RatSWD.

German Data Forum

German Data Forum's Position Paper on the Data Strategy of the German Federal Government

The German Data Forum (RatSWD) welcomes the federal government's initiative for a national data strategy. As an institutionalised forum for facilitating exchange between researchers and data producers, the German Data Forum (RatSWD) is looking forward to contributing to the upcoming process.

The **era of digital transformation opens up great opportunities and potential for innovation** by creating new data sources that can be utilised by civil society actors, associations, businesses, science and research, and public administration: in addition to economic opportunities, new methods of collecting and using data bear an immense potential for scientific research to tackle the challenges society faces today. Scientific use of data improves the information base for society and policy. It allows effective and efficient policy evaluation and thus the targeted use of public funds. Consequently, the federal government should carve out an active role for science and research in its data strategy from the onset.

Already for many years, the German Data Forum (RatSWD) has been committed to an infrastructure that facilitates flexible data access for science and research in accordance with data protection regulation. We have made important progress in the past years. Nevertheless, we see four areas with a concrete need for action:

1) Creating data access paths

The German Data Forum (RatSWD) has been working to improve **access to official microdata for independent researchers** for many years. Most recently, it championed the recently passed **Digital Healthcare Act** and, with it, the establishment of a research data centre (RDC) for care data from statutory health insurance providers. The German Data Forum (RatSWD) also demanded the provision of official statistics data on education and migration.[1] It routinely pays close attention to the specifics of anonymisation and pseudonymisation of sensitive, personal microdata as well as their long-term preservation and replicability.

Crucial existing information sources for scientific research continue to be inaccessible or accessible only to a limited extent for researchers. The area of administrative data lacks an **integrated data infrastructure** with many data lying dormant in data silos. Consequently, this data can neither be used by the administration itself nor by science and research.

[1] For a summary of RatSWD activities, see: https://www.ratswd.de/en/info/topic-access-research-data

Some areas of government action are insufficiently covered by specialised registers and statistics. Where they exist, they often do not provide clearly regulated access to researchers. This is particularly true for **the population registers, the Central Register of Foreigners, the register of education, the national mortality register, tax data, transaction data in real estate, as well as criminal and legal data**. A lack of data provision hampers data-driven research in Germany; data from Germany are also of only limited use for international comparisons. This applies, for example, to the 'EU Statistics on Income and Living Conditions' (EU-SILC panel data) and the crime-related surveys of the United Nations.

The RatSWD-accredited RDCs at official statistics agencies, e.g., the RDCs of the Federal Statistical Offices and those of the *Länder*, show that state actors are capable of continually providing sensitive microdata and **making the state a trailblazer in data provision**. The **RDC model** could be a way for many other public agencies to provide official microdata to researchers in strict accordance with data protection regulation. The success story of the network of now 34 RatSWD-accredited RDCs, which are currently facilitating access to 3,940 datasets,[2] can serve as a role model for others. This is also true for the model used by the Federal Ministry for Education and Research (BMBF) to provide project-based start-up funding for expediting the establishment of RDCs.

Important next steps include the establishment of (a) an RDC at the Federal Criminal Police Office, which makes the police crime statistics accessible to researchers, and (b) an RDC at the Federal Ministry of Finance to take care of the central preparation of tax data from the finance agencies of the *Länder*.[3] RDCs require sufficient human resources and technical infrastructures to effectively perform their tasks. The outstanding work done by the Federal Statistical Office should (c) be strengthened by providing it with more human and financial resources and by explicitly making research and service provision one of its tasks by law. This would enable it to independently generate methodological and technological innovation in the field of data provision.

Official statistics providers are also very limited by the fact that, compared to other countries, Germany lags behind in developing a **register-based population census**. Improving identity management, as part of the modernisation of population registers in general, is particularly important for making their data more re-usable and for enabling data linking. Science and research can provide important input for the planning and implementation of a register-based

[2] The German Data Forum's recent Activities Report gives an overview of the progress this infrastructure has made: RatSWD [German Data Forum] (2019): Activities Report 2018 of the Research Data Centres (RDCs) accredited by the German Data Forum (RatSWD). Berlin, German Data Forum (RatSWD). https://doi.org/10.17620/02671.44. For a more comprehensive overview, see: Bug, Mathias; Liebig, Stefan; Oellers, Claudia; Riphahn, Regina T. (2018): Operative und strategische Elemente einer leistungsfähigen Forschungsdateninfrastruktur in den Sozial- und Wirtschaftswissenschaften. Under Debate; In: Journal of Economics and Statistics 2018; 238(6): 571–590; De Gruyter; Oldenburg; https://doi.org/10.1515/jbnst-2018-0029.

[3] This would require an obligation for the finance agencies of the Länder to forward information to federal agencies that goes beyond those regulated in § 21 Abs. 6 FVG. The legal basis for linking with other research data centres could be included in tax statistics law.

population census and should be closely involved in the process. Making these census data available to science and research should be planned early on in the process.

Another issue that remains generally unresolved is how to organise access to **'commercial' data** – e.g., data from 'new economy' companies – for science and research. To date, access depends on making use of personal contacts. Here, it is important to explore how to facilitate regulated access to these data – without jeopardising the data owner's commercial interests. Solutions could include the **establishment of a data centre for digital company data and/or a trust agency**. The German Data Forum (RatSWD) has already developed a framework for a big data trust agency.[4] This framework consists of ideas as well as possible model solutions for developing the concept of data trusteeship that the federal government is planning.

2) Technical obstacles: cloud-based and remote access solutions

Digitisation is creating vast amounts of new data. Making these analysable requires European-level **cloud solutions**. GAIA-X and the European Open Science Cloud (EOSC) are steps in the right direction, which the federal government's data strategy should draw upon. Based on these initiatives, data providers such as the statistical offices, the German Pension Insurance, the Federal Employment Agency, Bundesbank, and many others could safeguard the storage of their data and secure their usability.

The technical possibilities for accessing official data in Germany, although already accessible in principle, lag behind the standards of many other EU member states and limit the data's use and potential for innovation: in most cases, it is only possible to access data on site or via controlled data processing. These access paths strain the human and organisational resources of data users and data providers. The German Data Forum (RatSWD) is therefore committed to creating flexible data access paths. Enabling off-site access even to sensitive microdata, so-called **remote access solutions**,[5] are the most promising option. Here, role models include the national statistical institutes of the Scandinavian states and other European countries, which have created the legal and technical requirements for this early on. The federal government's data strategy should also lay down the tracks to facilitate this. Remote access to official statistics data will require changing laws. Specifically, the RatSWD is pushing for changing §16 Abs. 6 of the Federal Statistical Act to make possible remote desktop access to formally anonymised official statistics microdata. The planning of timely pilot projects to bring about these changes should closely involve science and research.

[4] RatSWD [Rat für Sozial- und Wirtschaftsdaten] (2019): Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement. RatSWD Output 4 (6). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.39.

[5] RatSWD [Rat für Sozial- und Wirtschaftsdaten] (2019): Remote Access zu Daten der amtlichen Statistik und der Sozialversicherungsträger. RatSWD Output 5 (6). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.42.

3) Securing data quality

Science, research and administration need data of high quality and at a highly level of disaggregation and detail. Such data are a precondition for attaining robust scientific knowledge. The way we gather data, however, has changed in recent years. Increasingly, research data are collected using **new information technologies,**[6] **big data,**[7] or **online surveys**. These data bear immense potential for responding to existing and novel research questions. However, their quality does not always fulfil scientific standards. Unregulated use and dissemination of such data may not only jeopardise people's trust in science and research but may also be damaging to society and policy.

The federal government's data strategy should pay attention to the methodical and technical quality and, with it, the scientific value of these new data.[8] The introduction of, e.g., a data quality seal could be considered. The relevance of data quality could be strengthened by the targeted promotion of research on data quality.

In addition to safeguarding the quality of automated data collection, the federal government's strategy should make sure that administrative data conform to established quality standards. In its efforts to improve **Open Data Initiatives**, the federal government should invoke the pragmatic principles of the **FAIR criteria**,[9] which are internationally recognised principles for creating sustainable research data infrastructures. Administrative data should not only be provided in a machine-readable format but also be made findable by using **uniform metadata**.[10] Improving the data provided by GovData, launching the Competence Centre Open Data, and supporting municipal administrations by providing, for example, data-related training and resources (and developing curricula for digital education) are further measures that should be implemented soon.

4) Countering the culture of distrust and enabling data linkages

In principle, science and research pursue a quest for knowledge that serves to generate an added value for society. It does this by analysing group-based correlations, patterns, and social regularities. The aim of any scientific endeavour is never to re-identify individuals.

---

6 The German Data Forum (RatSWD) will soon publish a special handout on this issue.

7 RatSWD [Rat für Sozial- und Wirtschaftsdaten] (2019): Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement. RatSWD Output 4 (6). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.39.

8 See also: RfII [Rat für Informationsinfrastrukturen] (2019): Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel. Göttingen, Rat für Informationsinfrastrukturen. http://www.rfii.de/?p=4043.

9 The 'FAIR Data Principles' are a set of principles that sustainably reusable research data must adhere to and that research data infrastructures should implement as part of the services they provide. According to the FAIR principles, data should be '**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable.' In this context, issues of long-term preservation and replicability of research data are a necessary condition.

10 This applies to data such as official noise maps, which must be made public based on European law. In Germany, however, they are published in an array of formats based on various licenses etc., which complicates their re-use.

Scientific knowledge enhances administration and policy, especially where it uncovers a need for policy action and identifies possible improvements.

Limiting data access and data use makes it very hard to conduct data-driven scientific research. Particularly in Germany, there is a culture of distrust towards research with respect to data protection. In addition to limiting data access,[11] the fear of data misuse has generally made the conditions for **linking datasets so restrictive** that innovative research is nearly impossible or only possible by conducting expensive and time-consuming double data collections. The Federal Statistical Act, for example, restricts the project-independent combination of company data from the Federal Statistical Office with microdata from the Institute for Employment Research of the German Federal Employment Agency. This situation is obstructing and, in some cases, preventing a response to research questions that are relevant to the general public.

The German Data Forum (RatSWD) has been addressing this data protection distrust for many years, for example, by establishing itself as a point of contact and initiator for **establishing and consolidating standards in research ethics and data protection** and pushing for improving data competencies among both data users and data producers.[12] In addition to calling on government agencies and commercial actors to share their data, it also appeals to scientists and researchers to facilitate re-use by sharing their data in repositories.

The federal government's data strategy should create the conditions to foster a change in data culture towards more data sharing. Concrete measures may include **funding repositories** as well as **expanding the administrative and legal basis for linking data for research purposes**. Introducing the **principle of research secrecy** for independent researchers could also help overcome some of the current barriers. Extending privileges for researchers would also help prevent double surveys and thus satisfy the principle of data minimisation (*Datensparsamkeit*) enshrined in German data protection law. Lastly, data linking can save time and resources on the data user as well as the data producer side and preserve the integrity of data collections.

This envisioned change in Germany's data culture should build on European developments. Based on the European General Data Protection Regulation (EU-GDPR), for example, Finland has proposed the cross-national harmonisation of data access for researchers. This would result in a network of countries enabling transnational data access.[13] In the long run, a **European Data Area** could regulate researchers' access to public and, where possible,

[11] The panel data of the European Union Statistics on Income and Living Conditions (EU-SILC), for example, cannot be accessed by German researchers for data protection reasons. For the same reasons, the provision of the data of the Microcensus in the scientific use file format is considerably delayed.

[12] This includes Handreichung Datenschutz (RatSWD [Rat für Sozial- und Wirtschaftsdaten] (2017): Handreichung Datenschutz. RatSWD Output 5 (5). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.6) and material on research (data) ethics https://www.ratswd.de/en/topics/research-ethics).

[13] This would enable researchers from, say, a Danish university to work at their desks not only with Danish but also with Norwegian, Swedish, and Finnish official statistics data.

privately owned data and thus contribute to strengthening European research networks and international comparative research. This would unlock new potential for value creation, innovation and evidence-based policy making.

In sum, the German Data Forum (RatSWD) welcomes the federal government's initiative for a data strategy. It will actively **contribute to its success and give its full support** on issues such as creating legal frameworks, identifying data sources, securing data quality, designing innovative data access paths, anonymising and pseudonymising data as well as the technical design and development of the necessary infrastructures. The nascent national research data infrastructure project Nationale Forschungsdateninfrastrukur (NFDI) could serve as an umbrella for these activities.

We look forward to working on this together.

Established in 2004, the **German Data Forum (Rat für Sozial- und Wirtschaftsdaten, RatSWD)** is an independent council. It advises the German federal government and the federal states (Länder) in matters concerning the research data infrastructure for the empirical social, behavioural, and economic sciences. The German Data Forum (RatSWD) has 16 members. Membership consists of eight elected representatives of the social, behavioural, and economic sciences and eight appointed representatives of Germany's most important data producers. The German Data Forum (RatSWD) offers a forum for dialogue between researchers and data producers, who jointly issue recommendations and position papers. The council furthers the development of a research infrastructure that provides researchers with flexible and secure access to a broad range of data. The German Data Forum (RatSWD) has accredited 34 research data centres and fosters their interaction and collaboration.