in MARC21 and EAD will be normalized.

We also decided to change gender-specific names of professions. These are represented in the *Integrated Authority File* as two different entities with their own identifiers, i. e., male and female. Conceptually, however, it is one single entity with two versions, so these versions must be merged in the graph database and represented as a node. One challenge is to adequately display all information from the two versions without making the search more difficult. We are currently looking for a suitable solution.

5 Co-design workshop

Based on the practice of grounded evaluation (Isenberg et al., 2008), we aim to integrate domain experts closely into the data modelling as well as visualization process. Having HNA experts in our project team, all internal decisions are made taking the domain perspective into account. Besides that, the inclusion of external domain experts is another integral part of our research design. By conducting studies with researchers of various fields who are using the method of HNA, our aim is to improve the project’s outcome iteratively.

In the beginning of the project, it was important to stimulate discussions on the potential of bibliographic (meta)data for HNA as well as require-

ments for the visualization of historical networks. In order to identify the most central aspects to consider, we organised an initial co-design workshop. To gain new insights into historical network research and visualization, we invited domain experts, following the approach by Chen et al. (2014) and Henry and Fekete (2006).

5.1 Procedure

Ten persons participated in the co-design workshop, including four historical/social network practitioners as domain-experts, two project-internal information visualization designers/engineers, two people from our project-internal evaluation team, one person from our team of data scientists (responsible for the data transformation) and another external participant with background in design and experience with the co-design format. The goal of the interdisciplinary composition of the workshop group was to foster the discussion by offering a multitude of perspectives on the topic of HNA through the lens of HNA experts as well as fresh insights through the perspective of participants from other (project-relevant) domains. We aim to develop an infrastructure for HNA that can be used by researchers of all disciplines working with this method, so the participation of experts from fields other than history was explicitly welcome.

The workshop was scheduled for three hours in total. Similar to the approach by Fekete and Plaisant (2002), we started off with a small presentation of a broad range of recent network visualization possibilities and developments including some more novel and experimental approaches but without much focus on details.

In order to facilitate the process of conceptualizing network visualizations, we started the hands-on process with a short visualization exercise during which the participants were asked to visualize a very small social network (ten nodes) based on a data matrix we provided. After this warm-up, we gave a short introduction about the goals of our research as well as the data in our project. Afterwards, each participant was asked to create a collage about approaches for HNA research with our data and project in mind (see Fig. 2). For the collages, we provided a variety of materials (e. g., colorful paper, pencils and markers, sticky-notes). While Chen et al. (2014) provided visual material from their photographic collection, our data is more abstract and less visual. Therefore, to further support the collaging process with visual aids, we printed out and distributed further visual material including prints of an empty map, printed icons (e. g., as representations of network nodes) and a few printed scans from our full text data sources. We provided a few questions to kickstart the process, such as *“How would you like to move through*

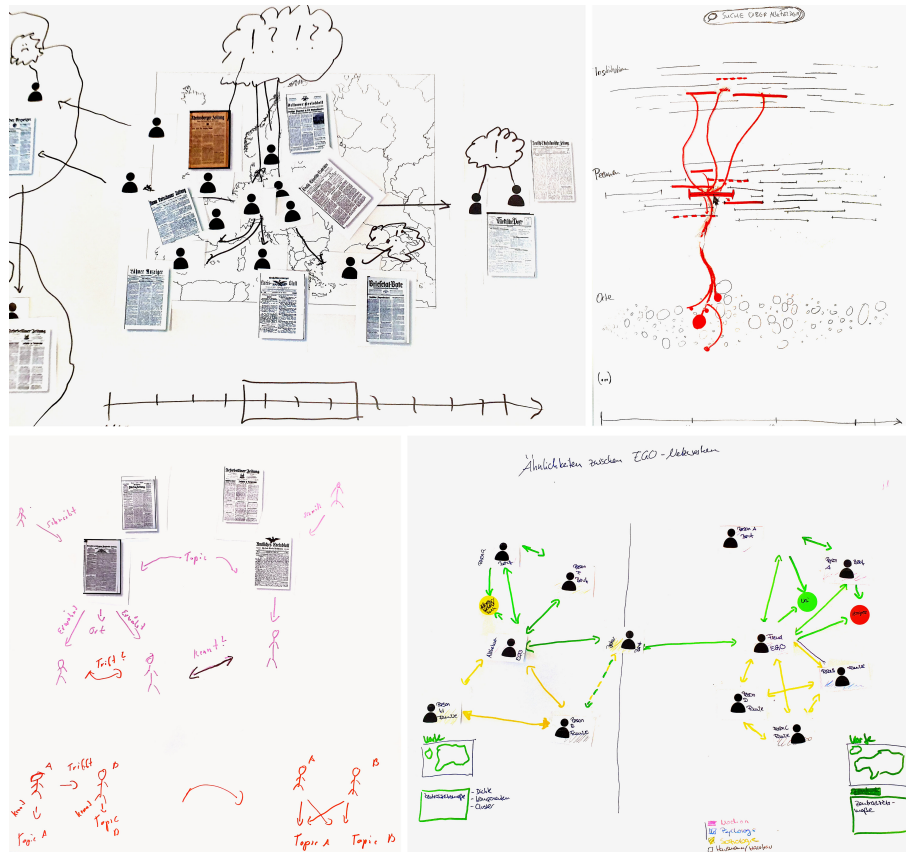


Figure 2: Selected collages created in the interdisciplinary co-design workshop.

the data?” – “*What role do data dimensions such as time, space or semantic relationships play?*”, but encouraged them to feel free to disregard them. In general, the task of the collaging-process was not to create wireframe-like sketches for a concrete user interface solutions but to envision general approaches, functionalities and entrance points to HNA research and our data.

After about 30 minutes, all collages were individually discussed in the plenary. First, the participants not involved in a collage were asked to interpret and speculate about what they were seeing. Afterwards, the creators of the collages were asked to give explanations and discuss their approach with the group. In this step, the typically occurring mis-interpretations of the results are meant to foster further discussions and novel ideas. In the last and recapitulating step, each participant was asked to give a closing statement about the most important insights from the process as well as the themes of the discussion most prominent to them.

For further analysis and documentation, the entire workshop was audio

recorded and the process was documented using still photography. The audio recordings were transcribed and encoded in a tool for qualitative data analysis. This allowed us to assess the scope of different qualitative aspects of the workshop discussions. The goal of the workshop was not to create functional wire frames or concrete interaction principles, but to stimulate discussions, attain a sensibility towards the domain and data, and highlight important domain-specific research aspects and challenges. The following Section will highlight some of the most relevant insights for our visualization process.

5.2 Results

We noticed two different kinds of statements. On a more abstract level, the participants expressed various information needs that commonly arise in the process of their research. However, in some cases, the conversation and the collages provided very concrete ideas on possible features an HNA infrastructure could offer in order to address these needs. As mentioned before, the latter were not regarded as immediate assignments for the visualization process, but rather as indicators for the participant’s reception attitude towards the user interface of an HNA infrastructure. Table 1 and 2 summarize the main aspects of the workshop discussions of both the needs and features.

| Need | Number of Mentions | Persons |
|-------------------|--------------------|---------|
| New Perspectives | 30 | 7 |
| Uncertainty | 25 | 7 |
| Data Potential | 27 | 4 |
| Graph Density | 26 | 4 |
| Entry Points | 17 | 4 |
| Data Explanations | 16 | 4 |

Table 1: List of the most frequently expressed needs by the workshop participants with the count of their mentions during the workshop and the number of persons ($n=10$) referring to them.

The most central topic in the discussions was the prioritized types of approaches supported by the infrastructure. Seven of the ten participants expressed the hope for *new perspectives* the visualizations could generate and thus create access to the data, which is hardly possible through non-machine-supported cognitive work. In this context, one participant explicitly emphasized the potential of visualizations to raise new questions:

“What kind of relationships you are looking for in the data, you of-

| Feature | Number of Mentions | Pers. |
|---------------------|--------------------|-------|
| Timeline | 22 | 4 |
| Tie Metrics | 18 | 4 |
| Other Filters | 16 | 3 |
| Export and Citation | 13 | 4 |
| Location Filter | 8 | 3 |
| Source Linking | 6 | 4 |

Table 2: Most frequently desired features with the count of their mentions during the workshop and the number of persons ($n=10$) referring to them.

*ten notice in the very moment you look at the pile for the first time.*³²

Since the participants were introduced to the fact that we have a very large amount of data, which can hardly be presented in its entirety (see Section 3), the discussion of possible entry points emerged. There was consensus about the importance of filter options, most importantly time filters.

“Without timelines, the visualizations are of no use to me – neither for the analysis nor for the presentation of results.”

In addition to timelines, other filters (e. g., node type and node source) were considered a prerequisite for data exploration. Three participants also mentioned the importance of location filters, e. g., through a map view.

Participants with more HNA experience explicitly stressed the essential role of a multi-layer approach. The capacity to display the evolution of relations (e. g., through time and location) was described to be the distinctive factor of the HNA method towards non-historical analysis of social networks. The sole option of static display was considered insufficient.

Along with possible entry points, another topic of discussion was data complexity. Introductions and explanations about the underlying data were considered to be crucial. Some participants suggested to address this with concrete use cases that could provide potential users with a more specific idea of the possibilities of the HNA infrastructure.

About half of the participants mentioned the ability to quantify network characteristics as graph metrics during the research process as a main motivation for using HNA methods. This includes indicators such as the clustering coefficient, closeness centrality, degree distribution, degree centrality, and betweenness centrality. Four participants also mentioned *density* within a selected sample of nodes to be a relevant indicator for the *data potential* to

³²All quotes translated from German into English.

refer to the possibilities a data set affords for network analysis. Following the first cluster of possible approaches, one participant highlighted the added value of graph metrics to the identification of anomalies in the data.

“What all these things are actually about: We are looking for patterns!”

Additionally, some participants stressed the potential of *tie metrics* to accommodate a variety of relation types and expressed the desire to have the weight of edge properties visualized.

“It is of course a big difference whether you are a family member [...] or whether you are a correspondence partner or whether you met at a congress during a coffee break. These are all relationships, but of course they have different weights in their interpretation. This is, for example, something we would like to see in the visualization.”

This statement is representative of another central topic discussed in the workshop, namely the visual marking of missing or uncertain information in the data, e. g., caused by inconsistencies in the metadata fields (see Section 4.2). The design expert considered this to be a desideratum.

“I think this is not done enough in current visualizations to show uncertainties of data.”

Other aspects mentioned in order to meet the scientific standards of HNA research, were those connected to *export and citation* of the visualizations. This, of course, requires unambiguous and persistent provenance links to the source of each data point as well as timestamps of the corresponding data import.

Many of the results from the co-design workshop match with current challenges in information visualization described in the literature. In the following Section, informed by the workshop results, we describe our prototyping approach and process.

6 Visual prototyping

The data set that we consider for this research comprises an amount of elements that far outreaches what can be perceptually or cognitively grasped at one glance. The amount of nodes and relations poses technological as well as visual challenges regarding the encoding (Fekete and Plaisant, 2002; Shneiderman, 2008). While some potential users of our technology might

have a fixed research question in mind, others – related to the phenomenon of serendipity (Thudt et al., 2012) – might want to use such an infrastructure in order to formulate questions. Our aim is to provide access points for a broad variety of motivations and research questions, including those we cannot anticipate yet. Therefore, the conceptualization of a visual representation as an access point to our data in the form of a data exploration interface can be described by a wide and diverse range of challenges and difficulties:

- *How to visualize tens of millions of nodes and hundreds of millions of edges?*
- *What are possible and meaningful entrance points to the data?*
- *How can we deal with uncertainty, missing data and varying data sources?*
- *How can we deal with multiple data dimensions?*
- *How can we provide a technology that is complex and open enough for a broad range of undefined research questions but simple enough to get casually used?*
- *How can we be transparent regarding used algorithms?*
- *How to move between overviews, detail views and egocentric views?*

Even though our workshop results, upcoming interviews and evaluation with domain-experts and existing task taxonomies (e. g., Lee et al., 2006; Keracher et al., 2015; Ahn et al., 2013) already offer a multitude of potential requirements, tasks and needs that should be addressed in our graph technology, we additionally see the prototyping process as a form of research through design (Zimmerman et al., 2007) that might confirm these requirements or even unveil new ones. Furthermore, in contrast to the mentioned task taxonomies, we are dealing with humanistic data and humanistic related research questions, where traditional visualization approaches are oftentimes considered unfit to the nature of the underlying data and research Drucker (2011). Therefore, concurrently to the data modeling process and besides the mentioned co-creation approaches, our visualization process can be described as a form of experimental and iterative rapid prototyping process and data exploration. In contrast to taking the potentially shortest path to a finished *tool*, our method resembles a curiosity-driven ‘sandcasting’ (Hinrichs et al., 2019). We understand experimental approaches and detours in the visualization process itself as a form of methodology for knowledge generation. By following this approach, visualizations created in the process are not necessarily created with the goal of adopting parts of them in a final prototype or concept, but also as a method to explore the data or individual facets of the data, for investigation of general challenges with the data or their encoding, or as a form of visual facilitator for cross-disciplinary

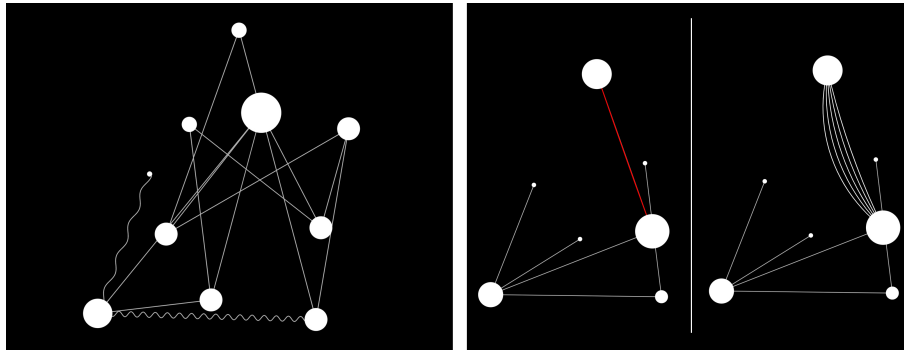


Figure 3: Two small design studies. Left: visualizing levels of uncertainty in edges between nodes by using waves and varying levels of frequency. Right: Concept for handling of multiple edges between two nodes. In the initial view multiple edges are combined into one edge (marked as the red line) to reduce overall complexity of a graph. A click allows to fan out the individual edges on demand, visually transitioning from one line to multiple arcs.

communication, provocation and novel approaches (Hinrichs et al., 2019).

Furthermore, the whole project started in an interdisciplinary and concurrent mode from the beginning, without delays between individual steps; data processing, case study developments, visualization and evaluation are taking place in parallel. Therefore, in the beginning of the project the data was neither processed for visualization nor was it accessible via some form of API, making it only possible to work with small subsets of selected data. While this makes it difficult to anticipate all facets and challenges of the real data, working with data subsets early on also lead to the possibility of having iterative influence on the data processing and the data model.

Instead of trying to combine all potential features and ideas in one prototype, in our process we gradually focus on many small separate problems and ideas through a multitude of many rough prototypes. We develop many design studies or prototypes in close collaboration with our HNA experts or based on results from the workshop, interviews or general evaluation; others are generally more experimental and result from spontaneous impulses. The upcoming examples were mainly designed with the data visualization library D3.js (Bostock et al., 2011), allowing the development of customized visualizations.

Figure 3, for instance, shows two small design studies from the beginning of the project, without using real data: the first one (left) dealing with visualization of levels of relation uncertainty and the other one (right) testing an interaction concept with the goal of reducing complexity by merging multiple edges and allowing to fan them out on demand.

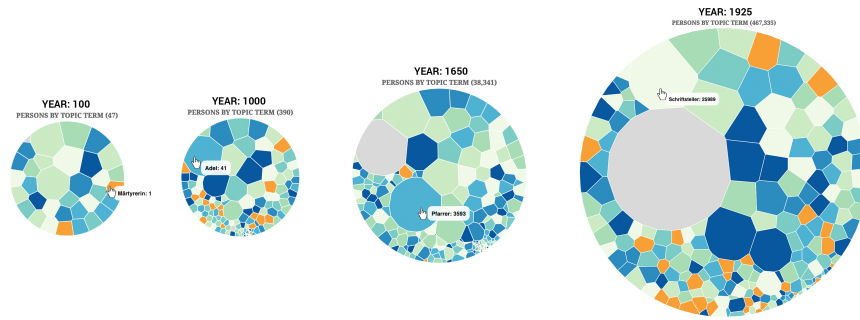


Figure 4: Prototype overviews of a specific data facet (here: topic terms related to Persons), based on a selected year. A Voronoi map displays distributions of topic terms connected to alive persons in a selected year. Orange represents female gendered terms.

As an example for when visualization influenced the data model afterwards, an early prototype, that clusters persons in a small subset of the data based on related topic terms (in most cases job titles), revealed that job titles in our base data (GND) are oftentimes gendered³³ and therefore men and women are oftentimes not related to the same topic term, even though they practice the same job. We did not expect this differentiation in the data and it is very relevant for search queries and the visualization, since there might be many cases where researchers do not differentiate by gender and may only use the male form that is traditionally considered to be generic. An effect of this differentiation in the data can be seen in another interactive prototype (see Fig. 4), where it is possible to select a specific year in the data with a slider, visualizing top topic terms related to person alive in this selection (female gendered topic terms are colored in orange). The goal of this prototype was to explore the potential of overviews to reveal specific aspects of the data that later might act as entry points for specific search interests or details on demand.

Another experimental prototype (see Fig. 5) of a small subset of our data also focuses on topic terms and temporality of the data, an aspect oftentimes mentioned by some of our HNA experts in the workshop. Here, the dimensionality reduction technique UMAP (McInnes et al., 2018) was used to map persons with similar topic term relations close to each other, effectively forming clusters for occupational domains (e. g., authors are clustered close to each other). A timeline on the right displays the general distribution of all nodes and a list next to it, displays all connected topic terms, ordered by oc-

³³Many German job terms exist in a male and a female gendered version, as with the English ‘actor’ and ‘actress’.

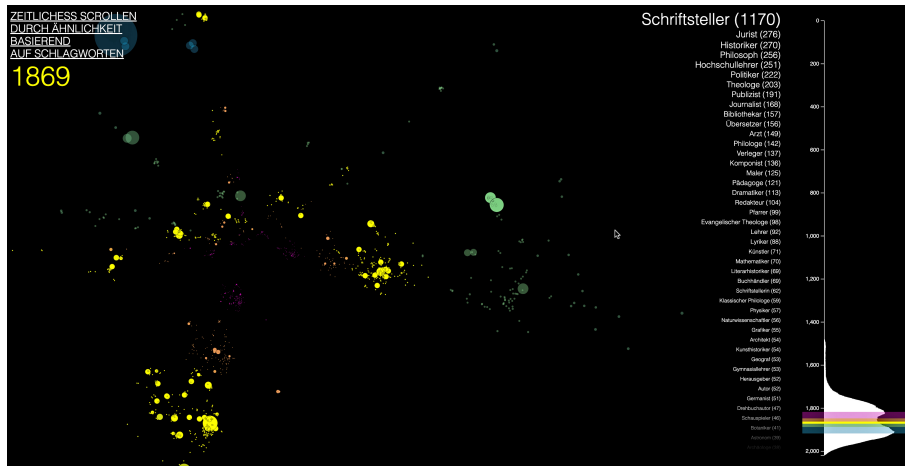


Figure 5: Experimental prototype that enables scrolling through time through a UMAP projection of a small subset of our data that arranges persons based on similarity across topic terms. Color and the sagittal (z) axis are used to encode temporal closeness of a node in relation to a selected year (in this example nodes that lie inside the selected year 1869 are colored in yellow).

currence. Scrolling enables a user to move through the temporal dimension of the network, aiming for the impression of moving through a time tunnel. Nodes that are part of a selected year are displayed in yellow. Temporally close nodes in the past are perspectively further away from the viewer and in red tones. Nodes that are temporally close but in the future are colored in green and blue tones and are perspectively closer to the viewer. One insight of this prototype was that the data model and processing again might need to be adjusted, to help make the data more accessible for use in visualizations, especially in regard to temporal filtering.

In some cases, as with Figure 6, we also developed prototypes out of curiosity for small specific research questions in mind, for example, “*are network communities in the data subset mostly composed of contemporary nodes or do communities stretch over multiple generations?*” Here, the prototyping process helped to test specific algorithm implementations and design strategies, while at the same time being able to provide deeper insights into the data.

While our research is still in progress, the experiences mentioned above illustrate the benefits of staying open for experimentation and curiosity throughout the analysis and visualization process. Even though many ideas and concepts oftentimes result from existing related work and, of course, the experience of our domain experts, we see additional value in experimenting with the data and generating a multitude of visual representations, even if

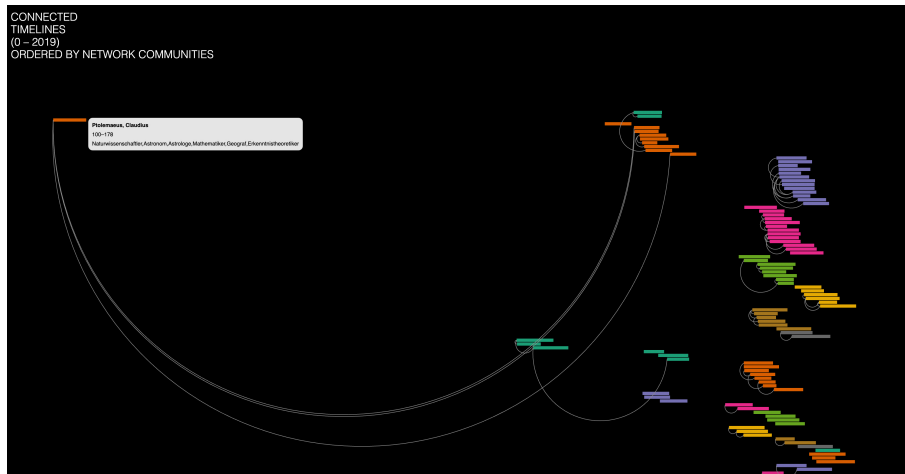


Figure 6: Prototype overviews of node relations to reveal relations and community clusters over time. First a community algorithm runs over the graph data. Afterwards nodes are ordered and colored based on the community algorithm results and the nodes are placed on a timeline, based on their dates of birth and death.

this means to knowingly take detours. It is these more experimental detours that can lead to new ideas for tools or insights into the data. The prototypes are non-incremental steps towards a final concept, iteratively informed by feedback from our evaluation with domain experts and other potential operators of a future tool.

7 Conclusion and further work

The converging of multiple heterogeneous data sources containing millions of nodes and edges for a graph-based research infrastructure, which enables historical social network analysis poses a multitude of multidisciplinary challenges, such as:

- Challenges that occur during the merging of heterogeneous data sources;
- Performance towards the given scope of the data,
- Creation of domain-customized interfaces, which are open and flexible with regard to unforeseen research questions,
- Integration of domain knowledge into the process,
- Visualization of millions of data points to provide explorable access points in addition to search interfaces.

We face these challenges by focusing on the tight and interdisciplinary collaboration and constant evaluation during the whole research and develop-

ment process between historical network experts, data visualization researchers, data scientists and experts on the evaluation of information infrastructures. Here, an initial co-design workshop with additional external HNA practitioners and other domain experts was used as one example to illustrate the collaboration within the project. Building on the contextual data from the co-design workshop, we will continue to follow a human-centered approach towards data modeling and visualization design.

In our next step, we aim to take a closer look at the individual process behind historical network research in one-on-one interviews with more domain experts on their approach to current HNA research. After finalizing the data model, next steps also include merging multiple visualization concepts into one prototype, joining views on the level of greater global overviews on our data with local views on specific individual networks inside it. Furthermore, we will provide exemplary use cases on a variety of historical topics making use of our data and the interface in collaboration with our HNA experts.

In this paper, we described the process of examining the potential of re-modeling and merging (bibliographic) big data from cultural heritage institutions into one single gathering point which is optimized for the use in historical network analysis. By providing insights into emerging challenges in the project as well as our approach to their solutions, we hope to foster more research and exchange in and with similar HNA related projects.

8 Acknowledgements

We would like to thank our participants for their attendance of the co-design workshop. Furthermore, we want to thank our project partners Heiner Fangerau, Thorsten Halling, Hans-Jörg Lieder, Gerhard Müller, Clemens Neudecker, David Zellhöfer and Josefine Zinck. This research is part of the research project SoNAR (IDH) and is funded by the DFG – German Research Foundation (project no. 414792379).

References

- Ahn, J.-w., Plaisant, C., and Shneiderman, B. (2013). A Task Taxonomy for Network Evolution Analysis. *IEEE transactions on visualization and computer graphics*, 20(3):365–376.
- Allison-Bunnell, J. (2016). Review of Encoded Archival Description Tag Library – Version EAD3. *Journal of Western Archives*, 7(1):6.
- Auer, S. and Mann, S. (2019). Towards an Open Research Knowledge Graph. *The Serials Librarian*, 76.

- Behrisch, M., Bach, B., Henry Riche, N., Schreck, T., and Fekete, J.-D. (2016). Matrix Reordering Methods for Table and Network Visualization. In *Computer Graphics Forum*, volume 35 (3), pages 693–716. Wiley Online Library.
- Bostock, M., Ogievetsky, V., and Heer, J. (2011). D³ Data-Driven Documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309.
- Cantara, L. (2005). METS: The Metadata Encoding and Transmission Standard. *Cataloging & classification quarterly*, 40(3-4):237–253.
- Chang, D., Ge, Y., Song, S., Coleman, N., Christensen, J., and Heer, J. (2009). Visualizing the Republic of Letters. *Stanford: Stanford University*. Retrieved April, 21:2014.
- Chen, K.-I., Dörk, M., and Dade-Robertson, M. (2014). Exploring the promises and potentials of visual archive interfaces. In *iConference 2014 Proceedings*. iSchools.
- Drucker, J. (2011). Humanities Approaches to Graphical Display. *DHQ: Digital Humanities Quarterly*, 5(1):1–21.
- Efer, T. (2016). *Graphdatenbanken für die textorientierten e-Humanities*. PhD thesis, Universität Leipzig, Augustusplatz 10, 04109 Leipzig.
- Fekete, J. and Plaisant, C. (2002). Interactive Information Visualization of a Million Items. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, pages 117–124.
- Gibson, H., Faith, J., and Vickers, P. (2012). A Survey of Two-Dimensional Graph Layout Techniques for Information Visualisation. *Information Visualization*, 12(3-4):324–357.
- Glaser, B. and Strauss, A. (1967). The Discovery of Grounded Theory. 1967. *Weidenfeld & Nicolson, London*, pages 1–19.
- Henry, N. and Fekete, J.-D. (2006). Matrixexplorer: a Dual-Representation System to Explore Social Networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):677–684.
- 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):i80–i99.

- Isenberg, P., Zuk, T., Collins, C., and Carpendale, S. (2008). Grounded Evaluation of Information Visualizations. In *Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization*, BELIV '08, New York, NY, USA. Association for Computing Machinery.
- Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PloS one*, 9(6):e98679.
- Jansen, D. and Wald, A. (2007). *Netzwerktheorien*, pages 188–199. VS Verlag für Sozialwissenschaften, Wiesbaden.
- Kerracher, N., Kennedy, J., and Chalmers, K. (2015). A Task Taxonomy for Temporal Graph Visualisation. *IEEE transactions on visualization and computer graphics*, 21(10):1160–1172.
- Kruk, S. R., Synak, M., and Zimmermann, K. (2005). MarcOnt – Integration Ontology for Bibliographic Description Formats. In *International Conference on Dublin Core and Metadata Applications*, pages 231–234.
- Layout, A. (2016). Text Object (ALTO) XML Schema. *Online*: <http://www.loc.gov/standards/alto>.
- Lee, B., Plaisant, C., Parr, C. S., Fekete, J.-D., and Henry, N. (2006). Task Taxonomy for Graph Visualization. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–5.
- Matschinegg, I. and Nicka, I. (2018). REALonline Enhanced. Die neuen Funktionalitäten und Features der Forschungsbilddatenbank des IMAREAL. *MEMO 2*, Digital Humanities & Materielle Kultur:10–32. doi: 10.25536/20180202.
- McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.
- Molina León, G. and Breiter, A. (2020). Co-creating Visualizations: A First Evaluation with Social Science Researchers. *Computer Graphics Forum*, 39.
- Munzner, T. (2009). A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928.

- Novak, J., Micheel, I., Melenhorst, M., Wieneke, L., Düring, M., Morón, J. G., Pasini, C., Tagliasacchi, M., and Fraternali, P. (2014). HistoGraph – A Visualization Tool for Collaborative Analysis of Networks from Historical Social Multimedia Collections. In *2014 18th International Conference on Information Visualisation*, pages 241–250. IEEE.
- Pienta, R., Abello, J., Kahng, M., and Chau, D. H. (2015). Scalable Graph Exploration and Visualization: Sensemaking Challenges and Opportunities. In *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 271–278. IEEE.
- Pienta, R., Hohman, F., Tamersoy, A., Endert, A., Navathe, S., Tong, H., and Chau, D. H. (2017). Visual Graph Query Construction and Refinement. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1587–1590.
- Purschwitz, A. (2018). Netzwerke des Wissens - Thematische und personelle Relationen innerhalb der halleschen Zeitungen und Zeitschriften der Aufklärungsepoche (1688-1818). *Journal of Historical Network Research*, pages 109–142 Pages.
- Rehm, G., Berger, M., Elsholz, E., Hegele, S., Kintzel, F., Marheinecke, K., Piperidis, S., Deligiannis, M., Galanis, D., Gkirtzou, K., Labropoulou, P., Bontcheva, K., Jones, D., Roberts, I., Hajic, J., Hamrlová, J., Kačena, L., Choukri, K., Arranz, V., Vasiljevs, A., Anvari, O., Lagzdinš, A., Melnik, J., Backfried, G., Dikici, E., Janosik, M., Prinz, K., Prinz, C., Stampler, S., Thomas-Aniola, D., Pérez, J. M. G., Silva, A. G., Berrío, C., Germann, U., Renals, S., and Klejch, O. (2020). European Language Grid: An Overview. In Calzolari, N., Béchet, F., Blache, P., Cieri, C., Choukri, K., Declerck, T., Isahara, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3359–3373, Marseille, France. European Language Resources Association (ELRA).
- 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*, pages 3–12. ACM.
- Thudt, A., Hinrichs, U., and Carpendale, S. (2012). The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries through Information Visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1461–1470.

- 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):953–960.
- von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J. J., Fekete, J.-D., and Fellner, D. W. (2011). Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges. In *Computer graphics forum*, volume 30 (6), pages 1719–1749. Wiley Online Library.
- Vorndran, A. (2018). Hervorholen, was in unseren Daten steckt! Mehrwerte durch Analysen großer Bibliotheksdatenbestände. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB*, 5(4).
- Warren, C. N., Shore, D., Otis, J., Wang, L., Finegold, M., and Shalizi, C. (2016). Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *DHQ: Digital Humanities Quarterly*, 10(3).
- Wintergrün, D. (2019). *Netzwerkanalysen und semantische Datenmodellierung als heuristische Instrumente für die historische Forschung*. doctoralthesis, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).
- 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*, pages 493–502.