**Task Area 4 – Integrating Data, TA-Lead: Jan Goebel (SOEP/DIW Berlin), Claudia Niessner/Katja Keller (KIT), Coordination: Janina Britzke / Alexander Jung (SOEP/DIW Berlin)**

**Übersicht:**

|  |  |  |
| --- | --- | --- |
| **M4-1:** BASiD – Administrative Data Linkage for Research and Policy Advice Profile: BASiD | BA, DRV, IAB | - |
| **M4-2:** Contract Templates Profile: Contract Templates | DZHW | - |
| **M4-3:** CrisesData – Establishing Infrastructures for Rapid Responses in Societal Crises Profile: CrisesData Category: Databases | U Berlin, WZB/RatSWD | Stefan Liebig, Andrés Saravia |
| **M4-4:** DisclosureControl – (Semi-)Automatic Disclosure Control for RDC Profile: DisclosureControl | BAMF, Destatis, IAB, KBA | - |
| **M4-5:** RDC Motor Performance and Physical Fitness – First RDC for Sports Science Data Profile: RDC Motor Performance (FDZ MoL) | KIT | Claudia Niessner, Katja Keller, Chris Rose, Hannah Zimmermann |
| **M4-6:** FirmData – Continuous Provision of Firm Data (A) Profile: FirmData | EBDC, RDC Ruhr at the RWI, ifo | Philipp Breidenbach, Sebastian Wichert, Philip Raatz |
| **M4-7:** Linking Data Profile: Linking Data | BAMF, Destatis, IAB | - |
| **M4-8:** QualidataNet Portal – Federated Archiving Infrastructure and Technical Solutions for Qualitative Data Profile: QualidataNet Portal | DIPF, DIW Berlin, DZHW, eLabour, IDS, RDC Qualiservice/ University of Bremen | Betina Hollstein, Kati Mozygemba |
| **M4-9:** StanDem Profile: StanDem | GESIS | Silke Schneider |

Task Area 4 "Integrating Data" (TA4) focuses on developing and implementing innovative technologies and methods for the integration, anonymisation, and harmonisation of data. These initiatives are crucial for improving data quality and ensuring secure access to important datasets. The primary goal is to facilitate comprehensive and seamless data integration, thereby enhancing scientific re-search and evidence-based decision-making through high-quality, accessible, and secured datasets.

**Contributors:** Both KIT and SOEP/DIW Berlin are involved in measures and jointly represent the role of co-spokespersons in TA4. In addition, other measure participation comes from IAB, DRV, DZHW, Destatis, BAMF, RWI, ifo, University of Bremen, and GESIS. Other contributors recur in their established roles: DIPF, FDZ-BO, IDS and eLabour to offer their expertise.

**Objectives & Measures:** TA4 aligns with five key KonsortSWD objectives: expanding data access (O1), automating and standardising RDM (O3), assuring reliability of services (O6), learning through collaborations (O8), and in particular, linking data across domains (O9).

To achieve these objectives, TA4 will implement nine targeted measures. These measures include providing centralised access to specific disciplines or data types through QualidataNet (M4-8), FirmData (M4-6), and FDZMoL (M4-5), as well as offering standardised contract templates (M4-2) and automating anonymisation and output control (M4-4). Additionally, TA4 will develop a framework for direct data linking (M4-7), link social security data through BASiD (M4-1), enhance the crises-related data platform (M4-3), and support standardised sociodemographic measurement through StanDem (M4-9). By advancing data integration, TA4 aims to provide researchers with high-quality, well-curated data, ultimately supporting robust scientific inquiry and thus indirectly policymaking. Be-low, measures are sorted alphabetically.

**Risks:** While TA4 aims to enhance data integration and accessibility, several risks must be ad-dressed. One major risk is the potential for reduction in the requested funding, which could hinder development and implementation. Additionally, recruiting and retaining skilled personnel under project-based financing is challenging, especially in a competitive job market for data scientists and IT professionals, potentially impacting project continuity and success.

On a technical level, the complexity of automating anonymisation processes and ensuring secure data linking involves inherent risks. Technical failures or delays in these areas could compromise the overall objectives of TA4. To mitigate these risks, we have continuous risk assessments. Implementing regular workshops within the Task Area can help keep the team updated on the latest technologies and best practices. For example, to prevent incomplete data integration due to varied for-mats, we provide comprehensive training for data providers (M4-4). To address firms’ data privacy concerns, we establish strong data privacy protocols and offer clear communication about security measures (M4-6). To overcome technical difficulties in providing remote access, we invest in robust IT infrastructure and offer continuous technical support (M4-5). To improve limited access to social security data despite bureaucratic hurdles, we partner with government agencies and streamline access procedures (M4-1 and M4-3).

**Cooperation:** We will work with TA1 to involve the community, in particular planned events and surveys. TA4 also cooperates with TA5 to incorporate the possibilities of business models for consolidated services. There is a connection to TA2 and TA3, as the QualidataNet framework is represented with measures in all of the three Task Areas from TA2 to TA4. Especially for providing FAIR access to integrated data, we cooperate with TA2.

**Conclusion:** While TA4 presents various challenges, targeted solutions, and proactive risk management strategies can effectively mitigate these risks, ensuring successful data integration and accessibility.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **M4-1: BASiD – Administrative Data Linkage for Research and Policy Advice Profile: BASiD** | | | | | Category: Data Curation | | | | |
| DFG Key Points | **A** | **B** | **C** | **D** | | **E** | **F** | **G** | **H** |
| **Target Group:** Researchers, RDCs | | | | | **Section: -** | | | | |
| **DFG KPI:** Number of datasets curated | | | | | **Partners:** BA, DRV, IAB | | | | |
| **Prop. KPI:** Number of users of datasets | | | | | **Basic Service: -** | | | | |
| **Internal Collab.: -** | | | | | **Status:** Development | | | | |
| **Consortia:** BERD@NFDI, NFDI4Health (planned) | | | | | **External Collab.:** SHARE, Public Employment Services | | | | |

**Added:** The proposed project adds substantial value by developing a harmonised multi-pur-pose research data product that combines invaluable administrative datasets from the Statutory Pension Insurance (DRV) and the Federal Employment Agency (BA), significantly enhancing re-search quality and enabling diverse research projects simultaneously. The best sources for research on employment, social security in old age, and unemployment are provided by the DRV and the BA. However, both lack information the other could deliver. For example, data from the Institute for Employment Research (IAB) lacks details on fertility, care duties, or divorce, while DRV data lacks in-formation on job search and work environment. A combined, harmonised dataset will enhance in-sights into human behaviour, social dynamics, and policy impact, providing evidence-based information for policymakers, practitioners, and other stakeholders. This measures’ final dataset will pro-vide cleaned data with German and English documentation, enabling easily reproducible and com-parable results with less time expenditure compared to individually linked datasets. The proposed data product renews the linkage of confidential social insurance data, similar to the existing dataset BASiD 2008 “Biographical Data of Selected Insurance Agencies in Germany” (Hochfellner et al., 2011). It includes a 1% representative sample of the DRV population and is enriched with adminis-trative data from the Federal Employment Agency (BA) provided by the IAB. BASiD has been widely used in numerous national and international research projects. The new project will build on this successful linkage of administrative data in a future-oriented and sustainable way by increasing the sample size to 2% (1.4 million insured persons). Other social insurance institutions can contribute their data to close knowledge gaps, and harmonisation scripts will enable future linkage projects to join BASiD.

**Existing user base:** Until mid-2024, 103 researchers from 56 projects have used BASiD at IAB. Additionally, 164 researchers in 66 international research projects used the scientific-use-file published by DRV. Researchers from institutions such as London School of Economics, University of Michigan, Oxford, Stanford, and Stockholm Universities have applied for BASiD. The international outreach will likely grow through this measure’s expanded data access points and improving remote data access at RDCs.

**Work Programme - Summary:** The project starts with conceptual work to harmonise data between the two major social insurances, as they store data differently. A second step includes data pre-processing, linkage, and creating standardised routines for periodic updates. Additionally, standard contracts for collaborating institutions and standardised programming routines for periodic updates will be implemented.

**Detailed Plan:** The project starts after the completion of preparatory works, such as cooperation contracts, exchange with data protection officers, data preparation, and data access between the cooperation partners. This is planned for late 2025 and early 2026. 2026 also includes population sampling, data linkage on individual level, harmonisation of parallel events like employment and care work, and the creation of a first linked dataset. In addition, the PhD students involved in the project start their empirical work with a research question to verify the harmonisation of the linked data. The second year, 2027, focuses on four modules. Firstly, a concept of sustainability of the data product in terms of expansion and periodic updates, independent of funding. The second module is a concept for extension to other social security institutions such as accident insurance or health insurances (in cooperation with NFDI4Health). The third module will focus on test projects with researchers, which will be open to different disciplines within and outside NFDI. Fourthly, the PhD students will present their own work at conferences, and the project team will promote the upcoming data product at various events, concentrated in the research community and the European social security infrastructure as well as interested NFDI consortia, CESSDA and networks like the International Data Access Network. As a result of these modules, a business model for periodic updates of BASiD and linked data components will be established (Q4/27). The third year, 2028, will focus on feedback from the test projects through community workshops (Q2/28), updating the data documentation, and releasing the data in line with the FAIR principles to the international research community (Q3/28).

**Resources:** To deliver the results described above, KonsortSWD requests 45 person months for Doctoral researchers and comparable plus direct costs for expenses (cf. 3.4.6).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **M4-2: Contract Templates Profile: Contract Templates** | | | | | Category: Support/Consulting | | | | |
| DFG Key Points | **A** | **B** | **C** | **D** | | **E** | **F** | **G** | **H** |
| **Target Group:** Researchers, RDCs, Data professionals | | | | | **Section: ELSA** | | | | |
| **DFG KPI: -** | | | | | **Partners:** DZHW | | | | |
| **Prop. KPI:** Number of RDCs and other types of institutions using parts of or complete contract templates | | | | | **Basic Service: -** | | | | |
| **Internal Collab.:** M3-7, M3-9, FDI Committee | | | | | **Status:** Operation | | | | |
| **Consortia: -** | | | | | **External Collab.: -** | | | | |

**Existing user base:** At least four RatSWD-accredited RDCs and one other party are already using the contracts. Five more accredited RDCs plan to incorporate them in their next contract revisions. Some RDM advisory centres use the templates in trainings and for legal guidance on data intake and access. High download numbers on Zenodo (1,596 for data provision documents and 2,233 for data use documents in July 2024) hint at a broad perception.

**Work Programme - Summary:** Feedback will be gathered from the RDCs and other users already working with the contract templates. These templates and associated legal documents will be ex-panded with input from a specialist law firm to address new legal requirements and resolve issues which have arisen in practice. The revised and finalised documents will then be translated and pro-moted within the community in collaboration with the specialist law firm and the community measures M1-3 and M3-2.

**Detailed Plan:** Feedback collection starts in September 2026 to allow RDCs to gather practical experience with the current model contracts beforehand. This initiative aims to update and expand contract templates and accompanying documents based on user feedback, potential new users, and legislative changes. The above-mentioned law firm will assist in supporting and legally securing the contract templates. Afterwards, we start collecting feedback on the contract template for data provision. The project supported by the law firm will revise the contract template and explanatory document, followed by a workshop for further input (Q3/27). Based on the gained feedback, these documents will be finalised, resulting in a report on future revision needs and possibly recommendations from a working group organised via FDI Committee (Q4/27). The revision process will then repeat for the data use contract template and explanatory document, including another workshop on the data use contract template (Q2/28). As soon as the contract templates for data provision and data use have been finalised, the mutual dependencies between the two contract templates are thoroughly reviewed, taking RDC processes into account, and adjusted if necessary (Q3/28). The tem-plates and explanatory documents will then be translated into English (Q4/28), followed by promotional efforts through workshops and e-mail lists. Lessons learned from the initial funding phase show that obtaining legal feedback requires considerable preparation time involving RDC personnel and legal experts. To address this, the work programme steps accommodate these timelines. Feedback indicated a preference for earlier involvement of RDC staff, data protection officers, and legal staff via workshops, which we’ve implemented in the revision phase. Contract template promotion will involve RatSWD, section ELSA, similar consortia, and other RDM stakeholders. For contracts, an operating model is not essential. Nevertheless, it seems useful to discuss issues and gaps with contract users at longer intervals and, if necessary, to revise the model contracts together in order to promote the harmonisation of contracts and processes.

**Resources:** To deliver the results described above, KonsortSWD requests 11.5 person months for Doctoral researchers and comparable plus direct costs for expenses (cf. 3.4.6).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **M4-3: CrisesData – Establishing Infrastructures for Rapid Responses in Societal Crises** | | | | | Category: Databases | | | | |
| DFG Key Points | **A** | **B** | **C** | **D** | | **E** | **F** | **G** | **H** |
| **Target Group:** Researchers | | | | | **Section: -** | | | | |
| **DFG KPI:** Number of visits | | | | | **Partners:** FU Berlin, WZB/RatSWD Working Group | | | | |
| **Prop. KPI:** Number of downloads of survey instruments; number of cooperation partners | | | | | **Basic Service:** - | | | | |
| **Internal Collab.:** M1-1, M4-9 | | | | | **Status:** Development | | | | |
| **Consortia:** NFDI4Earth, NFDI4Health | | | | | **External Collab.:** | | | | |

Function: The project, which has been funded through the consortium’s network development grants since December 2023, (1) facilitates the access to crisis-related data from different scientific disciplines and non-academic contexts, (2) provides a collection of tested, crisis-related survey instruments, and (3) establishes a structure to enable the coordination between researchers, practitioners, and decision-makers in immediate crisis and disaster situations.

Value Added: The CrisesData service adds value by strengthening the infrastructure for analysing the vulnerability, preparedness, and resilience of individuals and social structures in crisis and dis-aster situations. This measure is particularly relevant as it addresses the need for improved findability, accessibility, and cross-disciplinary linkage of crisis-related data from various sources, building on the project “Better results through interoperability and standardised research data management (BestFDM)” (Lenzner et al., 2022; RatSWD, 2023d, 2023f). By collaborating with other NFDI con-sortia, such as NFDI4Earth and NFDI4Health, and emphasising a social science perspective in crisis research, the service supports the interdisciplinary exchange of competencies and facilitates more efficient contributions from KonsortSWD and external partners, ultimately making vital data accessi-ble to new stakeholders like civil protection organisations and policymakers. The service also aims to improve data interoperability by harmonising crisis-related survey instruments and establishing a coordination centre for rapid knowledge exchange.

Existing user base: Service in development

Work Programme - Summary: In the second funding phase, the crisis-related data platform and survey instruments will be consolidated and improved. A coordination centre at the RatSWD office (M1-1) will unite researchers, practitioners, and decision-makers in immediate crisis and disaster situations, improving quick access to quality-secured data.

Detailed Plan: In Q4/25 and Q1/26, the information platform and survey instruments will be evaluated through a user survey on usability, missing information, and improvement suggestions. The findings will be documented in an evaluation report (Q2/26) with recommendations for modifications. Technical expertise will guide suitable modifications to the website, including structural design and content expansions. The survey instruments will be enhanced for usability and additional topics, and integrated into the ZIS Open Access Repository where possible. A coordination centre concept will be developed with a RatSWD working group accompanying this project, outlining procedures to unite researchers, practitioners, and decision-makers in crisis and disaster situations, ensuring quick access to quality-secured data and rapid responses (Q3/26). The concept will formulate general steps and specific crises scenarios. A stakeholder workshop will review the concept in terms of feasibility and possible integration into existing structures, followed by the launch of the coordination centre (Q4/26). To consolidate the network and strengthen cross-sector linkages, a crisis data network meeting will be organised after one year (Q4/27). This meeting will discuss current developments in crisis research and management and their implications for data infrastructures. In the case of further continuation of the service after 2027, the network meeting should take place once a year. With the support of M5-1, the project will develop a business model (Q4/26-Q2/27) to consolidate the service and ensure long-term sustainability. This process will begin at least one year before funding concludes, incorporating the advice of the RatSWD working group and NFDI project partners, while considering existing structures in crisis and disaster research and management. Over the entire du-ration of the project, its three pillars (information platform, collection of survey instruments, coordination centre) will be promoted among the stakeholders relevant to crisis and disaster fields, including different scientific communities and data-interested practitioners. Over the entire duration of the project, it will be promoted among the stakeholders relevant to crises and disasters. Existing collab-orations with the Federal Office for Civil Protection and Disaster Assistance (BBK), the German Committee for Disaster Risk Reduction (DKKV) and notably NFDI4Earth, and NFDI4Health will be leveraged and expanded through collaborative opportunities. The service will closely cooperate with the Public Relations Office (M1-4) and develop further strategies of dissemination.

**Resources:** To deliver the results described above, KonsortSWD requests 27 person months for Postdoctoral researchers and comparable plus direct costs for expenses (cf. 3.4.6).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **M4-4: DisclosureControl – (Semi-)Automatic Disclosure Control for RDC Profile: DisclosureControl** | | | | | Category: Workflow/Pipeline | | | | | |
| DFG Key Points | **A** | **B** | **C** | **D** | | **E** | **F** | **G** | **H** |
| **Target Group:** Researchers, RDCs | | | | | **Section: -** | | | | | |
| **DFG KPI:** Number of executions | | | | | **Partners:** BAMF, Destatis, IAB, KBA | | | | | |
| **Prop. KPI:** Number of stakeholders implementing the solution, number of data users benefiting from the solution | | | | | **Basic Service: -** | | | | | |
| **Internal Collab.: M2-6** | | | | | **Status:** Development | | | | | |
| **Consortia:** NFDI4Health, BERD@NFDI, GHGA, NFDI4Memory | | | | | **External Collab.:** UNECE, IDAN | | | | | |

disclosure control practices across RDCs, where manual disclosure control is currently a time-consuming and resource-intensive task that binds highly-skilled staff. By semi-automation, researchers can receive results more swiftly, thus minimising waiting times and ensuring that output is consistently checked against standardised rules, ultimately enhancing the overall quality of services pro-vided by RDCs**.**

**Existing user base:** Service in development

**Work Programme - Summary:** The measure develops a concept for semi-automating the disclosure control of sensitive data and implements a tool that can be used by various stakeholders. First, requirements for disclosure control and (IT) infrastructure are collected incorporating stakeholder needs. Building upon existing partial solutions, code is developed. The open-source solution allows stakeholders to adapt it to institutional requirements, contribute improvements, and guarantee its sustainability. Transparent documentation and training tools enable researchers to understand and comply with the rules implemented. Re-use scenarios of the tool will be discussed with other consortia handling sensitive data.

**Detailed Plan:** The measure develops a tool that (semi-)automates disclosure control. Most data providers have developed routines and rules for manual or even partially script-based disclosure control. Therefore, the measure first collects pre-existing solutions and rules, as well as the stake-holder’s needs for a semi-automated solution. Then, a concept including criteria for the programming of the solution is designed (Q3/26). During this stage, a close collaboration with the stakeholders is necessary to ensure that the solution to be developed is useful to a broad target group. The concept is then programmed and tested. For the solution to work for as many stakeholders as possible, the program has to be open source and easily adaptable. Thus, a suitable software has to be chosen. Once a minimum viable product (MVP) is designed and documented, the solution will first be tested and debugged within the group of partners in this measure (Q4/26). Thereafter, stakeholders identified in the beginning of the work package will be asked to participate in the testing to further develop the solution. Throughout the programming and testing phase, transparent documentation for data users is developed resulting in a final report with a business model (Q4/27). The documentation’s goal is to make the rules of disclosure control transparent to data users and foster a deeper understanding of its importance. Additionally, training tools (for example videos, documents, etc.) are developed to further enable data users to code in a way that makes disclosure control as little restrictive as possible (Q2/28). For long-term sustainability, all programs, documentation, and training tools are made available via open source such that all potential stakeholders can benefit and further develop the service (Q4/28).

**Resources:** To deliver the results described above, KonsortSWD requests 29.5 person months for Doctoral researchers and comparable, and 19.5 person months for other staff plus direct costs for expenses (cf. 3.4.6).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **M4-5: RDC Motor Performance and Physical Fitness – First RDC for Sports Science Data Profile: RDC Motor Performance (FDZ MoL)** | | | | | Category: Data Curation | | | | |
| DFG Key Points | **A** | **B** | **C** | **D** | | **E** | **F** | **G** | **H** |
| **Target Group:** Researchers, RDM professionals | | | | | **Section:** Infrastructure | | | | |
| **DFG KPI:** Number of datasets curated | | | | | **Partners:** KIT | | | | |
| **Prop. KPI:** Number of datasets published, downloads | | | | | **Basic Service: -** | | | | |
| **Internal Collab.:** M2-4, M3-6, M3-7, M4-8 | | | | | **Status:** Development | | | | |
| **Consortia:** NFDI4Health | | | | | **External Collab.:** Leibniz Science Campus DiTraRe | | | | |

**Function:** This RDC supports sports scientists in sharing and reusing sensitive data and in acquiring expertise in research data management.

**Value Added:** As an interdisciplinary field, sports science bridges the humanities, social-, and natural sciences, creating synergies that benefit all stakeholders. In summer 2023, the German Sports Association’s ad hoc committee “RDM in sports science” (co-chairperson: Katja Keller) surveyed 122 researchers at different career stages about RDM. The results illustrate that researchers work with multiple data types and follow a multiple-storage strategy, mainly using business computers and institutional servers for data storage. With the first repository (MO|RE data) and RDC in sports science, the RDM group at the KIT’s sports institute serves as the primary contact for sports science researchers and from related fields in KonsortSWD, like education, economics, or social sciences.

**Existing user base:** The MO|RE data repository underpins the RDC with public use file access (Klemm, Bös, et al., 2024). It stores 41 datasets (19 aggregated, 22 raw) by 58 authors, totalling 63,117 items. Datasets include motor performance data of 23 harmonised test items and additional data as age, gender and body composition and were published between 1970 - 2024.

**Work Programme - Summary:** Building on the foundation of the RDC for Sports Science in the first funding period (Klemm, Kron, et al., 2024), with initial access to sensitive data from 2024-25, the following goals are being pursued: further development of interdisciplinary data use (mapping/harmonisation, metadata), enhancing RDM support for sports scientists, creating a dashboard for RDC data, and developing remote access to analyse sensitive data without data exchange.

**Detailed Plan:** A key deliverable is a robust database that will house all the data collected and processed by the RDC (Q4/27). This database will serve as a valuable resource for researchers and stakeholders involved in the project. Among the project’s milestones is the development and implementation of a revised mapping scheme (Q3/26). This will involve updating the existing mapping scheme to enhance data visualisation and interpretation. We will also create a revised metadata scheme, in line with KonsortSWD practice, that will improve the organisation, accessibility, and us-ability of the data stored in the RDC database (Q3/26). A significant milestone is the launch of a Data Dashboard (Q4/26). This interactive platform will allow users to easily navigate, analyse, and interpret the data stored in the RDC database. Finally, we plan on implementing a concept for remote access (Q3/27), (M2-4). This will enable researchers to access the RDC database remotely, increasing accessibility. We also plan to develop a comprehensive business model (Q4/27) that will be implemented post-2027. This model will outline the strategies, revenue streams, and operational plans to ensure the sustainability and growth of the project.

We regularly present the RDC’s progress in digital exchange formats such as NFDI Talks and at relevant national and international conferences in the fields of RDM, sports, and health sciences.

**Resources:** To deliver the results described above, KonsortSWD requests 27 person months for Postdoctoral researchers and comparable plus direct costs for expenses (cf. 3.4.6).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **M4-6: FirmData – Continuous Provision of Firm Data (A) Profile: FirmData** | | | | | Category: Data Curation | | | | |
| DFG Key Points | **A** | **B** | **C** | **D** | | **E** | **F** | **G** | **H** |
| **Target Group:** Researchers | | | | | **Section:** Industry | | | | |
| **DFG KPI:** Number of datasets curated | | | | | **Partners:** EBDC, RDC Ruhr at the RWI, ifo | | | | |
| **Prop. KPI:** Usage of datasets | | | | | **Basic Service:** - | | | | |
| **Internal Collab.:** M2-4, M3-7 | | | | | **Status:** Development | | | | |
| **Consortia:** BERD@NFDI | | | | | **External Collab.:** Stifterverband | | | | |

**Function:** Continuous Provision of Firm Data is an extended service based on the project "Access to Firm Data" from the first funding phase, also turning from a *support* to a *data curation service*. It provides data through cooperations with firms as well as a systematic collection of research (data) cooperations between academia and firms using well-established German firm surveys.

**Value Added:** The measure provides sustainable, long-term, data access to firm datasets, usually inaccessible to research (Gottschalk et al., 2023), and documents firms' data-cooperation with academia.

**Existing user base:** Service in development

**Work Programme - Summary:** The FirmData measure focuses on the sustainable and long-term provision of FAIR data products for the social and economic sciences, derived from firm cooperations. We develop semi-automatic curation routines to streamline future data deliveries from firms, minimising the cost of data provision. Additionally, established ifo Surveys will gather detailed over-views on cooperations between firms and researchers.

**Detailed Plan:** "Access to Firm Data" initiates firm cooperations and provides obtained data to the scientific community, offering significant value for empirical research across various fields. This typ-ically inaccessible data is curated and documented in a FAIR form. However, the initial data provision will become outdated over time and therefore of limiting use for future policy-relevant research. For this reason, the first work package of the extended service focuses on the sustainability and extension of the data provision initiated during the first funding period. On the one hand, firm cooperations are to be reorganised in such a way that sustainable data provision is made possible (Q1/26). On the other hand, reusable coding pipelines and routines are to be developed (Q2/26) to enable the low-cost, nearly standardised, and automatic curation of additional (data) waves, even after the end of the project. In the second work package, the measure aims to create a systematic and long-term

overview of existing data cooperation between German companies and the scientific community through expert interviews and data from ifo Business Survey and Management Surveys (Q4/25 and Q1/26). Our goal is to provide up to three datasets from firm cooperations on a long-term and sustainable basis (Q3/26). To make them known throughout the desired research community, we pro-mote the new data products on websites and social media platforms, e.g., with the Public Relations Office (M1-4) as well as Forum4MICA (M3-4) (Q4/26). The two work packages are linked to the initiatives mentioned in the profile table above and will each develop a suitable business model for the data products (Q4/27).

**Resources:** To deliver the results described above, KonsortSWD requests 16.5 person months for Postdoctoral researchers and comparable plus direct costs for expenses (cf. 3.4.6).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **M4-7: Linking Data Profile: Linking Data** | | | | | Category: Data Curation | | | | | |
| DFG Key Points | **A** | **B** | **C** | **D** | | **E** | **F** | **G** | **H** |
| **Target Group:** Researchers, RDCs | | | | | **Section: -** | | | | | |
| **DFG KPI:** Number of datasets curated | | | | | **Partners:** BAMF, Destatis, IAB | | | | | |
| **Prop. KPI:** Number of linkage projects, downloads of templates | | | | | **Basic Service: -** | | | | | |
| **Internal Collab.: -** | | | | | **Status:** Development | | | | | |
| **Consortia:** BERD@NFDI, NFDI4Health | | | | | **External Collab.:** German Record Linkage Center | | | | | |

**Function:** This measure enables the linkage of sensitive data for the same individuals held by dif-ferent RDCs and other data owners for the scientific community.

**Value Added:** Value is added by creating a vast data treasure that opens new analysis potential for the scientific community while providing essential support to facilitate the creation of linked datasets, effectively breaking down existing data silos. This is especially relevant as linking sensitive datasets from different data holders is currently limited by legal frameworks and discrepancies in methodology and definitions among data providers, hindering the full potential of data analysis and research. By developing a generic framework to facilitate the practical implementation of specific linkage projects, preparing for future projects through coordinated workflows, and providing funding for specific link-age projects, this measure addresses the practical, financial, and methodological barriers to linking datasets, ultimately enabling the scientific community to open up new analysis potential and advancing research in the field. Further linkage projects can be modelled from NFDI4Health or BERD@NFDI. This approach is analogous to linkage projects funded by the Austrian Academy of Science.

**Existing user base:** Service in development

**Work Programme - Summary:** Based on extensive experience from related projects (German Record Linkage Centre by IAB), a pilot project between the Deutsche Bundesbank and the Federal Statistical Office (Statistisches Bundesamt (Destatis), 2023), as well as an EU-funded project to link

migration register and survey data (BAMF), this measure will provide a generic and versatile frame-work to facilitate the practical implementation of different linkage projects. It will also support the implementation of selected linkage projects. The service “Linking Data” supports data-holding institutions by providing organisational expertise concerning data linkage projects, legal expertise and financial support for these, as well as best practices for the technical process of linkage data from different data holders.

**Detailed Plan:** The service “Linking Data” provides a versatile basis for linking data from different sources via two actions. As a basis, it enables data-holding institutions to link their data amongst each other. In addition to flexible financial support for linkage projects, action 1 will provide a generic framework, developed as part of the work package, to facilitate the practical implementation of specific linkage projects and re-use their data. Both aspects are substantial challenges for administrative data in Germany. Initially, organisational, technical and legal challenges will be identified. Given the partners’ highly specialist experiences from previous linkage projects, earlier observed challenges can be avoided. The aim is to create and publish essential documents that are necessary for the realisation of linkage projects, such as declarations of consent for linkage permission or templates for contracts between data owning institutions (Q1/27). A further focus lies on developing workflows (Q1/27) that are required to make the linked data products re-usable for the scientific community via RDCs in the long term. The service thus aims to overcome existing hurdles and prepare for the expansion of linkage projects that will follow the implementation of the planned Research Data Act. Against this backdrop, it will also develop a business model to sustain linkage projects re-using the methods developed here (Q3/27). Action 2 will support promising linkage projects through application and selection processes (Q1/26). This includes tender texts or templates for the expression of interest. The project will draw on experiences from other NFDI consortia for these steps. Moreover, a catalogue of evaluation criteria will be developed (Q1/26). The selection of projects will be aligned with the selection of flex funds (cf. 3.4.5) whenever possible.

**Resources:** To deliver the results described above, KonsortSWD requests 66 person months for Doctoral researchers and comparable plus direct costs for expenses (cf. 3.4.6).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **M4-8: QualidataNet Portal – Federated Archiving Infrastructure and Technical Solutions for Qualitative Data Profile: QualidataNet Portal** | | | | | Category: Web Application | | | | |
| DFG Key Points | **A** | **B** | **C** | **D** | | **E** | **F** | **G** | **H** |
| **Target Group:** Researchers, RDCs, Data professionals | | | | | **Section:** ELSA, EduTrain, Metadata, Infrastructure | | | | |
| **DFG KPI:** Number of visits | | | | | **Partners:** DIPF, DIW Berlin, DZHW, eLabour, IDS, RDC Qualiservice/ University of Bremen | | | | |
| **Prop. KPI:** Number of downloads | | | | | **Basic Service: -** | | | | |
| **Internal Collab.: -** | | | | | **Status:** Operation / Extension | | | | |
| **Consortia:** NFDI4Health, Text+, NFDI4Biodiversity, NFDI4Culture | | | | | **External Collab.:** SSHOC, ISDFPN | | | | |

**Function:** Promote and facilitate access to and findability of qualitative data by providing a coordi-nated, user- and service-oriented infrastructure. Beyond service operation and network extension, QualidataNet will be extended to include: a) a data protection-compliant anonymisation tool suited for the complex anonymisation and pseudonymisation of text data and b) a survey on the need for common technical service components for the preparation, archiving, and provision of qualitative research data at the RDCs and data holders.

**Value Added:** For the first time, QualidataNet offers a central point of entry for finding, managing, sharing, and re-using sensitive qualitative data (Mozygemba et al., 2023). QualidataNet connects relevant actors and addresses the needs of different qualitative communities in a coordinated manner, thus counteracting fragmentation and parallel developments in this area. The integration of further qualitative archives and RDCs increases the scope of the network. Researchers get advice on archiving options and find the archiving partner that suits them best. The joint platform increases the FAIRness and visibility of qualitative data. This FAIRness is also optimised by the controlled vocabulary “QualiTerm”. Low-threshold networking in QualidataNet also enables smaller archives to participate in exchange and joint development of quality criteria and standards. QualidataNet also main-tains and provides an existing data protection-compliant anonymisation tool as service for textual data for researchers and RDCs. This tool will be maintained and developed further as part of Kon-sortSWD’s portfolio. The tool significantly facilitates the complex anonymisation and pseudonymisation of text data by offering flexible replacement levels and transparent documentation. It reduces costs and increases reliability of the time-consuming processing of qualitative data. Exportable re-placement concepts enable researchers to share topic-specific replacement schemes, which can be further developed by RDCs and scientific communities. Previously unmet needs of the communities (e.g. implementation of semi-automatic recognition in the marking of sensitive information, options for teamwork on an anonymisation project, the integration of the anonymisation of audio files etc.) will be addressed and possibilities for implementation will be examined. Finally, a survey on the need for technical solutions and service components for preparing, archiving, and providing qualitative data will lay the groundwork for advancing the services of qualitative RDCs.

**Existing user base:** Launched in December 2023, the QualidataNet website had 700 visits in its first month. Several RDCs using qualitative data have integrated elements of the “QualiTerm” vocabulary into their metadata descriptions. The QualidataNet network consists of six data providers, with others interested in joining. Since 2021, around 500 users have downloaded the anonymisation tool, with regular requests for use and development from various disciplines. RDC Qualiservice ap-plies the tool in data curation, and it was presented in an international webinar at the International Association for Social Science Information Service & Technology (IASSIST) qualitative working group.

**Work Programme - Summary:** The measure continues and expands QualidataNet as the central point of entry for qualitative research. Activities include network coordination, exchange on quality standards for data curation workflows, services for researchers, advancement of FAIR metadata, and controlled vocabularies for qualitative data. Services for researchers include information and advice on suitable archiving partners, qualitative datasets, and aspects of data sharing. The network continuously integrates new partners. The measure also includes the representation in committees and the creation of a business model for sustainable operation.

**Detailed Plan - Operation:** Action 1 provides the continuous operation and further expansion of QualidataNet, which was successfully implemented as the central point of entry for qualitative data. It coordinates the exchange within the network and joint reconciliation of quality standards as well as the integration of further network partners beyond KonsortSWD. Researchers from various disciplines receive support in finding suitable archiving partners, which will be further facilitated by an archive-wizard (Q3/27). Data-holding institutions receive support in questions of archiving and the provision of qualitative research data. The measure includes the continued operation, maintenance, and expansion of the search portal and the harvester, the integration of additional datasets, the optimisation of the searchengine, and the support of data holders in questions of metadata standards for the description of qualitative data as well as the maintenance and provision of the controlled vocabulary QualiTerm in order to refine the search results and increase the FAIRness of qualitative (meta-)data. The measure will also develop a business model for the sustainable and long-term operation of the service (Q4/27).

**Detailed Plan - Extension:** Action 2 adds new functionality to QualidataNet. To support anonymisation, QualidataNet is extended though an existing tool and related user support. The tool will be maintained developed further in line with user requirements (e.g., implementation of semi-automatic marking of certain entities) (Q2/28). Prior to new versions, the anonymisation concepts and replacement schemes developed by RDCs and the communities will be made accessible via QualidataNet (Q4/26), allowing researchers to apply and develop existing replacement-concepts that fit their thematic focus. A business model will also be developed for this tool (Q4/27). Action 3 carries out a needs analysis. It includes 1) developing an overview of workflows, problems and efforts in the preparation, archiving, curation and provision of qualitative research data through workshops (Q1/27); 2) members of qualitative RDCs and selected data producers in different user communities will be in-terviewed and possible services will be discussed in focus groups; 3) the data is analysed, needs, requirements, and user preferences are identified, published, and discussed with researchers and infrastructure providers (Q3/27).

**Resources:** To deliver the results described above, KonsortSWD requests 29.4 person months for Postdoctoral researchers and comparable, 25.5 person months for Doctoral researchers and com-parable plus direct costs for expenses (cf. 3.4.6).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **M4-9: StanDem Profile: StanDem** | | | | | Category: Support/Consulting | | | | | |
| DFG Key Points | **A** | **B** | **C** | **D** | | **E** | **F** | **G** | **H** |
| **Target Group:** Researchers, RDCs | | | | | **Section: -** | | | | | |
| **DFG KPI: -** | | | | | **Partners: GESIS** | | | | | |
| **Prop. KPI:** Number of projects supported | | | | | **Basic Service: -** | | | | | |
| **Internal Collab.:** M3-4, M3-5, M4-3, “AG Demographische Standards” through M1-1, FDI Committee | | | | | **Status:** Operation | | | | | |
| **Consortia:** NFDI4Health | | | | | **External Collab.:** ZIS by GESIS | | | | | |

solutions to the documentation. An annual StanDem roundtable meeting will be organised, following the positive experiences from the first funding period. The measure will continue to bring interested parties from KonsortSWD and relevant other NFDI consortia (e.g., NFDI4Health) together to discuss challenges, present solutions and identify gaps in the existing documentation of sociodemographic standard items and variables. Challenges and gaps will be documented, and funding for potential extensions may come from KonsortSWD’s flex funds. Identified needs for new standard questionnaire items will be communicated to the “AG Demographische Standards”. A business model will be developed in the first two years of the project as well (Q4/27). At the end of the funding period, an updated and extended version of the existing documentation (Schneider et al., 2023) will be published, including tightened specifications using lessons learned during the second funding phase (Q4/28). An indirect output will be further datasets with sociodemographic standard variables published by RDCs.

**Resources:** To deliver the results described above, KonsortSWD requests 19.5 person months for Postdoctoral researchers and comparable plus direct costs for expenses (cf. 3.4.6).