Project Description – Project Proposals in the Area of Scientific Library Services and Information Systems (LIS)

LIS Funding Programme or Call: e-Research Technologies (Implementation)

SoNAR. Social Network Analysis and related Research.

- Prof. Dr. Achim Bonte, Berlin, Director General, Berlin State Library – Prussian Cultural Heritage Foundation (Staatsbibliothek zu Berlin – PK, SBB-PK)
- Prof. Vivien Petras, PhD, Berlin, Director, Berlin School of Library and Information Science, Humboldt University Berlin (Humboldt-Universität zu Berlin, IBI/HU)

Project Description

1 Starting Point

The Berlin State Library and the Berlin School of Library and Information Science jointly propose the implementation of an advanced research data market for Historical Network Research (HNR), named *SoNAR, Social Network Analysis and related Research*. The proposal builds on a concept elaborated during the preliminary, successfully evaluated pilot project[1]. SoNAR will leverage HNR as it addresses a key obstacle: access to suitable data. Since the 1940s, Social Network Analysis has gained widespread acceptance in the Social Sciences. Over time, methods and theories have been adapted for historical research, giving rise to HNR – a promising interdisciplinary approach to investigate the past today (Petz 2021, 33 ff.). Yet, HNR has remained somewhat niche, due to the labor-intensive process of acquiring a data set that is large enough for a research topic. Robert Gramsch-Stefest notes that while there are also small forms of network research that offer much through visualization techniques, the manual curation of a critical amount of data hinders HNR from reaching its full potential (2020, 9). This is where the project steps in: SoNAR will scale the gathering of network data by integrating extensive heterogeneous data sets from cultural heritage and research communities to infer the statements about agents and agent relations. The resulting social knowledge graph will significantly reduce both data acquisition and data sharing efforts of research projects like *PDB18*[2] or *Correspondence of Early Romanticism*[3]. In addition, SoNAR will help to further data services of research service providers like libraries and archives and to spread data science methods in History, (Digital) Humanities, and the Social Sciences.

The mass digitization campaigns of recent decades contributed to lower access barriers to cultural heritage resources and paved the way for new digital methods to study the past. Notwithstanding, curating data sets for a specific research topic remained a challenge as the data are still randomly isolated, even hidden across data repositories. SoNAR, designed as a unifying access point, can be described best as a research data market (Putnings 2021, 143) with a social knowledge graph at its heart and a focus on three key services elaborated with the pilot project (Appendix 4):

1. ***Data integration:*** 1) to aggregate, archive, and transform data, 2) to cluster same entities identified by different identifiers of various authority files, and 3) to infer explicit and implicit statements about agents and agent relations from aggregated data sets. The service *data integration* will ensure that metadata about data provenance and processing are available.

---

[1] SoNAR (IDH). 2019-2021. https://gepris.dfg.de/gepris/projekt/414792379, https://github.com/sonar-idh [2024-04-05]

[2] Portal "Der deutsche Brief im 18. Jahrhundert". https://www.pdb18.de [2024-04-05]

[3] Prof. Dr. Aline Deicke, project manager of the DFG-funded *Correspondence of Early Romanticism*, will accompany the integration of research data (Appendix 2). https://correspondences-early-romanticism.uni-mainz.de [2024-04-05]

2. ***Data access:*** to explore, curate, and download network and respective provenance data with data science solutions like R. SoNAR will offer an Application Programming Interface (API) for that purpose. A User Interface (UI) will assist with curating network data sets and substantiating research questions with preconfigured data visualizations and statistics.
3. ***Research support:*** to incorporate basic, essential research requirements like traceability of network data to an original data source, reproducibility of downloaded historical network data sets, and publishing network data sets with research papers by also assigning Digital Object Identifier (DOI)[4].

These services were derived from use cases established with an in-depth analysis of stakeholder demands that govern the objectives of the proposed implementation project (ch. 2.2).

The project team will implement lean project management including SCRUM methods (ch. 2.3.1) to ensure the implementation will meet the needs of stakeholders: research and cultural heritage communities but also of SBB-PK responsible for sustainably operating SoNAR. The work program (ch. 2.3.2) is broken down into two work packages: The development is focused in WP1, and the supporting activities, in particular evaluation and community relations, are allocated to WP2. Both work packages are interrelated through the common objective of developing SoNAR with the user value at its center. The project takes measures accordingly to collect and evaluate user feedback:

1. Interviews, standardized surveys, and usability studies,
2. Community consultations like workshops and conferences, and
3. Use case analyses for best practice tutorials and educational materials.

Feedback on (interim) results will be carefully weight to maximize user experience and user value. However, feedback out-of-scope of SoNAR will be documented in a pool of topics for requirement analysis by subsequent projects. An advisory board will support decision making (Appendix 2).

The initial implementation will integrate data of well-established data providers who combine over 55 million records describing cultural heritage resources and related entities: meta- and authority data of libraries, archives, and museums. HNR data sets stored in research data repositories and Wikidata will complement integrated knowledge about agents and agent relations (Appendices 3, 5). Moreover, the project should continue the evaluation of latest technologies for extracting and integrating linked data about agents and relations from text-digitized resources like newspapers (WP2.3): While the pilot project focused on evaluating the quality of entity recognition and linking (Menzel 2021), this project will give attention to the tasks and efforts related to implementing and operating artificial intelligence / machine learning technologies to extent the knowledge graph.

The concept underlying SoNAR matured over the last decade after first ideas were discussed at a conference on personal papers and manuscript collections at the SBB-PK in 2014[5]. The project proposal originates from an implementation concept, analysis of cultural heritage ontologies, and visualization concepts – such as unfolding edges to visualize multimodal relations (Bludau et al. 2023) – developed during the pilot project (Appendix 4). The project results were discussed at multiple occasions (e.g. Müller 2022a, b) and have received encouraging feedback. The scope of SoNAR was reevaluated and partly consolidated with the proposal development (Appendices 5, 6). Furthermore, SBB-PK also completed a self-funded internal project to specify tasks and efforts associated with operating SoNAR in 2023. The operational concept covers maintenance and

[4] The Berlin State Library will open EMPORION, a research data repository for social and economic history, for Historical Network Research (data sets and related papers): https://emporion.gswg.info [2024-04-05]

[5] https://kalliope-verbund.info/files/HANSKPEKonferenz2014-finalPubl.pdf [2024-04-05]

advancement of the research data market (Appendix 7). The pivotal approach focuses on reducing average support and maintenance expenses by reusing the components for the portfolio of data-centered services that are operated by SBB-PK. The future advancement, in turn, requires sustaining high user value and will be thus coordinated with follow-up research and development projects. Minor feature updates will be considered with the maintenance of components.

The applicants possess substantial knowledge about data, infrastructure, and related standards of both the cultural heritage and research communities. SBB-PK and IBI/HU have been devoting considerable efforts to raising awareness about research data, data literacy, and digital methods. They have been contributing to projects of the *National Research Data Infrastructure (NFDI)* initiative, in particular to the *Memorandum-of-Understanding Group*[6], *QUADRIGA*[7], funded by the Data Literacy Supporting Program of the Federal Ministry of Research and Education, and among others, *Human.Machine.Culture – Artificial Intelligence for the Digital Cultural Heritage*, a project developing and evaluating intelligent document analysis methods[8]. Further, the applicants have connected with international initiatives, namely *Social Networks and Archival Context* cooperative *(SNAC)*[9], University of Virginia, the *Center for Humanities Computing (CHC)*[10], Aarhus University, and *Language Technology and Data Analysis Laboratory*[11] *(LADAL)*, University of Queensland. These initiatives – together with the Letters of Intent received from scholars, research service and data providers (Appendix 2) – exemplify the broader historical data science context and highlight the substantial support SoNAR has already attracted. More general, SoNAR will support:

- *Researchers* who study past events by utilizing HNR methods and theories and are thus in need of historical network data sets. The project will coordinate research services with research service and data providers to enable complementing the knowledge graph.
- *Research service providers* like libraries and archives who support research projects with data gathering services like digitizing and describing archival and related resources which are maintained even by different cultural heritage organizations[12].
- *Data providers* who might reuse historical social network data to complement respective data services like catalogs, portals, digital editions etc. with contextual information about the creation, usage, and maintenance of cultural heritage resources.

SoNAR will bridge historical research and data science to unlock hidden narratives and deepen our understanding of the past through agents and agent relations documented with the cultural heritage. The proposal is based on sound technological and organizational concepts elaborated and adjusted during the last years. As a research data market, SoNAR embraces the vast amount of diverse data to further the application of data science methods in History, (Digital) Humanities, and Social Sciences. SoNAR will be also a case example for examining benefits and conditions of utilizing advanced digital methods and thus contributes to the scientific discourse on methods. The applicants have been reaching out to both the research and the cultural heritage communities and will continue to do so to stand up to expectations and ensure a high user value.

---

[6] NFDI4Culture, NFDI4Memory, NFDI4Objects, Text+

[7] QUADRIGA. https://www.uni-potsdam.de/de/pogs/ueber-uns/projekte [2024-04-05]

[8] Human.Machine.Culture (Mensch.Maschine.Kultur). https://mmk.sbb.berlin/ [2024-04-05]

[9] Social Network and Archival Context. https://snaccooperative.org/ [2024-04-05]

[10] Past Social Network Reconstruction from Material Culture Data. https://chc.au.dk/research/past-social-network-reconstruction-from-material-culture-data [2024-04-05]

[11] Best Practice Tutorials of LADAL, e.g. Network Analysis using R: https://ladal.edu.au/net.html [2024-04-05]

[12] Here exemplified by the project “Cataloguing of the correspondence and biographical documents of Veit Ludwig von Seckendorff (1626-1692)” involving 32 organizations (https://gepris.dfg.de/gepris/projekt/426320490)

2 Objectives and work programme

2.1 Anticipated total duration of the project

The project is scheduled for 24 months. The proposal aims at implementing the results of the pilot project in line with the funding program e-Research Technologies that covers three stages: testing, implementing, and consolidating e-research technologies (DFG form 12.19 – 01/21, 3).

2.2 Objectives

The applicants propose the implementation of an advanced research data market for historical network data to support HNR. SoNAR aims to streamline gathering historical network data sets and subsequent research activities like data analysis and publication by:

1. Establishing a service for integrating dispersed or isolated data sets from cultural heritage and research communities to infer statements about agents and agent relations.
2. Providing interfaces to explore, curate, and download network data sets for a subsequent analysis with suitable data science solutions like R, Jupyter notebooks, or Gephi.
3. Implementing essential research requirements to trace the inferred data to their respective origin and reproduce downloaded network data sets to support good research practice.
4. Supporting researchers with state-of-the-art educational materials, best practice tutorials, and, among others, documentation of API and UI.
5. Encouraging cultural heritage organizations to reuse SoNAR data to provide more detailed contextual information about agents who created and used (archival) resources[13].

Additionally, the proposal suggests continuing work on a concept to implement and operate text-analytic technologies for also integrating data about agents and agent relations from text-digitized resources like newspapers. The result will be a recommendation on an implementation strategy that covers a) technology, quality, and stakeholder analysis, and b) a cost-benefit analysis.

Besides these functional objectives, SBB-PK ensures that the infrastructure (backend, data, and interfaces) is operational as well as secure and reliable. The applicants will make the development process and final results transparent including:

- System architecture, data processing, and components
- Transformation rules and a historical network ontology
- Decision making regarding product and sprint backlogs
- Feedback collected through (online) consultations

To reach out to stakeholders and make the content broadly accessible website, reports, tutorials etc. will be published in English. The website will conform to web content accessibility standards.

2.3 Work program and proposed methods

2.3.1 Project Management

The applicants will give particular attention to user experience and user value. In this regard, the project adapts to lean project management and SCRUM methods. The lean project management follows a classical approach but brings the user value of a product or service into focus by avoiding activities not contributing to this end. The user value is supported best by engaging stakeholders

[13] Beispielhaft für eine Nachnutzung wären Handschriftenportal (https://handschriftenportal.de), Archivportal-D (https://www.archivportal-d.de) oder „Natural Law 16252-1850“ (http://naturallawdatabase.gotha.digital/).

in an open communication with the development team – a strength of SCRUM. The development process is designed as a chain of feedback loops while managing, controlling, and closing of the project will follow a traditional (lean) project management approach.

Work packages will self-organize, and team members act self-responsible to avoid bureaucracy. To facilitate communication, SCRUM-based meetings will be established:

- Regular project team meetings to share information and coordinate in short intervals
- Monthly review meetings of project team and project management to discuss progress, discuss lessons learnt, and decide on both sprint and project related issues
- Joint quarterly reviews of project team, project management, and the advisory board to also discuss methods, feedback conclusions, and product related issues in more detail

The project team will be constituted by project and permanent staff members working on the work packages together. The project management will be representatives of both IBI/HU and SBB-PK. They will be responsible for strategic stakeholder management including reporting and outcome. The project management monitors progress and meets at review meetings or ad-hoc if necessary. The advisory board is constituted by scholars, research service and data providers (Appendix 2).

The applicants will set up standard tools to communicate and publish interim results. The meeting minutes (summaries) will be shared to monitor progress and make decision making transparent. The project team will publish interim results for continuous evaluation after each monthly review.

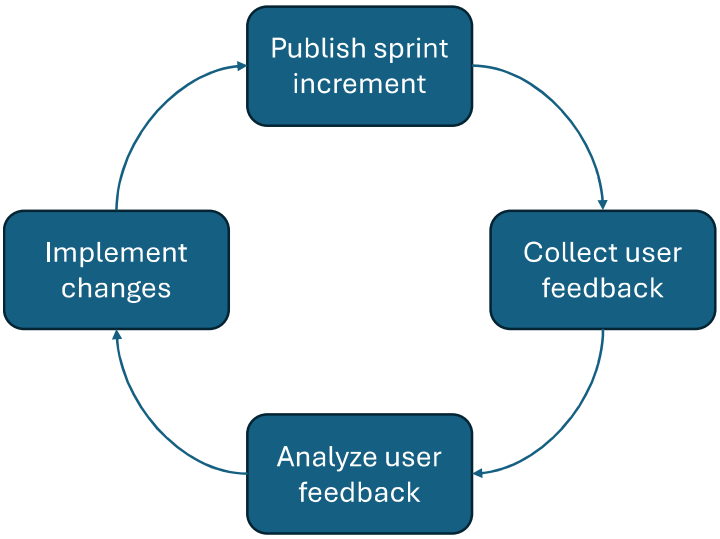


Figure 1: Feedback loop

The project will collect feedback from stakeholders to ensure high user value. To gather feedback, the project applies different approaches like online survey, usability test, or interview. The collected feedback will be analyzed and added to a product backlog. The product owner coordinates product backlog items with the development team for a subsequent sprint. Feedback not registered with the product backlog will be added to a pool of topics for later requirements analysis.

There will be three milestones, which the project team will work on: an alpha release scheduled for the 3rd quarter, a beta release for the 5th quarter, and a final release at the end of the project. Features of the alpha release are outlined in Appendix 5.2. Collected feedback will further adjust the development.

Finally, the project sets up a risk management plan (Appendix 8). The project coordinator monitors risks based on interim reports and coordinates, supported by the project management, measures at an early stage to avoid negative impacts on the overall project outcome.

2.3.2 Work packages and project schedule

The project will implement two work packages (WP): Implementation infrastructure (WP1) and Product and project management (WP2). Both WPs are interrelated by feedback loops. The Gantt outlines three project stages: On-Boarding (1st quarter), development (2nd to 7th quarter), go-live (8th quarter). The development stage will be structured by three product release cycles: Minimum Viable Product (MVP), beta release, and final release. The product release cycles will determine the activities of both WPs. Colored fields relate the activities with the timeline. The efforts, in turn, are enumerated in Appendix 1.3 (justification of personnel).

Iterations (sprints)	Q1	Q2			Q3			Q4			Q5			Q6			Q7			Q8		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
	On-boarding	MVP stage 1			MVP stage 2			Beta stage 1			Beta stage 2			Final stage 1			Final stage 2			Go-live stage		
WP1 Implementing infrastructure and services (Lead: SBB)																						
Preparation stage																						
M1 MVP phase																						
M2 Beta phase																						
M3 Final phase																						
M4 Start-up phase																						
WP2 Product and project management (Lead: IBI/HU)																						
WP2.1 Product evaluation and project coordination (HU)																						
Preparation stage																						
T2.1.1 Expectations and assessment of the advisory board																						
T2.1.2 Short-term feedback on sprint result																						
T2.1.3 Prototype usability tests (MVP, beta release)																						
T2.1.4 Future development roadmap																						
T2.1.5 Project coordination activities																						
WP2.2 Community management (IIE/SBB)																						
Preparation stage																						
T2.2.1 Feedback analysis and product backlog																						
T2.2.2 Social media outreach and website content strategy																						
T2.2.4 Ccommunity engagement in R&D																						
T2.2.5 Product strategy																						
T2.2.6 Go-live of SoNAR																						
WP2.3 Technology evaluation and digital educational materials (IDM/SBB)																						
Preparation Stage																						
T2.3.2 Evaluation of full-text aggregation services																						
T2.3.2 Evaluation of network data enrichment services																						
T2.3.3 Education material consulting and best practice tutorials																						

Figure 2: Key project activities related with a timeline. Due to SCRUM based methods (feedback loops) most tasks are repetitive. Strategies will be conceptualized and refined during the project.

During sprints, a four-week period, WP1 will produce increments (product versions). A MVP will be published within the 3rd quarter. It will include all the components of the final release (Appendix 5.2). Subsequent sprints will complement the features, data, and optimizations. Optimizations are as crucial as both features and data. Thus, product management and development team will give particular attention to the software quality for a sustainable, efficient operation and maintenance during the products' life cycle, in particular security and reliability. Feedback loops with customers (WP2.1) on (interim) results will support decision making of the project team.

<u>WP1 Implementing infrastructure (Lead: IDM/SBB-PK)</u>

WP1.1 Backend Infrastructure, WP1.2 Transformation models and Historical Network Ontology, WP1.3 User Interface

The implementation is divided into three activities: 1) establishing the backend infrastructure, 2) building the intellectual framework of transformation rules and a historical network ontology – an in-depth set of deduction rules to infer the explicit and implicit statements about agents and agent relations from integrated data –, and 3) implementing the web-based machine and user interfaces (API, UI). The interfaces of SoNAR will allow for exploring, curating, and downloading of historical network data, thereby supporting research requirements like specifying a research question, reproducing a network data set, or verifying data by tracing the respective data sources. WP2 will support the development process through collecting and analyzing feedback and developing best practice tutorials. The advisory board will give advice on the interim results through joint quarterly

reviews, a historical network ontology, and workflows to integrate network data sets from research projects. Ontology maintenance will be coordinated with the research community by WP2.2.

WP1.1 Backend infrastructure (IDM/SBB-PK)
The backend infrastructure is divided into two parts: 1) components constituting a highly efficient ETL process (Extract, Transform, Load) to aggregate, integrate, and load data sets 2) into a graph database optimized for inferring explicit and implicit data about agents and agent relations. The backend infrastructure considers key research requirements, in particular maintaining data about the origin of inferred data, ensuring accessibility to aggregated data and transformation rules, and reference integrity. By iteratively collaborating with experts, the implementation aims at ensuring that data provided by SoNAR will be reliable and ready to be used in historical network research.

T1.1.1 Data harvesting
Implement a scalable, configurable data harvesting system. It will fetch data from multiple, diverse data sources (Appendix 3) via standard protocols like OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) and SRU (Search/Retrieve via URL) as well as customized interfaces for SNAC (Social Networks and Archival Context) and Museum Digital. To integrate the historical network data gathered by research projects, SoNAR will be capable of harvesting research data repositories. EMPORION, a research data repository operated by both SBB-PK and the Common Library Network, will serve as a case example: To establish and demonstrate a suitable workflow, a network data set created by the project *Correspondence of Early Romanticism* will be uploaded to EMPORION and integrated by SoNAR (Appendix 2). Based on this case example, SoNAR will share guidelines about data encoding and harvesting protocols with research data repositories and HNR research projects. A metadata manager will be provided with features to configure and manage connected data repositories efficiently. The harvesting methods ensure a steady flow of data to continuously update and extent the data set provided by the research data market SoNAR.

T1.1.2 Archiving and versioning
The harvested data will be archived to allow for reproducing network data sets downloaded from the research data markets' interfaces. Archived data will be preserved for a period of three years. In addition, SoNAR will archive both the transformation rules and the historical network ontology. To enable users to identify the relevant archived versions, SoNAR will automatically generate and provide provenance metadata with downloaded network data sets. Provenance data include links to the archived data sources and versions of the respective ETL configurations. Customers will be able to archive provenance metadata with a network data set in a research data repository.

T1.1.3 Transformation system
Develop a flexible transformation system within SoNAR that allows a metadata manager to define and manage transformation rules using a Domain-Specific Language (DSL, e.g. Metafacture Flux, Catmandu Fix, or Airflow DAG). Such a system will

- support conversions of conventional (XML) data of the cultural heritage communities into RDF, based on the domain-specific models (T1.2.1),
- complement the source data with information from sources like Wikidata, and
- seamlessly merge data from various source systems.

The last-mentioned includes clustering same entities identified by identifiers of different authority files used by the various harvested data sets. The clustering will be based on identifier crosswalks like VIAF or Wikidata. A staging system for reviewing transformed data by the metadata manager will be established to maintain the integrity and quality of data.

T1.1.4 Knowledge graph

Implement a solid, efficient system for loading RDF data into a graph database. This system must efficiently update data, integrate new data, and incorporate comprehensive provenance data. The objective is to maintain an accurate and richly annotated graph database. The knowledge graph utilizes a historical network ontology (T1.2.1) to infer explicit and implicit statements about agents, agent properties and relations encoded with the integrated and uploaded RDF-domain data. The knowledge graph will be made up of millions of agents and respective data sources documenting each inferred statement. The evolving SoNAR knowledge graph will unify access to the otherwise widely scattered, hidden knowledge about agents and thus significantly contribute to HNR.

WP1.2 Transformation models and ontologies (IIE/SBB-PK)

The ETL process (T1.1.3) requires transformation rules encoded in a domain-specific language (DSL). The transformation rules define crosswalks from traditional well-established data formats of the cultural heritage domain to respective RDF models. The inference of the explicit and implicit statements about agents and agent relations from the integrated data requires a comprehensive ontology tailored for historical network research. Such an ontology defines classes of entities, e.g. agents like persons and corporate bodies, relationship types, and inference rules to deduce the statements about agents, agent properties, and agent relations. It enables the metadata manager to tailor data processing to the needs of historical network research. WP1.2 requires an extensive collaboration with experts from the cultural heritage and research communities to test, refine, and verify both the transformation rules and the historical network ontology.

T1.2.1 Transformation models

Develop and encode a set of transformation rules by using a DSL established by T1.1.3 to convert incoming aggregated data into an intermediate graph structure (RDF), e.g.: MARC21 to the RDA-Ontology (bibliographic data), MARC21 to GND-Ontology and Agent Relation Ontology (authority data), EAD and EAC-CPF to Records in Context Ontology, and LIDO to CIDOC CRM Ontology. These intermediate graph structures, still rooted in the domain context of libraries, archives, and museums, form the basis for a unified graph of historical networks. The transformation rules must ensure that the intermediate structures accurately cover complexities and relations encoded with the data of the cultural heritage domain. This includes transformations of the (semantics) of data formats and controlled vocabularies like MARC relator codes[14]. The project aims at reusing well-established transformation rules and sharing adjustments[15].

T1.2.2 Historical network ontology

Elaborate a comprehensive ontology tailored for historical network research that defines entities and relationships, as well as the rules to deduce both explicit and implicit relations about agents and agent relations from the intermediate graph structures (T1.2.1). A historical network ontology reuses vocabularies and inference rules of ontologies like GND-O[16], AgRelOn[17], or RiC-O[18]. A broad coverage of the historical network ontology in research is key to integrate historical network data from the scholarly domain efficiently and for cross-domain data exchange (WP2.2).

WP1.3 Public web interfaces (IDM/SBB-PK)

[14] https://www.loc.gov/marc/relators/relaterm.html [2024-05-04]

[15] This might include transformation rules that come with ETL solutions like Catmandu, Metafacture, or are provided by domain experts, e.g. https://ica-egad.github.io/RiC-O/examples.html [2024-05-04]

[16] https://d-nb.info/standards/elementset/gnd [2024-05-04]

[17] https://d-nb.info/standards/elementset/agrelon [2024-05-04]

[18] https://www.ica.org/resource/records-in-contexts-ontology/ [2024-05-04]

The research data market SoNAR focuses on easy to use means to access and interact with the historical network data. These data will be disseminated through a web application programming interface (API) that will support various data formats. The API will be used equally by a web-based graphical user interface (UI) and the various data science solutions like R, Jupyter Notebooks, or Gephi. The API allows for searching, curating, and downloading historical network data sets. The UI provides means to select and visualize data subsets for exploring and clustering data based on agents and relation properties. An accessible and user-friendly web page will provide extensive documentation about data, functionality, and best practice tutorials. A coherent design across the research data market will ensure consistent visual identity, enhance usability, and encourage user engagement with the interfaces. The public web interface will be accessible without registration.

T1.3.1 Web application programming interface
Develop a public facing Web API that interfaces with the graph database. The Web API enables both UI and data science solutions like R to query and retrieve data. The API must support multiple data formats (JSON, CSV, GraphML) and provide the provenance data containing at least links to the source data and transformation rules. It will incorporate a graph query language such as CYPHER for Neo4j or SPARQL for RDF stores, thereby catering to diverse research needs.

T1.3.2 Data selection interface
Develop a user-centric, web-based GUI for the research data market that is designed primarily to support users with exploring, curating, and downloading data from the graph database with ease and precision. This involves integrating advanced data exploration features, e.g. search, cluster, facet, and filter options, tailored for historical network data. The GUI must also facilitate in-depth views of selected entities and relationships, including their historical contexts and provenance.

T1.3.3 Graph visualization interface
Develop and integrate visualization capabilities with the data selection interface that must allow for visualizing network data along the demographic, temporal, or subject properties to identify and display clusters of a social network. These visualization tools should offer users a comprehensive view of the data's historical context and relationships, facilitating deeper insights into the dynamics of historical networks, e.g. development of a persons' network over time or cohort-based contexts.

T1.3.4 Data download
Implement functionalities that allow users to effortlessly download selected historical network data in a data format suitable for data science solutions like R. This should enhance the research data markets' interoperability and extend its utility by enabling users to analyze data with tools suitable for a particular research purpose.

T1.3.5 Web page
The web page welcomes and guides users. It encourages exploration and usage in research. It introduces services, functionality, and data. Best practice tutorials provide users with use cases and demonstrate the capabilities of SoNAR. It will provide also extensive, detailed, and accessible documentation for the interfaces and in-depth information about the data sources, latest updates, and the frequency of data harvesting. T1.3.5 provides a content management, design framework for publishing and maintaining the content. The web page content will be provided by T2.2.2.

T1.3.6 Corporate design
The development and implementation of a cohesive, distinctive corporate design of the research data market UI will reflect the academic and research context of SoNAR. The design facilitates a seamless, intuitive user experience and promotes both ease of use and a strong visual identity.

WP2 Product and project management (Lead: IBI/HU)

WP2.1 Product evaluation, WP2.2 Stakeholder management, WP2.3 Data extension and usage

The project aims at maximizing the user experience and user value of the research data market. WP2 will establish a product and stakeholder management that provides for a viable, sustainable solution to meet users' needs and to support development and maintenance over the life cycle of SoNAR. Product and stakeholder management makes sure a tangible user value will be delivered by connecting with customers for respective feedback. It makes certain that accurate requirement descriptions and related context information will be shared with the development team. In addition, product and stakeholder management ensures that a feasible, user-centered product strategy will be available at the end of the project. WP2 will also coordinate the transition from the development team to the support and maintenance team. An advisory board that is constituted by researchers, research service and data providers (Appendix 2) will give advice on HNR methods, requirements, and building a sustainable community of customers and both research service and data providers. Finally, WP2 continues the evaluation of latest text-analytic technologies for a recognition, linking, and integration of structured data about agents from text-digitized resources like newspapers.

WP2.1 Product evaluation (IBI/HU)

The project requires active and reciprocal relationships with the intended user communities of the SoNAR infrastructure: a) researchers from several disciplines with a broad variety of use cases and research scenarios utilizing HNR methods and b) both research service and data providers facilitating data gathering, acquisition of data sets, or data analysis solutions for research projects. WP2.1 will leverage the advisory board members, who are involved in respective communities of users, as a first proxy to the user population and to maintain feedback loops for evaluating user value of design, features, and data. Online surveys will allow for reaching out for feedback from communities that will be recommended by the advisory board members or identified by T2.2.2 via Social Media outreach. The user interface design, available with the first two milestones, will be examined in a usability lab. Results from surveys and usability tests will be shared with the product owner for analysis and coordination with the development team. Furthermore, WP2.1 will provide a future development roadmap deduced from product backlog, pool of topics, and evaluations.

T2.1.1 Expectations and assessment of the advisory board

The advisory board will meet as a group three times to provide structured and concerted feedback about the projects' progress at the beginning, in the middle, and at the end of the project. Members will be interviewed at the beginning of the project to build a set of user expectations that will be translated into product and infrastructure development steps. The interviews will also serve as a mechanism, whereby case studies of HNR research scenarios will be collected during the project. The case studies will be used as template for envisioning user interactions with the product. The advisory board will be apprised of updates and invited to joint quarterly reviews. Members will be asked to test prototypes, especially public interfaces.

The results will be shared with WP2.2 for feedback analysis and product backlog maintenance.

T2.1.2 Online consultation for short-term feedback on increments

This task constantly connects the development team with the user communities. Each iteration of WP1 will provide sprint outcomes (increments), e.g. UI mockups, prototypes, updates. WP2.1 will set up online surveys to regularly reach out for timely feedback. Online surveys will be modified depending on an increment and the required feedback. The consultations will also serve as a communication tool to stay in contact with the user communities and update them on the progress of the project. The expectation is to build up a user community before the product will go-live.

The results will be shared with WP2.2 for feedback analysis and product backlog maintenance.

T2.1.3 Prototype usability tests
The prototype usability tests will be applied to the first two milestones of the project. Envisioned are evaluations of the Minimum Viable Product and the beta release. The testing population will be recruited with support of T2.2.2, by reaching out to participants of online surveys, and with the advice given by the advisory board. The project will evaluate usability and user experience. The tests will utilize the Usability-Lab of IBI/HU: researchers will test interaction features via simulated tasks based on the case studies collected from advisory board members (T2.1.1). Prototype user experience tests will, in some cases, include interviews about use cases, topics, the context of research, and reservations regarding SoNAR or even HNR methods and theories.

The results will be shared with WP2.2 for feedback analysis and product backlog maintenance.

T2.1.4 Future development roadmap
This task will compile expectations for the future product development, e.g. the integration of new data sets or interface features. The future development roadmap will be based on an analysis of:

- assessments of the advisory board members
- product and sprint backlogs, pool of topics (WP2.2)
- online consultations and prototype usability tests
- workshops and the HNR conference (WP2.4)

The roadmap will be coordinated with the advisory board and the product owner (WP2.2).

T2.1.5 Project coordination
The overall project management strategy is described at the beginning of this chapter. The project coordinator will organize and moderate project meetings and reviews and will coordinate project documentation and the sprint reviews. A regular exchange with all members of the project team is essential and will be supported by internal communication and documentation tools.

WP2.2 Stakeholder management (IIE/SBB-PK)
WP2.2 takes responsibility for stakeholder management. It coordinates the product development and takes measures to build a lasting community of scholars and both research service and data providers in collaboration with WP2.1. Activities include amongst others the planning, testing, and implementing outreach strategies as well as training concepts for a) a dissemination of outcomes, b) the community cooperation, and c) educational materials to support subsequent operation and user guidance. The Product Owner (PO) ensures high user value through feedback analysis and requirements implementation coordination with the development team. The PO will arrange the transition from implementation to operation. It is an objective to build strong ties with international stakeholders, particularly scholars and research service and data providers to encourage the use of SoNAR and sustainable community support. Moreover, WP2.2 will conceptualize and organize both an international conference on historical network research and two workshops on 1) artificial intelligence and semantic web knowledge bases, and 2) research library services supporting data acquisition and assist research with complementing and using the SoNAR research data market.

T2.2.1 Feedback analysis and product backlog
The Product Owner (PO) sets up a product backlog and maintains user stories based on analysis and prioritization of feedback that will be either considered with the backlog or added to a pool of topics. The PO coordinates sprint backlogs with the development team. The initial sprint backlogs aim at a MVP pre-defined by a set of user stories (Appendix 5.2). The components will cover the

whole data processing chain including public interfaces. Subsequent sprints will add functionality, data, and enhance the quality of the research data market. The prioritization will be supported by the advisory board, whose members will be interviewed at the beginning of the project and share feedback at joint quarterly reviews (T2.1.1).

T2.2.2 Social media outreach and website content strategy
To reach out to the communities and users that might benefit from SoNAR, a professional Social Media Outreach (SMO) strategy will be drafted and implemented. The drafted strategy will include standard measures to track SMO activities through data analytics to adjust the strategy. The SMO strategy will consider the advice given by advisory board members about respective communities, and contributes to raising awareness about SoNAR, HNR, and related topics. The implementation of SMO will be tested by student assistants. The experience from the project will be consolidated in a SMO guide to assist the support and maintenance team that will take over after the project. SMO also supports WP2.1 by reaching out to communities for online consultation (T2.1.2).

SBB-PK sets up a project website to inform about progress, findings, workshops, and conferences of the project. The website will support the SMO strategy, and it will be optimized for indexing by search engines including on- and off-site measures. The PO will develop a website content concept for both project and product. The product website will be developed by WP1.3 and should inform about the integrated data repositories, related standards, partnerships, and among others, educational materials like best practice tutorials (T2.3.3). The content strategy considers content maintenance by the support and maintenance team and will be accordingly coordinated with SBB-PK and the advisory board members. The website content creation will be supported by student assistants and the project team members. The project website will be taken offline two years after the end of the project. The product website will continue to inform about the project.

T2.2.3 Community cooperation
The applicants have been connecting with scholars, projects, and initiatives to discuss respective support for the implementation project (Appendices 2 and 3). A strong and dedicated community support is a prerequisite for the SBB-PK to develop and operate SoNAR. Besides the stakeholder involvement through feedback loops (T2.1.1) and a digital outreach strategy (T2.2.2), SoNAR will contribute to the discourse on digital methods. To do so, the project will take three approaches:

1. *SoNAR as a case study:* SBB-PK concluded agreements with the projects QUADRIGA, HERMES, and NFDI4Memory (Appendix 2) to work with public interfaces of SoNAR, e.g. to elaborate advanced educational materials, to examine methodological (pre-) conditions and related research requirements regarding data and interfaces.
2. *SoNAR in research and teaching:* Advisory board members will impart the research data market in respective lectures and seminars. The PO will provide supporting materials and presentations to introduce SoNAR and to encourage students to test and share feedback on the public interfaces and the research data markets' underlying approaches.
3. *Community research workshops*: The PO will plan and organize two public workshops on 1) artificial intelligence and semantic web-based knowledge graphs and 2) establishing of a supporting international community of scholars and research service and data providers.

The focus of the first workshop is on the expectations and the advice of the research community regarding the integration of agents identified in text-digitized resources into the knowledge graph, in particular confidence and semantics (T2.3.1, T2.3.2). The first workshop will also discuss, more generally, the training of subject-related language models with knowledge graphs.

The second workshop will discuss the contribution of research service and data provider to assist research projects with using SoNAR, to complement data not yet available for a research project, and among others, to curate, analyze, and publish historical network data sets.

Both workshops will be open to a broad audience and conceptualized and coordinated by the PO.

The PO will organize – supported by WP2.1, WP2.3, advisory board members, and *The Historical Network Research Community* – an international conference on SoNAR and HNR.

T2.2.4 Product strategy
The product strategy lays the foundation for SoNAR after the implementation by specifying future objectives including the product and service profile. It will be a basis for the marketing, community cooperation, research and development projects as well as a strategic planning at SBB-PK. The PO will commence work on the product strategy at an early stage for its development is a process that requires experience from the project and continuous business analytics:

1. based on an analysis of deliverables (Appendix 6) and interim results and the
2. coordination with stakeholders involved in the strategic planning and research.

The product strategy development will be supported by T2.1.4, T2.3.1, and T2.3.3, and it will be explicitly considered as key outcome of the project along with the technical implementation (WP1).

T2.2.5 Go-live coordination
The PO will coordinate the transition of SoNAR from the project, especially the development team, to the support and maintenance team. The transition involves permanent staff members (ch. 5.2). Personnel that will be responsible for support and maintenance of SoNAR will be provided with documentation and handouts to solve technical issues and provide advice to users (Appendix 5).

WP2.3 Educational materials and technology evaluation (IDM/SBB-PK)
Technologies such as artificial intelligence (AI), machine learning (ML), and related expectations of scientific communities change rapidly. Therefore, the applicants propose to continue evaluating technologies that significantly contribute to scale the gathering of network data by also processing text-digitized resources (OCR/HTR) like newspapers, manuscripts, and letters. It covers methods of a) entity recognition and linking and b) knowledge graph enrichment through recognizing and qualifying the type of a relation between agents identified with text-digitized resources. The focus is on 1) examining and assessing respective technologies that benefit the user requirements best, 2) the most suitable approaches to apply said technologies for SoNAR, and 3) related costs and efforts required to operate them. Furthermore, WP2.3 will elaborate open educational materials that document and showcase the potential use of SoNAR to prospective users and provides them with clear and practical educational resources on how to use SoNAR effectively.

T2.3.1 Evaluation of full text aggregation services
To expand the data provided through SoNAR, this task will evaluate technologies to extract the relevant information from full text resources obtained by OCR, HTR, or transcriptions (scholarly digital editions). It will leverage state-of-the-art methods from natural language processing such as Named Entity Recognition (NER) and Entity Linking (NEL) based on AI/ML that have been developed by SBB in prior projects[19] (Labusch 2024). Through the identification and classification of named entities and their disambiguation and linking to authority files such as Wikidata or GND, the scope of data offered by SoNAR can not only be significantly enlarged, but full texts can also provide additional contextual information for entities already included in SoNAR, thereby

[19] OCR-D, https://ocr-d.de/en/ [2024-04-05]; MMK, https://mmk.sbb.berlin/ [2024-04-05]

improving their information and analysis value. The potential contribution of these technologies will be evaluated with regard to quality and performance and carefully assessed against requirements and cost for their future integration and productive use within SoNAR.

T2.3.2 Evaluation of network data enrichment services

Various AI/ML technologies have been considerably improved in recent years and can contribute enormous added value for researchers using SoNAR. Relationship extraction can qualify and expand information about the relations between entities (e.g., "*John* married *Catherine* in *Rome*"). Coreference resolution (CR) can help uncover relations previously not discoverable (e.g., "After he married his wife, they moved to Rome."). Furthermore, approaches such as Graph Neural Networks and Retrieval Augmented Generation offer great potential for advanced exploration and novel analysis methods of the network data in SoNAR. These technologies will be evaluated with regard to quality and performance and carefully assessed against the requirements and cost for their future integration and productive use within SoNAR.

T2.3.3 Open educational materials

SoNAR aims at furthering innovative ways to do research with data based on the cultural heritage resources. Therefore, it is of the utmost importance to accompany outreach with comprehensive documentation and best practice tutorials to guide users. The availability of educational materials and case studies for varying skill levels will have an impact on usage and development of research projects based on SoNAR. In turn, this requires a close collaboration with the research community as it is already coordinated with initiatives like QUADRIGA and NFDI4Memory (Appendix 2). This task will establish high-quality resources to ensure that users will be enabled to quickly and easily make use of the full potential of SoNAR, such as

- Comprehensive documentation and examples: detailed API documentation, usage guide, example queries, best practices with data science solutions like R or Jupyter Notebooks. The design must be understandable by both developers and scholars.
- Guidelines for external tool usage: Provide users with documentation and guidelines on how to use downloaded data with popular tools like Gephi including advice on data import, visualization, and analysis to enhance the usability of the downloaded data.
- Training and outreach: Conduct training sessions and workshops with potential API users to demonstrate effective API use for data access and integration into research workflows. Develop outreach materials to promote the API's capabilities for the research community.

This task will closely coordinate with WP2.1 and WP2.2 to include user expectations and evaluate educational materials to improve satisfaction of scholars with different data analysis skills.

3 Project- and subject-related list of publications

Bludau, Mark-Jan et al: SoNAR (IDH). Datenschnittstellen für historische Netzwerkanalyse. In: Schöch, Christian (Hg.): DHd2020. Spielräume. Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts. 7. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. Paderborn. (2020). 360-362

Bludau, Mark-Jan / Dörk, Marian / Tominski, Christian: Unfolding Edges for Exploring Multivariate Edge Attributes in Graphs. In: Byška, Jan / Jänicke, Stefan / Schmidt, Johanna (Hg.): EuroVis 2021. Posters. The Eurographics Association. Zürich. https://doi.org/10.2312/evp.20211070 (2021)

Bludau, Mark-Jan / Dörk, Marian / Tominski, Christian: Unfolding Edges. Adding Context to Edges in Multivariate Graph Visualization. In: Computer Graphics forum. https://doi.org/10.1111/cgf.14831 (42, 2023), 3. 297-309

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

Labusch, Kai / Schneider, Sophie / Neudecker, Clemens: Automatisierte semantische Anreicherung von historischen Texten. Erkennung und Verknüpfung von Entitäten mit Wikidata und Wikipedia. In: B-I-T online. (27, 2024), 3 (https://b-i-t-online.de/heft/2024-03-fachbeitrag-labusch.pdf), 232-241

Menzel, Sina et al.: Named Entity Linking mit Wikidata und GND. Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten. In: Franke-Maier, Michael (Hg.): Qualität der Inhaltserschließung. https://doi.org/10.1515/9783110691597-012, 2021

Müller (a), Gerhard: SoNAR – Metadata and networks. Workshop at the KOOP-LITERA 2022. 9th symposium of the German Literary Archives and related institutions. https://www.dnb.de/EN/Kulturell/Veranstaltungsrueckblick/20220622KoopLitera.html [2024-04-05]

Müller (b), Gerhard / Schmid, Larissa / Ostrowski, Felix: Von Entitäten-Relationen zu Forschungsdaten für die Historische Netzwerkforschung. In: Becker, Irmgard Christa et al. (Hg.): Nutzung 3.0. Zwischen Hermeneutik und Technologie? Beiträge zum 25. Archivwissenschaftlichen Kolloquium der Archivschule Marburg. 2022

Petz, Cindarella: On Combining Network Research and Computational Methods on Historical Research Questions and its Implications for the Digital Humanities. Munich, 2022

Putnings, Markus: Datenmarkt. In: Putnings, Markus / Neuroth, Heike / Neumann, Janna (Hg.): Praxishandbuch Forschungsdatenmanagement. Berlin, 2021. 141-145

Stegbauer, Christian / Häußling, Roger: Handbuch Netzwerkforschung. https://doi.org/10.1007/978-3-658-37507-2, 2024

Appendix

1 Budget table and commentary

1.1 Budget overview

2 Advisory Board

The advisory board will give advice and support to the project. The board will be constituted by representatives of the research community as well as research service and data providers. They will contribute at least to one of the following three aspects depending on respective backgrounds:

- Provide feedback on interfaces developed to explore, to curate, and to download historical network data sets for subsequent data analysis, e.g., with R or Jupyter notebooks.
- Exchange of experience about educational materials and research services to encourage the use of digital methods in History, (Digital) Humanities, and Social Sciences.
- Support dissemination and quality management, e.g., by recommending interviewees and channels for distributing surveys or by imparting SoNAR in lectures and workshops.

Advisory board members will contribute to an international Historical Network Conference hosted by the applicants to promote SoNAR and digital methods in History, (Digital) Humanities, and the Social Sciences. To monitor the progress of SoNAR, the advisory board will be invited to the joint quarterly reviews (ch. 2.3.1) to share thoughts on deliverables, lessons learned, and feedback.

Data providers will give advice about the data repositories (Appendix 3) (interfaces, data formats, controlled vocabularies, data about agents and agent relations) and contribute to the discussions about approaches to furthers enhance the SoNAR user value for the research and cultural heritage communities, e.g., by reusing data describing social contexts of cultural heritage resources.

The constitution of the advisory board is based on the need 1) for an international cooperation to access data, 2) for evaluating the user value of interfaces for a broad set of research cases, and 3) for advice on educational material, best practice tutorials, and research services.

List of Advisory Board members:

Name	LoI ID	Organization	Projects and Contributions
Dr. Bettina Buchholz	A01	University of Potsdam. Potsdam Graduate School	Project: QUADRIGA. Center of Data Literacy Competence (BMBF) - SoNAR as a case example for the project QUADRIGA (development of educational materials) - Support dissemination
Prof. Dr. Marten Düring	A02	University of Luxembourg, Luxembourg Centre for Contemporary and Digital History (C2DH)	Historical Network Research Community, Projects impresso I / II (newspapers) - Provide feedback also from students / seminar - Exchange experiences - Support dissemination
Prof. Dr. Martin Schweinberger	A03	University of Queensland	Project: LADAL. Language Technology and Data Analysis Laboratory (including Historical Network Data) - Exchange experiences on educational materials - Discuss sharing data among research data market - Support dissemination
Dr. Cinderella Petz	A04	Leibniz Institute of European History	Historical Network Research Community, Modern and Contemporary History - Provide feedback on SoNAR interfaces based on guiding research questions - Exchange experience - Support dissemination
Prof. Dr. Aline Deicke	A05	Philipps University of Marburg. Marburg Center for Digital Culture and Infrastructure (MCDCI)	Historical Network Research Community, Project: HERMES. Center of Data Literacy Competence (BMBF), Correspondences of Early Romanticism (DFG)

Name	LoI ID	Organization	Projects and Contributions
			- Support best practice workflows for integrating data from research projects (Correspondence of Early Romanticism) - Support dissemination - SoNAR as a case example for the project HERMES (development of educational materials)
Dr. Heimo Stiemer	A06	Technical University of Darmstadt. Institute of Linguistics and Literary Studies	Literary Studies, in particular periodicals and story analysis with machine learning methods and Historical Network Research - Provide feedback on SoNAR interfaces based on guiding research questions - Exchange experiences - Support dissemination
Dr. Iza Romanowska	A07	Aarhus University. School of Culture and Society. Center for Humanities Computing	Historical Network Research Community, Archaeology - Provide feedback on SoNAR interfaces based on guiding research questions - Exchange experiences, especially on educational materials and application of advanced technologies - Support dissemination
Prof. Dr. Heiner Fangerau	A08	Heinrich-Heine University (Düsseldorf). Department of the History, Theory and Ethics of Medicine	Historical Network Research Community, Medical History, Project: Correspondence in “Vester’s Archives” (Pharmacy) - Provide feedback on SoNAR interfaces based on guiding research questions - Exchange experiences - Support dissemination
Prof. Dr. Torsten Hiltmann	A09	Humboldt-University of Berlin, Digital History	Project NFDI4Memory, Lead TA5 (Data Culture) - Exchange experiences - Support dissemination - In-depths analysis of impact of HNR and SoNAR on research (data analysis pre-conditions, NFDI4Memory)
Prof. Dr. Frederick Elwert	A10	Ruhr University Bochum. Collaborative Research Center. SFB1475	Historical Network Research Community, Religious Studies, Project: SeNeReKo - Provide Feedback on SoNAR interfaces based on guiding research questions - Exchange experiences - Support dissemination
Florence Clavaud	A11	Archives Nationales (Frances). Lab	Executive Member of ICA EGAD (Records in Context – Ontology), TS-EAS of the Society of American Archivists - Exchange experiences (as research service provider) - Support dissemination - Give advice on ontologies, especially “Records in Context”
Dr. Hendrikje Carius (Deputy Director)	A12	Forschungsbibliothek Gotha. Benutzung und Digitale Bibliothek	Europäische Gelehrtennetzwerke (EMLO, Nodegoat) - Exchange experiences (as research service provider) - Support dissemination - Give advice on best practices for imparting digital methods by medium-sized research service provider
Joshua Ramon Enslin	D01	Freies Deutsches Hochstift. Digital Collection Management	Museum Digital (MD), Senior Manager and Data Scientist - Exchange experiences (as data provider) - Support dissemination - Give advice on data, data exchange, and ontologies

Name	LoI ID	Organization	Projects and Contributions
Frank Scholze (Director General)	D02	German National Library. Integrated Authority File	Integrated Authority File (GND) - Exchange experiences (as data provider) - Support dissemination - Give advice on data, data exchange, and ontologies
Prof. John R. Hott (recommended by Daniel Pitti, SNAC Director)	D03	Assistant Professor in the University of Virginia. Computer Science Department	(former lead developer of SNAC), focus on graph technologies and data modelling - Exchange experiences (as data provider) - Support dissemination - Give advice on data, data exchange, and ontologies

3 Data Providers and Repositories

The initial implementation will integrate particularly well-established data repositories from cultural heritage communities. Selection criteria are:

- Size of data sets,
- Representation of library, archives, and museum communities,
- Metadata and authority records.

Repositories must meet few quality criteria: persistent URI identifying entities and compliance with domain data standards. Data repositories, which had been processed during the pilot project SoNAR (IDH), were given priority. To foster an international orientation, the collaboration with the Social Network and Archival Context Cooperative, which was already agreed upon during SoNAR (IDH), has been taken up again.

SoNAR will also integrate historical network data gathered by researchers. The workflow will be exemplified by a data set created by the project "Correspondence of Early German Romantism" (Appendix 2). The data set will be uploaded to the research data repository "EMPORION"[20] and harvested by SoNAR (Appendix 4). EMPORION is operated by SBB-PK. Alternative research data repositories will be connected to SoNAR on request and based on research data repository standards.

The SoNAR infrastructure will make use of established SameAs-Services of Lobid.Org, VIAF, and Wikidata to cluster entities that are identified by different authority control numbers of various authority files. Identified entities will be also enriched by respective data available at Wikidata (Appendix 4).

The table below names repositories whose data will be integrated during the initial implementation project. It includes brief descriptions of the type of data, size of the repository (number of records, file size, estimated annual average growth rate), and both data format and interfaces available to harvest the repositories.

SoNAR will aggregate approx. 55.000.000 records from cultural heritage communities (not considered the data harvested from research data repositories), an estimated total size of 107 GB. The average annual growth rate of the data repositories is estimated to be 7.16 percent (as of April 2024).

List of Data Providers:

Data repository and provider	ID	Type of data	Size	Data Access
Kalliope Union Catalog (KPE), SBB-PK (https://kalliope-verbund.info)	01	Archival metadata (personal papers, manuscript collections, publisher's archives)	6.000.000 records, 12 GB, average growth p.a.: 10 percent	EAD, Solr API
Integrated Authority File (GND) cooperative, German National Library (https://gnd.network)	02	Authority records (persons, corporate bodies, subjects, geographical names, work titles, corporate bodies, conferences)	9.700.000 records, 37 GB, average growth p.a.: 5 percent	MARC21, OAI
German National Library (DNB), Catalog (https://katalog.dnb.de/)	03	Bibliographic metadata (printed books, e-resources, maps, music)	31.000.000 records, 100 GB, growth p.a.: 10 percent	MARC21, OAI
Union Catalog of Serials (ZDB), SBB-PK, German National Library (https://zeitschriftendatenbank.de)	04	Bibliographic metadata (periodicals like scientific journals, newspapers)	2.120.000 records, 11 GB, average growth p.a.: 4 percent	MARC21, OAI
Museum Digital (MD), Community (https://www.museum-digital.de/)	05	Metadata cover museum objects like paintings, films, photographs, three-dimensional (vases, cars etc.)	1.050.000 records, 5 GB, average growth p.a.: 12 percent	LIDO, MD API
Social Network and Archival Context (SNAC), Cooperative (https://snaccooperative.org/)	06	Authority records describing 884352 corporate bodies, 2454641 persons, and 195561 families (CPF) and over 2106332 related archival records	5.640.000 records, 25 GB, average growth p.a.: 2 percent	EAC-CPF, SNAC API

[20] EMPORION. Forschungdaten-Hub für Sozial- & Wirtschaftsgeschichte. https://emporion.gswg.info [2024-04-05]

4 Key results from the pilot project

The results of the pilot project (2019-2021) are published on GitHub https://github.com/sonar-idh – The following enumerated results are considered by the work program of the proposal:

Concept on implementing and operating SoNAR (2021, in German only): https://github.com/sonar-idh/reports/blob/main/AP1-SBB-1-Implementierungskonzept-short.pdf

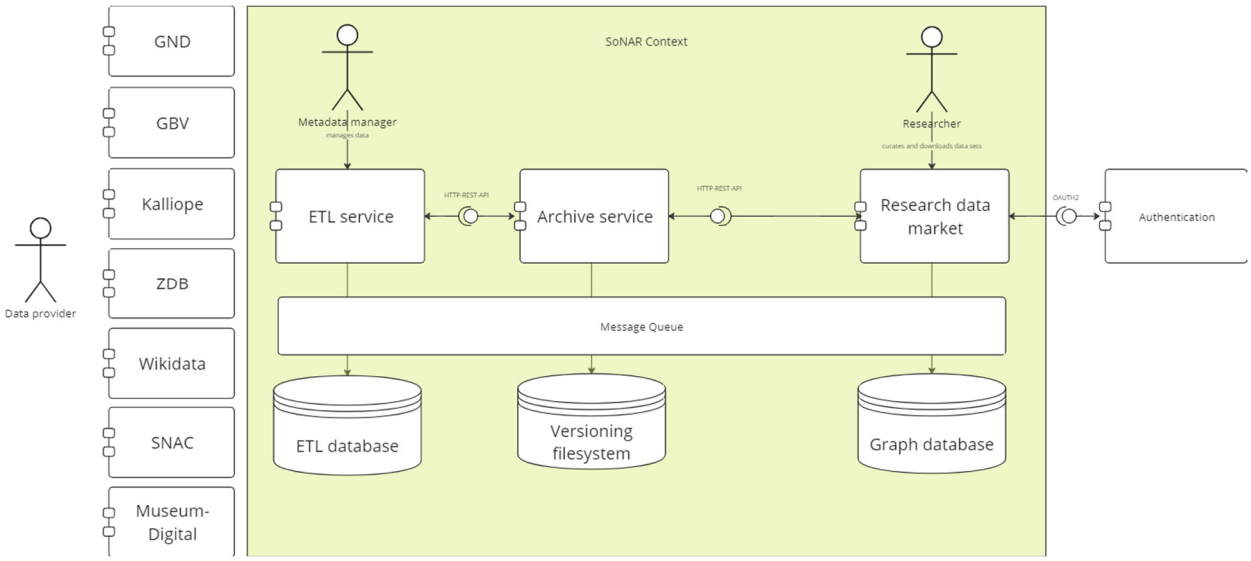
- Appendix 1: Status-quo analysis of historical network research https://github.com/sonar-idh/reports/blob/main/AP1-SBB-2-Umfeldanalyse.pdf
- Appendix 2: Use cases and systems description https://github.com/sonar-idh/reports/blob/main/AP1-SBB-3-Systembeschreibung.xlsx
- Appendix 3: Domain ontologies analysis: https://github.com/sonar-idh/reports/blob/main/AP1-SBB-4-Datenmodellskizze.pdf

Additional supporting results considered with the proposal for implementation:

- Exemplary research design (2021, in German only):: (https://github.com/sonar-idh/reports/blob/main/AP2-UDK-Projektdokumentation.pdf)
  - Chapter 1: Exemplary research design
  - Chapter 2: User requirements regarding visualization and user interface design
  - Chapter 3: Report on a Historical Network Research workshop 02/2021
- Visualization and interface design concepts
- Jupyter-Curriculum on Historical Network Research (HNR) https://github.com/sonar-idh/jupyter-curriculum
- Scientific report on visualizing historical networks and multimodal relationships: https://github.com/sonar-idh/reports/blob/main/AP3-FHP-Projektdokumentation.pdf
- Documentation of preliminary prototype visualization concepts: https://github.com/sonar-idh/visualization-prototypes
- Evaluation reports
  - Data quality: (https://github.com/sonar-idh/reports/blob/main/AP4-HU-4-2-1-Bericht-Evaluierung-I.pdf
  - Exemplary research design (https://github.com/sonar-idh/reports/blob/main/AP4-HU-4-4-2-Evaluierung-III.pdf)
  - Report on an initial “walk through”-Workshop 2019: https://github.com/sonar-idh/reports/blob/main/AP4-HU-4-5-1-Auswertungsbericht-Visualisierungsworkshop.pdf
  - User interview analysis (https://github.com/sonar-idh/reports/blob/main/AP4-HU-4-5-3-Evaluierung-IV.pdf)

5 Systems architecture and commentary

5.1 Key components, workflows, and data repositories



All products considered for SoNAR components will be reevaluated during an on-boarding project phase. The products mentioned below are snapshots and might be subject of change due to the rapid technological advancements anticipated by the applicants to take probably place between the proposal submission and a start of the project:

- Archive service: The archiving service provides efficient data storage and versioning. An implementation candidate is LakeFS (https://lakefs.io/). LakeFS is designed to manage large datasets like Git, and it is providing robust versioning control enabling the SoNAR service to store data modifications over time, track historical versions of datasets, and ensure that every version is accessible and citable to meet a key research requirement for data integrity and reproducibility.
- ETL service: The ETL service must be highly adaptable, expandable, and suitable for a broad set of data transformations. There are at least three candidates for implementation in form of well-established frameworks like Metafacture, Catmandu or Apache Airflow, all offering a robust set of tools for processing structured data. The tools will be evaluated during the preparation phase. Performance, cost-efficiency, and sustainability (provided by a strong open-source community) will be major criteria for choosing a particular approach. Only in case the established tools will not meet the requirements, a custom approach will be implemented using a plugin-based architecture, allowing for tailored functionalities and dynamic adaptation to evolving requirements.
- Message queue: To ensure reliable data communication between ETL processes and web presentation layer of the SoNAR platform, a message queue based on Apache Kafka will be implemented. Apache Kafka is an established service at SBB-PK. It is well-regarded for high throughput, built-in partitioning, replication, and fault tolerance. It is an excellent solution for managing data flows within complex systems like SoNAR.
- Graph database: The project will give either Neo4J or Blazegraph priority. Neo4J is a popular graph database known for its high performance, extensive query capabilities, and strong community support. Blazegraph, an RDF triple store, is known for the ability to deal even with very large RDF data sets, and it supports SPARQL query. Both databases are robust solutions for complex, graph data structures. A decision depends on technological state-of-the-art at the time of the project.
- Web UI: The web-based UI of SoNAR requires a high-performance framework for graph visualization, such as Cosmos (https://github.com/cosmograph-org/cosmos). Additionally, the UI development will use standard technologies like React.js for building UI

components and Node.js for server-side operations. Both ensure that SoNAR manages and displays complex networks efficiently and delivers a responsive user experience.

- Web API: The API will be designed primarily as a REST-based wrapper that provides standardized access to the graph database and ensures a streamlined, intuitive data interaction. In addition, the API will expose the graph query language to users and thus will allow for complex, precise data manipulations and queries. This dual approach enables both ease of use for common tasks and advanced capabilities for sophisticated data analysis. It makes the API a versatile tool for working with historical network data.

5.2 Narrative for a Minimum Viable Product (MVP)

The story below describes the functionalities and data that will be available with a first MVP nine months after the proposed project commenced its work. Both functionalities and data repositories will be extended and consolidated with subsequent iterations (sprints) during the project:

Max Mustermann studies History. He is working on a topic for his master thesis. For some time he has been familiarizing himself with the "Frankfurt School" (Frankfurter Schule), and therefore he enters "Theodor W. Adorno" into the search field of SoNAR, a service he learnt about during a seminar on „Digital Methods in History" only recently. After clicking the search button, he selects "Theodor W. Adorno (1903-1969)" from a hit list. Max sees detailed information about Adorno that are gathered from multiple data repositories and clicks on "historical social network". SoNAR now displays an agent-centered historical network graph of Adorno who relates to hundreds of first-degree agents by multiple types of edges (visualized property graph). The graph provokes Max's curiosity, and he clicks on different nodes, each representing a uniquely identified agent. He determines, unsurprisingly, that most agents are also related with the "Frankfurt School" – a dense social network of intergenerational scholars. A bit at the sideline, Max recognizes the name of "Roberto Mangabeira Unger", a person he has not yet heard about in the context of the "Frankfurt School". Max Mustermann now continues exploring Unger by expanding the network of agents based on the social relations of Unger. He discovers that the networks of Adorno and Unger have a clear overlap but are detached from each other. On the SoNAR website, Max learnt that SoNAR is still an MVP with data from the Integrated Authority File of the German National Library indexed only. However, the data serves Max's purpose, and he decides to dig deeper into the data to search for betweenness-centrality. He downloads the data set he has curated, choosing the GraphML. After that, Max continues his work with Gephi, a standard open-source software solution for network analysis. While analyzing the data set, Max discovers even more key persons attributed to "Critical Legal Theory". He then concludes to write his master thesis about the origin of „Critical Legal Theory". During his work on this topic, he creates multiple network visualizations with Gephi – based on a then already expanded data set provided by the research data market SoNAR. When finalizing his work, Max uploads his data set to EMPORION, a research data repository provided by SBB-PK for Social and Economic History that provides a Digital Object Identifier (DOI) which he uses to reference his data. While reviewing the thesis, the supervisor wants to know more about the data used by Max Mustermann for visualizing the networks and the network statistics. She clicks the DOI and gets a description of the data set from EMPORION along with the URL to download the data set. She proceeds to download the data set which contains provenance data referencing the source data and transformation routines. For testing purposes, the professor looks for the archived data sources linked to by the metadata describing the data set of Max. She downloads the archived version of the Integrated Authority File from four months ago when Max was downloading the network data set from SoNAR for his thesis. She also reads a description of the SoNAR system. The documentation is good enough, she does not start reproducing the data set and the data analysis of Max Mustermann.

5.3 Use Case Descriptions

The table documents use cases underlying both implementation and budget planning. It provides a short but likewise precise overview of the components of SoNAR that will be implemented with the implementation project.

ID	Epic	Use Case	Description	Agents
1	Data harvesting	Set up OAI-PMH data source	As a metadata manager, I would like to set up a new data source (e.g. GND MARC21) based on OAI-PMH technology.	Metadata manager
2	Data harvesting	Set up SRU data source	As a metadata manager, I would like to set up a new data source (e.g. Kalliope EAD) based on SRU technology.	Metadata manager
3	Data harvesting	Set up Museum-Digital data source	As a metadata manager, I would like to set up a new data source (e.g. LIDO) based on Museum-Digital technology.	Metadata manager
4	Data harvesting	Set up SNAC data source	As a metadata manager, I would like to set up a new data source (e.g. EAC-CPF) based on SNAC technology.	Metadata manager
5	Data harvesting	Set up Wikidata data source	As a metadata manager, I would like to set up a new data source based on Wikidata technology.	Metadata manager
6	Data harvesting	Set up MyCORE data source	As a metadata manager, I would like to set up a new data source (e.g. Emporion) based on MyCORE technology.	Metadata manager
7	Data harvesting	Remove data source	As a metadata manager, I would like to remove a data source.	Metadata manager
8	Archiving and versioning	Archiving a data source	As a metadata manager, I would like to configure the archiving of a data source	Metadata manager
9	Archiving and versioning	Archiving a data transformation	As a metadata manager, I would like to set up the archiving of a data transformation	Metadata manager
10	Archiving and versioning	Archiving the HNA ontology	As a metadata manager, I would like to set up the archiving of the HNA ontology	Metadata manager
11	Archiving and versioning	Access original data records	As a researcher, I would like to access the original datasets on which the datasets in the SoNAR system are based in order to trace the provenance of the data	Researcher
12	Transformation system	Set up XML data transformation	As a metadata manager, I would like to set up an XML-to-RDF data transformation for a source.	Metadata manager
13	Transformation system	Set up data merging	As a metadata manager, I would like to set up a data transformation across two sources to merge certain values	Metadata manager
14	Transformation system	Set up data transformation as a multi-stage process	As a metadata manager, I would like to set up a multi-step transformation for a source,	Metadata manager

			composed of several existing transformations	
15	Transformation system	Set up data pipeline	As a metadata manager, I would like to set up a data pipeline from a source via the transformation to a data sink	Metadata manager
16	Transformation system	Delete data pipeline	As a metadata manager, I would like to delete a data pipeline	Metadata manager
17	Transformation system	Validate data pipeline	As a metadata manager, I would like to check a data pipeline for errors	Metadata manager
18	Transformation system	Set up data validation after transformation	As a metadata manager, I would like to set up a validation for a transformation step	Metadata manager
19	Transformation system	Set up JSON data transformation	As a metadata manager, I would like to set up an JSON-to-RDF data transformation for a source	Metadata manager
20	Transformation system	Remove data transformation	As a metadata manager, I would like to delete a configured transformation	Metadata manager
21	Knowledge graph	Configure graph database as data sink	As a metadata manager, I would like to configure a graph database as a data sink	Metadata manager
22	Knowledge graph	Remove data sink	As a metadata manager, I would like to delete a data sink	Metadata manager
23	Knowledge graph	Configure data sink for HNA ontology	As a metadata manager, I would like to configure a graph database as a data sink for the HNA ontology	Metadata manager
24	Knowledge graph	Validate data sink	As a metadata manager, I would like to validate the data before transferring it to a data sink	Metadata manager
25	Web application programming interface	Obtain data via a REST API	As a researcher, I would like to select data programmatically and obtain it as JSON / CSV / XML	Researcher
26	Data selection interface	Select data through targeted searches	As a researcher, I would like to select a partial dataset	Researcher
27	Data selection interface	Select data through browsing exploration	As a researcher, I would like to select a partial dataset through exploratory browsing	Researcher
28	Data selection interface	Visualize temporal information and relationships	As a researcher, I want to visualize temporal information of the data I have selected or the relationships between them	Researcher
29	Data selection interface	Visualize personal relationships "egocentrically"	As a researcher, I would like to visualize relations between persons and other entities (persons, corporate bodies, keywords, etc.) "egocentrically", i.e. with a concrete entity as the center of its network	Researcher

30	Data selection interface	Interrupt and continue work / share link	As a researcher, I would like to interrupt my work and continue the research I have started after a break by saving or sharing a link	Researcher
31	Data download	Download data for external analysis	As a researcher, you would like to carry out a data analysis outside the SoNAR system	Researcher
32	Data download	Download section of the HNA graph data	A third-party system wants to download specific parts of the database via an interface in order to use or further process the data	Third-party system
33	Data download	Download a complete dump of the HNA graph	A third-party system would like to download the entire database via an interface in order to use or further process the data.	Third-party system
34	Data download	Trace the genesis of data at a given point in time	As a researcher reading an article that cites data from SoNAR, I want to be able to understand how this data was obtained	Researcher
35	Web page	Information about SoNAR	As a researcher, I would like to find out more about working with the SoNAR system, the aggregated data sets and forms of visualization	Researcher
36	Web page	Best practice tutorials	As a researcher, I would like to deepen my understanding of the SoNAR platform through best practice tutorials	Researcher
37	Web page	Documentation (UI and API)	As a researcher, I would like to read documentation about the UI and API functionalities	Researcher
38	Corporate design & identity	Coherent corporate design and identity	As a community manager, I want to address the community in a coherent way, both visually and linguistically	Community Manager

6 Deliverables WP1 and WP2

6.1 Deliverables of WP1

ID	Title	Description
1.1.1.1	OAI-PMH Client	A system capable of fetching data from sources using the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protocol
1.1.1.2	SRU Client	A system capable of performing targeted searches and retrieving data from sources using the SRU (Search/Retrieve via URL) protocol.
1.1.1.3	Custom API Client for SNAC	A system capable of fetching data from the SNAC (Social Networks and Archival Context) database via its custom API.
1.1.1.4	Custom API Client for Museum Digital	A system capable of integrating data from Museum Digital through its custom API.
1.1.1.5	Wikidata Client	A system capable of integrating data from Wikidata.
1.1.1.6	MyCoRe Client	A system capable of integrating research data from MyCoRe-based repositories, with EMPORION as a specific test case.
1.1.1.7	Source Management Interface	An authenticated interface for the metadata manager to configure and manage all data sources efficiently.
1.1.2.1	Archive Management System for Source Data	Implement an archive management system that retains versions of source data for 3 years. This system should ensure the integrity and accessibility of archived data over this period, enabling users to download and verify datasets cited in scientific publications.
1.1.3.1	DSL for Transformation Rules	Design and implement a DSL that allows for the intuitive specification of transformation rules from record-based data formats to graph-based data models. This DSL should be powerful enough to handle complex transformation logic while remaining accessible to users without extensive programming experience.
1.1.3.2	Transformation Rule Management System	Create a system for managing transformation rules. This system should retain all revisions of transformation rules, making them accessible for review and reuse, ensuring transparency and continuity in data processing practices.
1.1.3.3	Data Merging and Integration Framework	Implement a framework that supports the merging and integration of data from multiple source systems according to specified transformation rules. This framework should handle conflicts and duplications intelligently, ensuring a coherent and unified graph structure in the output data.
1.1.4.1	Efficient Data Loading and Updating Mechanism	Create a mechanism that can seamlessly load new or transformed graph data into the graph database, detecting and updating existing data where necessary. This mechanism should handle large volumes of data with minimal performance impact, ensuring data freshness and relevance.
1.1.4.2	Provenance Data Integration	Implement a system for including detailed provenance data within the graph database. This system should capture and link information about the origin, version, and transformation rules applied to each data set, providing a clear audit trail for data sources and processing steps.
1.2.1.1	Domain-Specific Transformation Rules	Craft a detailed set of transformation rules that accurately map incoming record-based data to graph structures reflective of the specific domains of libraries, archives, and museums. These structures should preserve the original context and relationships present in the source data. Engage with domain experts from libraries, archives, and museums to validate the transformation rules and ensure they effectively capture the domain-specific nuances and relationships critical for subsequent historical network analysis. Transformations shall be publicly shared for reusing by tools like Catmandu or Metafacture.
1.2.1.2	Seamless Integration with Graph Database Schema	Facilitate the seamless integration of the domain-specific graph structures into the SoNAR platform's graph database schema,

		ensuring that these structures are accurately represented and stored within the database.
1.2.2.1	Ontology Design for Historical Network Analysis	Create an ontology that comprehensively covers the key entities relevant to historical network analysis (e.g., persons, places, events, organizations) and the types of relationships that can exist between them, ensuring the ontology's alignment with historical research practices and data sources.
1.2.2.2	Definition of Inference Rules	Define a set of logical inference rules within the ontology that allows for the automatic derivation of relationships from the transformed domain-specific data. These rules will enable the SoNAR platform to identify and represent implicit connections within the historical data.
1.2.2.3	Publication and Promotion of the Ontology	Publish and promote the developed ontology within the academic and research communities to encourage adoption and collaboration. This may involve presenting the ontology at conferences, publishing articles, and engaging with relevant research networks and projects.
1.3.1.1	API with Multiple Data Format Support	Design and implement an API that allows data retrieval in JSON, CSV, and GraphML formats, ensuring compatibility with a wide range of research tools and applications.
1.3.2.1	Data Selection Interface	Develop a GUI that prioritizes data selection, enabling users to navigate, explore, and pinpoint historical network data through tailored search, clustering, faceting, and filtering mechanisms based on the ontology. The GUI should be intuitive and focused on facilitating the data selection process, making it accessible and engaging for all users, regardless of their background in historical network analysis and technical skills.
1.3.2.2	Accessible Provenance	Ensure that each data selection is accompanied by accessible and comprehensive provenance information, reinforcing the credibility and context of the data, and aiding users in making informed selections.
1.3.3.1	Temporal Visualization Tools	Implement a timeline feature that enables users to visualize selected data across different historical periods, allowing for the exploration of temporal patterns, trends, and events within the historical network data.
1.3.3.2	Cluster Visualization Features	Introduce tools to visualize clusters within the selected data, highlighting significant groupings, relationships, or patterns that emerge from the analysis, thereby providing users with critical insights into the structure and dynamics of historical networks.
1.3.3.3	Descriptive and Network Statistics	The network visualization provides statistics about agents and agent relations properties to inform users about frequencies and related statistics, e.g. distribution of sex, frequency of occupations.
1.3.4.1	Support for Multiple Download Formats	Ensure the download functionality supports multiple data formats, such as CSV, JSON, and GraphML, catering to the diverse requirements of external analysis and visualization tools like Gephi.
1.3.4.2	Interface for Download	Develop an intuitive interface for the download functionality that guides users through the process of choosing a data format and initiating the download, ensuring ease of use regardless of the user's technical expertise.
1.3.5.1	Overview of Available Data and Features	Present an overview of the types of historical network data available on the platform and highlight key features, including search, visualization, and download capabilities, to inform users about the platform's utility and encourage exploration.
1.3.5.2	Updates and Harvesting Frequency Information	Regularly update the documentation to reflect the latest changes and enhancements made to the platform, including new features and data sources. Provide clear information on the frequency of data harvesting from each source, helping users gauge the freshness and relevance of the data they are working with.
1.3.5.3	Comprehensive Documentation	Develop documentation that covers all aspects of the SoNAR platform, including data selection, visualization, downloading, the ontology, and the API. The documentation should cater to both

		novice users needing guidance on platform basics and advanced users looking for detailed information on specific functionalities.
1.3.6.1	Development of a Visual Identity	Create a visual identity for the SoNAR platform, including a color scheme, typography, logo, and other visual elements that reflect its focus on historical network analysis and appeal to its academic audience.
1.3.6.2	Consistent Design Implementation	Apply the corporate design consistently across all elements of the user interface, including the start page, documentation, complementary pages, and data exploration tools, ensuring a unified user experience that reinforces the platform's brand identity.

6.2 Deliverables of WP2

Number	Title	Description
T2.1.1	Report on expectations and assessment of the advisory board	Documents and summarizes interviews with the advisory board members on initial expectations and assessments (the feedback at joint quarterly reviews, respective conclusion on user value or planned HNR project).
T2.1.2	Report on online consultations for short-term feedback on increments	Documents survey methods and aggregated feedback of online surveys on sprint increments.
T2.1.3	Report on prototype usability tests (MVP, beta release)	Documents MVP and beta release usability tests: methods, in-depth analysis of results, and executive summary.
T2.1.4	Future development roadmap	Outline of requirements – features, data, quality – based on expectations and usage context identified during the project but could not be considered with the product development and might influence future user value (relates to T2.2.4).
T2.1.5	Report on project coordination activities	Documents the progress of the project and related activities, in particular project meetings (part of final project report).
T2.2.1	Report on feedback analysis and product backlog	Documents and summarizes feedback analysis and product development related backlogs (product, sprint, pool of topics).
T2.2.2	Report on social media marketing and website content strategy	Documents the social media marketing and website content strategy and related outreach and dissemination activities.
T2.2.3	Report on community cooperation	Documents the HNR related activities like project workshops and conference including conclusions drawn from community coordination and research discourse on project development.
T2.2.4	Product strategy	Starting from the implemented product, the product strategy specifies product purpose and objectives based on business analytics, especially stakeholder and respective added value evaluation (the larger ecosystem context of SoNAR) including impact of changes like technological advancement on needs and possible actions (relates to T2.2.4, T2.3.1, and T2.3.2).
T2.2.5	Report on go-live	Documents transition activities from the implementation to operating SoNAR.
T2.3.1	Report on the evaluation of full text aggregation services	Documents activities and summarizes the outcome of the evaluation of AI/ML technologies to enhance user value
T2.3.2	Report on the evaluation of network data enrichment services	Documents activities and summarizes the outcome of the evaluation of AI/ML technologies to enhance user value
T2.3.3	Report on educational materials	Documents activities and summarizes the outcome of the work on educational materials and best practice tutorials

7 Operating the SoNAR research data market

8 Risks