SSOAR

Open Access Repository
www.ssoar.info

Activities Report 2020 of the Research Data Centres (RDCs) accredited by the German Data Forum (RatSWD)

Veröffentlichungsversion / Published Version
Tätigkeitsbericht, Jahresbericht / annual report

Funded by the German Research Foundation (DFG) - Project number: 442494171

Empfohlene Zitierung / Suggested Citation:
Rat für Sozial- und Wirtschaftsdaten (RatSWD). (2022). *Activities Report 2020 of the Research Data Centres (RDCs) accredited by the German Data Forum (RatSWD).* Berlin. https://doi.org/10.17620/02671.69

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by/4.0/deed.de

Terms of use:
This document is made available under a CC BY Licence (Attribution). For more Information see:
https://creativecommons.org/licenses/by/4.0

gesis
Leibniz-Institut
für Sozialwissenschaften

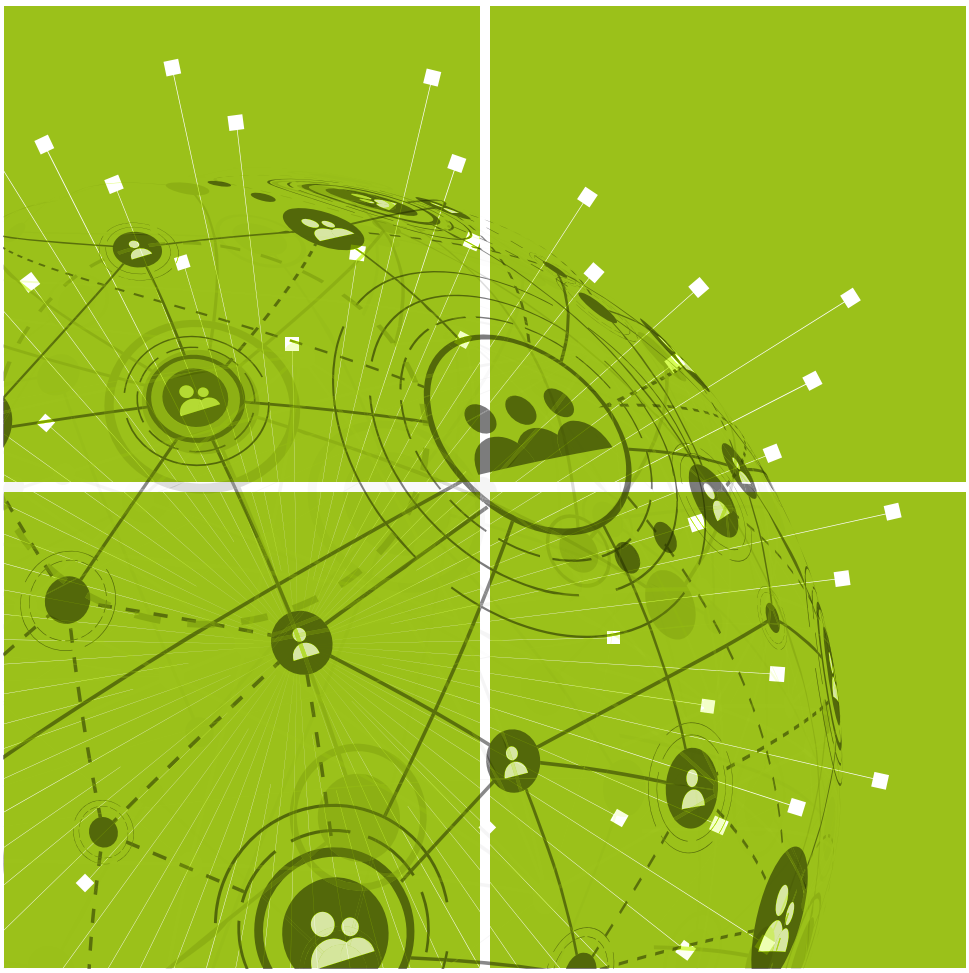
Mitglied der
Leibniz-Gemeinschaft

RatSWD.

Rat für Sozial- und
WirtschaftsDaten

Activities Report 2020

of the Research Data Centres (RDCs) accredited by the German Data Forum (RatSWD)



German Data Forum (RatSWD)

Activities Report 2020

of the Research Data Centres (RDCs) accredited by the German Data Forum (RatSWD)

Contents

Preface by the chairs of the FDI Committee 6

1 Overview of the research data infrastructure of the German Data Forum (RatSWD) 7
Current key figures 7
The starting point for establishing the research data centre network 8
Evolution and consolidation of the research data infrastructure 8
Accreditation process and quality assurance 9
New accreditations 10
Integrating the evolved infrastructure into the National Research Data Infrastructure 13
Challenges during the COVID-19 pandemic 13

2 Structure of the research data centres (RDCs) 14
Staff 14
Cooperation and research activities 16
Academic publications 17

3 Rotating topic: Securing data quality 18

4 Available data and data use 21
Range of available data 21
Time of availability of datasets and fees 21
Research output based on RDC research data 23
Scope of data use 24

5 Established data access paths and service concepts 26
Advertising the data offering 26
Access paths in data provision 27
Processing time after a signed agreement was received 29
Provision of tools 30
Services for data users 31
Service quality assurance 33

6 Current development of the research data infrastructure in the social, behavioural, and economic sciences 34
Internationalisation 34
Innovation and improvement of the research data infrastructure 35
Further developing the research data infrastructure 36

7 Special topic: Metadata standards 37
Findability 38
Accessability 40
Interoperability 42
Re-usability 44
CONCLUSION: FAIR (meta) data? 47

8 Complaints management 48
Current complaints procedures in the 2020 reporting year 48

Appendix

Appendix A: Development of the RatSWD's research data infrastructure and RDCs 50
Appendix B: Index and categories of data of the research data infrastructure oft he RatSWD 54
Appendix C: The monitoring commission 60
Appendix D: Contributors to the 2020 Activities Report 61

Preface by the chairs of the FDI Committee

The Activities Report of 2020 underscores how successful the research data centres (RDCs) have been in tackling the challenges of the pandemic. The key figures on data products that were created and the orders of these data show that the core business continued uninterrupted at most RDCs. It also highlights the benefits of the far-reaching digitisation of services, on the one hand, and a highly motivated staff on the other, who quickly adapted to new communication paths and formats. The many publications put out by RDC staff, too, are evidence of the creativity and motivation to be found at the RDCs.

By adding five new RDCs, the range of data offered by the RDCs was expanded in 2020. On the one hand, new data sources for social and economic research facilitate improved explanatory models for future science. Among these were pharmaceutical data from health insurance accounts provided by the German Pharmacoepidemiological Research Database (GePaRD) and data on settlement structure and land use from the Monitor of Settlement and Open Space Development (IOER Monitor). On the other hand, the RDC landscape was complemented by service facilities that are committed to making available existing data from research projects. Both the Research Data Centre of the German Centre for Integration and Migration Research (DeZIM) and the Research Data Centre for audio-visual data of qualitative social research (RDC-aviDa) make data accessible for secondary use and therefore increase the value of already existing data troves. Integrating European datasets to foster research that extends beyond national borders continues to be a challenge. The Research Data Center of the Leibniz Institute for Financial Research SAFE will facilitate access to harmonised European data in the field of economics in future. Beyond the 2020 reporting year, the expansion of the data offering was continued by adding two more RDCs in 2021: The Research Data Centre of the Federal Institute for Occupational Safety and Health (RDC-BAuA) facilitates access to new survey data in the fields of work and health; the main aim of the Research Data Centre of the Federal Office for Migration and Refugees (BAMF-RDC) is to make available data products from the Central Register of Foreign Nationals.[1]

We welcome all these new RDCs to a dynamic work environment that thrives with continuous exchange, one in which we all learn from cooperating and get inspired to create new products and services for researchers.

Dr. Daniel Fuß
Chair of the FDI Committee

Tatjana Mika
Chair of the FDI Committee

1 The RDCs GePaRD, aviDa, SAFE, and BAMF-FDZ received preliminary accreditation because they have not yet taken up operation.

1 Overview of the research data infrastructure of the German Data Forum (RatSWD)

The FDI Committee oversees a dynamic and decentral network of 39 research data centres (RDCs) that are accredited by the German Data Forum (RatSWD), five of which were newly accredited in 2020 (as of December 2020).[2]

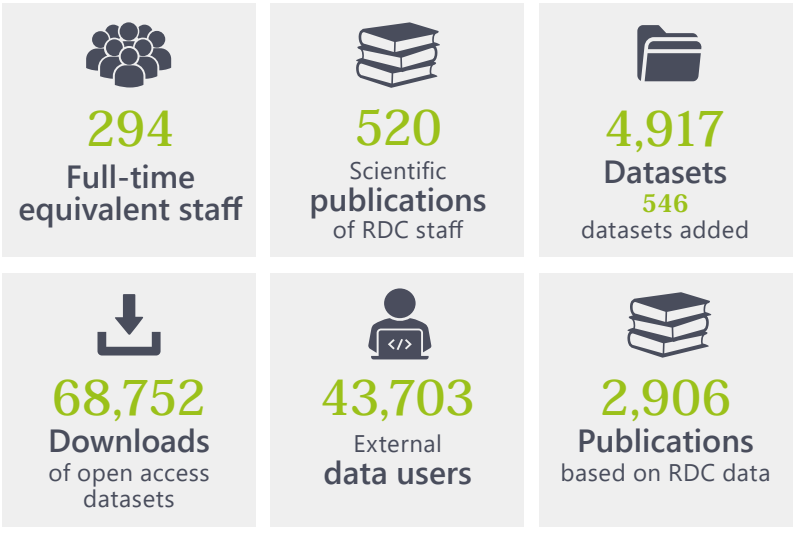
The RDCs archive data and make them available to researchers in accordance with data protection regulation and by using several access paths. The spectrum comprises survey data from various disciplines of the social, behavioural, and economic sciences as well as geographic and spatial data, financial data, statistical data from the federal and state level as well as register and social security data. Meanwhile, access includes not only quantitative but also qualitative data. The network is continuously expanded by including more RDCs and thus also enhanced with additional survey methods, data types, and data formats. The steady increase of RDCs and increasing usage figures show that the RDC model is future-proof and facilitates empirical research and that RatSWD accreditation has established itself as a seal of quality among research funders and data users.

Current key figures

The key figures for 2020 underscore this successful development of the RDC landscape: On the cut-off date for this publication (31/12/2020), the RDCs employed a total of 294 staff in full-time equivalents (FTEs), about two thirds of which were academic staff. This staff put out 520 scientific publications in 2020. This is a success in many ways: (Re-)using in-house research data creates closer ties to the research community, strengthens the support competencies of RDC staff, contributes to quality assurance, and makes the datasets better known. At least 2,906 scientific publications are fully or partly based on the 4,917 datasets made available by the RDCs.

In 2020, 546 additional datasets were made available to data users – a significant increase on the previous year.

The potential for re-use of research data is growing. Data users made good use of the broad array of data available at RDCs in 2020: RDCs reported 43,703 data users and 68,752 free downloads.



2 Find an alphabetical list of RDCs and information of the data offered by the various RDCs (Categories: Social, Economy, Education, Health, Psychology, Qualitative, Other) in Appendix B.

The starting point for establishing the research data centre network[3]

The Commission on Improving the Informational Infrastructure (KVI) was established in 1999 as a response to initiatives from within the scientific community.[4] The committee presented a comprehensive report in March 2001. One of its key recommendations was to set up RDCs at major public data producers, including the federal and state-level statistical offices, the German Pension Insurance, and the Federal Employment Agency, with the aim of professionally archiving existing research data, making sure that they could be used for replication studies, to the extent possible, and to respond to new research questions. The Founding Committee leading to the RatSWD was set up that same year. With that, the cornerstone of today's RDC network around the RatSWD was created.

Info box 1

Tasks and structures of the RatSWD

Established in 2004, the German Data Forum (RatSWD) is an independent council advising the German federal government and the governments of the Länder in matters concerning the research data infrastructure for the empirical social, behavioural, and economic sciences. As an institutionalised forum for dialogue, it facilitates a continuous exchange between data producers and data users in science and research with the aim of improving access to high-quality and scientifically potent data. These data are supplied by public, private and scientific actors. Working together in the RatSWD (as of the 7th appointment period, 2020) are ten representatives of the social, behavioural, and economic sciences, legitimised by election, and ten representatives of the most important data producers. This mode of equal representation ensures the broad range and depth of expertise on the committee. The RatSWD plays a key role in developing research infrastructures in the social, behavioural, and economic sciences and is committed to creating research-friendly legal and political conditions.

Evolution and consolidation of the research data infrastructure

The research data infrastructure has been continuously developing since 2001. This included the founding of new RDCs, while the reasons for founding them were multifarious: Some sought to implement the recommendations issued by commissions like the KVI, the German Council of Science and Humanities, or scientific advisory groups. Other RDCs were commissioned by their parent institutions with the aim of promoting research. What they had in common was the aim to expand and strengthen the research data infrastructure in Germany by improving access to research data for the scientific community.

The RatSWD was founded in 2004 as a strategic committee. To promote a productive dialogue between the RDCs, the RatSWD set up the "Committee for Data Access" (FDI Committee). The main task of this committee is to continuously secure and improve the research data infrastructure: This included expanding the quality and quantity of the data offerings as well as developing and facilitating data access for the research community.

.

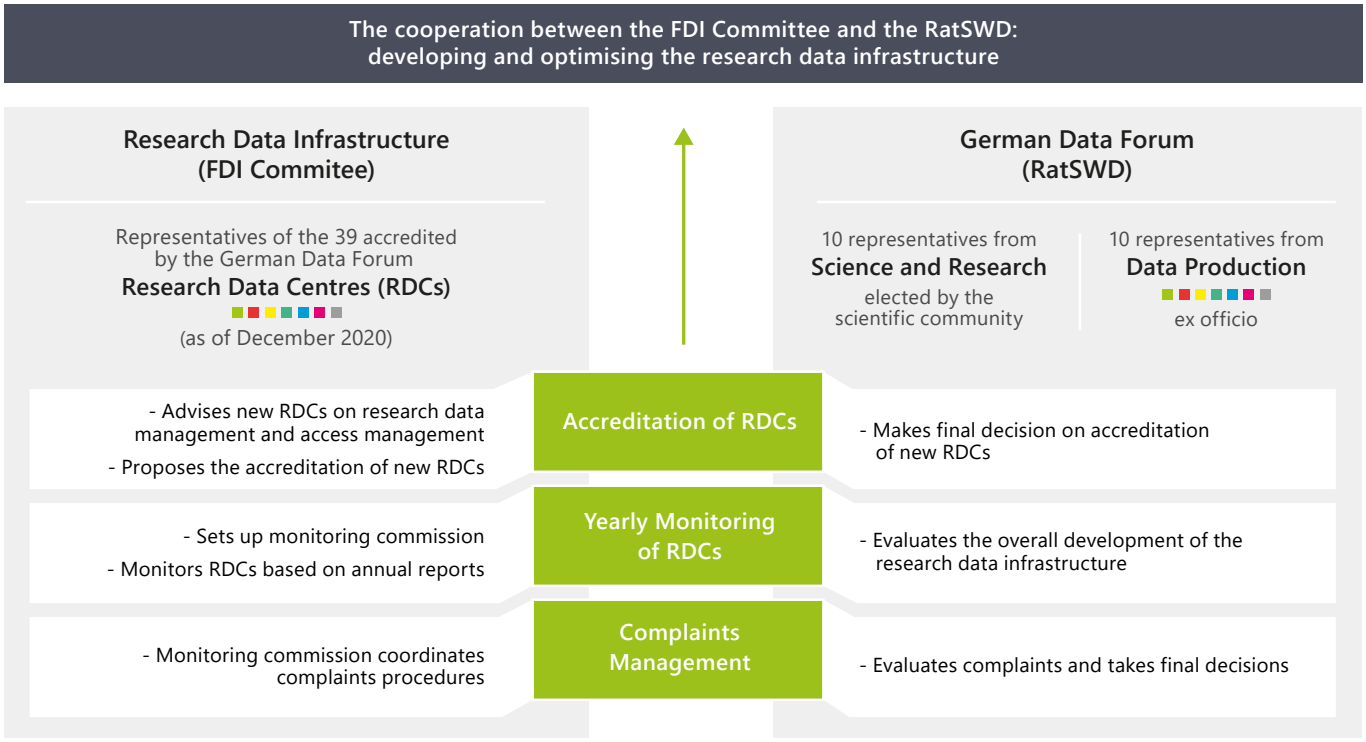
3 See Appendix A for a chronological overview of the development and the services of the RatSWD research data infrastructure.

4 Zapf, W. et al. (1996): *Memorandum zur Verbesserung der Zugangsmöglichkeiten zu Mikrodaten der amtlichen Statistik*. In: ZUMA-Nachrichten, 39, 172–175. / Hauser, R.; Wagner, G. G. & Zimmermann, K. F. (1998): Erfolgsbedingungen empirischer Wirtschaftsforschung und empirisch gestützter wirtschafts- und sozialpolitischer Beratung. In: Allgemeines Statistisches Archiv, 82, 369–379.

Accreditation process and quality assurance

To ensure the quality of the research data infrastructure, the RatSWD defined minimum standards and accreditation criteria in 2010.[5] Since 2015, they have been continuously adjusted to keep up with technological innovation and the methodological advances of the network's portfolio. Annual monitoring ensures adherence to standards and high service quality. In addition to annual monitoring, the monitoring commission, which is elected by the members of the FDI Committee, oversees the inspection of RDC accreditation documents.[6] The final decision on accreditation is made by the RatSWD. In addition, the RDCs take part in a complaints management system, which is managed by the RatSWD office and overseen by the monitoring commission (see also Chapter 8), covering cases where data provision problems cannot be solved bilaterally.

Fig. 1: Cooperation between the FDI Committee und the RatSWD



© RatSWD 2022

The results of the monitoring process are compiled in an internal as well as the publicly available Activities Report of the RatSWD-accredited RDCs at hand.[7] The RatSWD's accreditation is a seal of quality for the RDCs because it requires compliance with mandatory criteria: RDCs must have at least one data access path, must adequately document their datasets, and present a concept for ensuring long-term data availability. Moreover, accreditation depends on available information on tools and other material, quality assurance, the further development of the infrastructure, and adherence to data protection regulation.[8] Accreditation benefits RDCs in a variety of ways: They receive extensive support as well as information about best practice solutions to help guide the ongoing development of their own infrastructures and participate in an exchange of knowledge and experiences with other RDCs.

5 RatSWD [Rat für Sozial- und Wirtschaftsdaten] (2017). *Qualitätssicherung der vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) akkreditierten Forschungsdatenzentren (FDZ).* RatSWD Output 8(5). https://doi.org/10.17620/02671.4.

6 For background information and a list of members of the monitoring commission, see Appendix C.

7 The Activities Reports since 2015 are available at: https://www.konsortswd.de/aktuelles/publikationen/taetigkeitsberichte (Last retrieved: 31/01/2022).

8 Consult the accreditation criteria of the RatSWD [Rat für Sozial- und Wirtschaftsdaten] (2017). *Qualitätssicherung der vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) akkreditierten Forschungsdatenzentren (FDZ).* RatSWD Output 8(5). https://doi.org/10.17620/02671.4.

New accreditations

Five additional RDCs were accredited in 2020. These new RDCs significantly expanded the existing network by making medical-epidemiological data (GePaRD), financial data (SAFE), spatial data (IOER Monitor), and data on migration and integration issues (DeZIM) available to researchers in the social, behavioural, and economic sciences. With RDC-aviDa, an RDC offering videographical research data has been included for the first time.[9]

German Pharmacoepidemiological Research Database (GePaRD)

The research data centre FDZ GePaRD is a pharmacoepidemiological research database with data from statutory health insurance providers in Germany. Since 2004, the Leibniz Institute for Prevention Research and Epidemiology – BIPS has been working on the establishment and maintenance of GePaRD, which can be used to investigate research questions on the utilization and safety of drugs and vaccines in routine care.

GePaRD contains accounting data from four statutory health insurance providers and information from currently 25 million people, who have been insured through one of them since 2004. In addition to demographic information, GePaRD contains information on reimbursable drug prescriptions, outpatient and inpatient care, and diagnoses. It boasts information on roughly 20 % of the general population across all geographical regions in Germany. Ranging from obtaining the data, to preparing them, and making them available, processing takes 25 months, i.e., data from 2018 are ready to use the earliest in late 2020.

BIPS does not own the data and is thus not allowed to decide for which specific projects the data can be used. The approval of projects is based on the authorisation by the health insurance providers and the respective governing authorities. Approval for data use in accordance with § 75 SGB X depends on whether the public interest significantly outweighs the right to personal data protection of the persons concerned. The process of approval by the health insurance providers and the governing authorities usually takes at least three months.

Research Data Center of the Leibniz Institute for Financial Research SAFE

Research on German and European financial markets suffers from a lack of pan-European data sets. Also, existing data sets do not provide a standard identification of, for example, companies. Therefore, researchers often utilize data from the United States where the integration of different databases is more advanced. Consequently, empirical analyses are mostly based on non-European data However, because of the institutional differences, political recommendations that result from these analyses cannot – or only in a limited scope – be transferred to the European area.

To overcome this problem, the SAFE Research Data Center not only draws on the usual international data sources but also creates new European data sets, combines existing data sets and processes them. The aim is to place the five central research areas of SAFE on a common European data footing.

9 The RDCs GePaRD, aviDa, and SAFE received preliminary accreditation because they have not yet taken up operation.

Monitor of Settlement and Open Space Development (IOER Monitor)

The IOER Monitor is a service of the Leibniz Institute for Ecological Urban and Regional Development (IOER). It provides data and information for land cover and land use structure and change for the whole of Germany, particularly regarding sustainability. The data research centre allows for accessing the data via a browser-based viewer, through geoservices and downloads.

Data is provided on an annual basis and available in high resolution on grids and administrative levels. The data sets provided are based on research results of the IOER. The metadata (data sheet) of the individual data sets give insights on used methods, calculation and used data. Among others, geotopographic data (ATKIS Basis-DLM), land cover data (LBM-DE), official building footprints (HU-DE) and house coordinates (HK-DE) as well as other geospatial data are used as input data.

The Research Data Centre provides access to the data via a map viewer with comprehensive tools and via geo services and downloads. As required, the data for Germany can be obtained by interested scientists or for specified spatial delimitations and time periods. An overview of the more than 80 data sets including calculation methods, spatial and temporal reference and the corresponding export functions via geoservices is available at https://www ioer-monitor de/en/indicators.

Research Data Centre of the German Centre for Integration and Migration Research (DeZIM)

The German Centre for Integration and Migration Research (DeZIM) is a political and scientific initiative in order to strengthen excellent and internationally visible integration and migration research in Germany. The Research Data Centre DeZIM.fdz gives researchers the opportunity to access data collected within the scope of research projects of the DeZIM institute itself and of the institutes belonging to the DeZIM research community. Besides providing access to these data, the DeZIM.fdz also offers a comprehensive information database. This database allows for research on migration and integration studies archived in the DeZIM.fdz as well as in other research data centres. Moreover, the DeZIM.fdz offers support to data users and gives advice on selected methodological issues.

Research Data Centre for audio-visual data of qualitative social research (aviDa)

RDC-aviDa's service aims at sharing videographical research data created by researchers and making them available for re-use. It is aimed at researchers in the field of qualitative empirical social research working with videography. At aviDa, too, research data from primary research are made available to third parties and for re-use in research and teaching in a web-based form.

Based on DepositOnce, a repository for research data and publications from Technical University Berlin (TU Berlin), avida was established as a prototypical research data infrastructure for long-term and community-based digital preservation and re-use of audio-visual research data in July 2018. It was set up in cooperation with the Department of General Sociology of TU Berlin, the Chair of Cultural and Religious Sociology of the University of Bayreuth, TU Berlin's University Library as well as its Central Institution for Campus Management (ZECM). In doing so, the foundation was laid for a sustainable and stable web-based platform with high usability and efficiency, which seeks to serve the (science-internal) exchange of research data for subsequent use in research and teaching.

Fig. 2: Locations and guest researcher workstation of accredited research data centres in 2020

In addition, the FDZ BA at IAB has guest researcher workstations in Europe, Canada and the USA.

Integrating the evolved infrastructure into the National Research Data Infrastructure

Equipped with funding of the National Research Data Infrastructure (NFDI), the Consortium for the Social, Behavioural, Educational and Economic Sciences (KonsortSWD) began its work in October 2020. The services developed by the consortium are to further improve the services of RDCs in future. For this reason, the RDCs are integrated into all work packages and the consortium now regularly reports to the FDI Committee on the project's progress. Its activities on professionalising research data management and on increasing the connectedness of RDC services are particularly relevant to RDCs. They receive direct support on guideline and contract design as well as the qualification and training of their staff. From the user perspective, the increased connectedness of guest researcher workstations (GWAPs) is an important measure that will help to reduce the costs for data access. Find an overview of the consortium's services, on the following website: https://www.konsortswd.de/en/konsortswd/the-consortium/services/.

KonsortSWD
Consortium for the
Social, Behavioural, Educational
and Economic Sciences

Challenges during the COVID-19 pandemic

In 2020 (and beyond), the RDCs were faced with particular challenges brought about by the COVID-19 pandemic, particularly in the areas data access, contract design, and the organisation and implementation of workshops and trainings. Regarding data access, social distancing rules most strongly affected guest researcher workstations, which had to be closed during the pandemic. Opening of workstations for guest researchers depended, for one, on regional pandemic events in 2020 and the proper measures for reducing infection, but also on decisions made by the respective institutions to protect their staff. Individual RDCs created measures to bridge these gaps, e.g., remote data access as an alternative data access path. Using scientific use files is most often limited to the scientific institutions of data users by contract and is not possible from their homes via secure remote access procedures. Here, the RDCs facilitated temporary exemptions, including new regulation on data access when working from home. In contract design, necessary obtaining of original signatures led to delays in data access. This could be mitigated, for example, by introducing certified signatures. How rapidly RDCs made changes to adapt to the new situation is also shown by the increase in virtual trainings and workshops, which previously took place face-to-face.

2 Structure of the research data centres (RDCs)

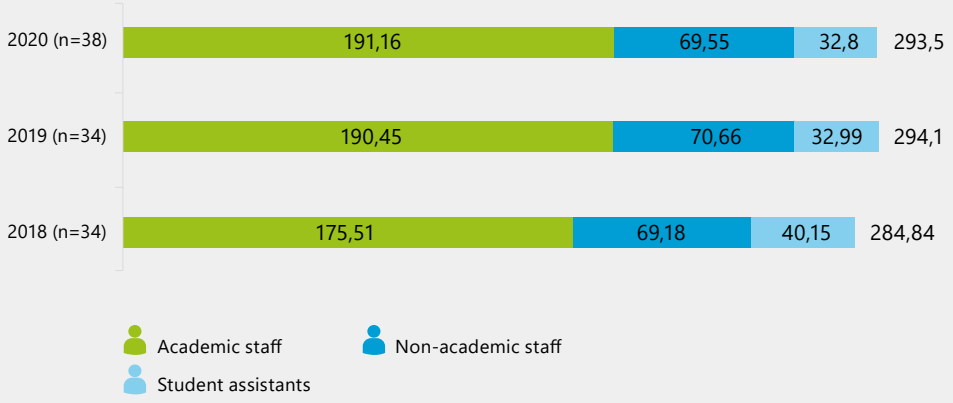
The information presented in the following chapters are gleaned from the RatSWD's annual monitoring, which was jointly developed by all RatSWD-accredited research data centres (RDCs) and which all RDCs therefore participate in. In the 2020 reporting year, 39 RDCs took part in the monitoring process. The RDCs of the federal and state statistical offices jointly responded to the questionnaire. Therefore, the following data refer to 38 responses from the RDCs.

Staff

The RDCs have been continuously expanding their staff over the past years. They have now entered a consolidation phase. Staff numbers at the RDCs have firmed up at a high level. The addition of four new RDCs in 2020 did not lead to an increase in staff. On the cut-off date (31/12/2020), the 38 RDCs employed a total of 293.5 staff in full-time equivalents (FTEs). Staff increases of previous years were mainly in the academic field. The proportion of non-academic staff and student assistants has hardly changed compared to the previous year (▸ Fig. 3). About a third of RDCs does not employ students at all.

The increases in staff of previous years have not continued.

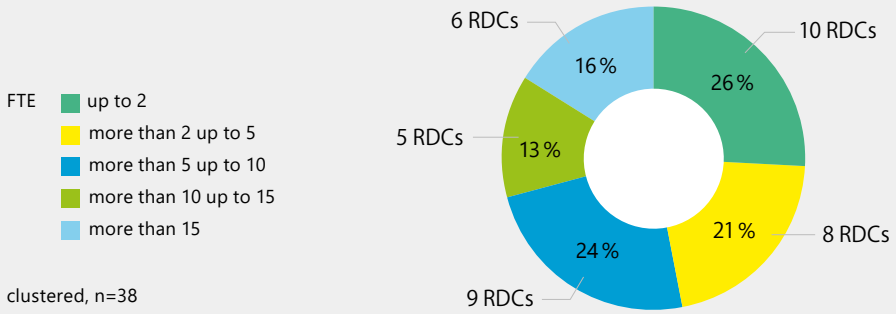
Fig. 3: RDC staff and its distribution in full-time equivalents (FTEs)



© RatSWD 2022

Staff numbers vary strongly across RDCs. They employ an average of 7.7 staff in full-time equivalents. However, they can range from less than one FTE to 34 FTEs. Smaller entities abound as large RDCs tend to be rare. Just under three quarters of RDCs have ten employees in FTEs at the most (▸ Fig. 4). If an RDC belongs to a parent institution, there is no direct correlation to the staff numbers of that institution.

Fig. 4: RDC staff in full-time equivalents (FTEs)



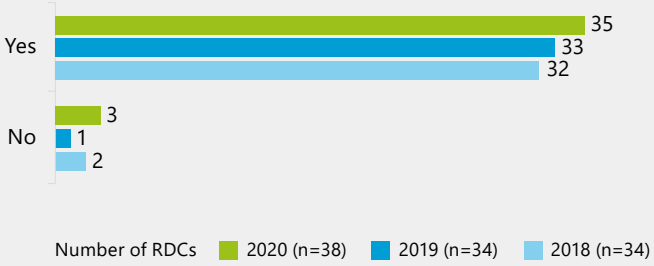
© RatSWD 2022

The staff at RDCs varies greatly. Most RDCs employ five or more FTEs.

In 35 of 38 surveyed RDCs, academic staff conduct their own research (see ▸ Fig. 5). These research activities can be contributions to the research agenda, deal with data methodologically, or focus on the technological and functional development of the data infrastructure. Data users ultimately benefit from these activities in the form of competent support based on the current state of research.

Fig. 5: Independent research by academic staff

Does your RDC's academic staff conduct their own research?



© RatSWD 2022

Academic staff conduct their own research at almost every RDC

What is the scope of these research activities? At 17 RDCs, the academic staff dedicate a fixed share of their work hours to research activities. On average, this share is just under a third, but there is a broad distribution across the RDCs. At seven RDCs, staff dedicate up to one quarter of their work hours to research. The share is higher at ten RDCs.

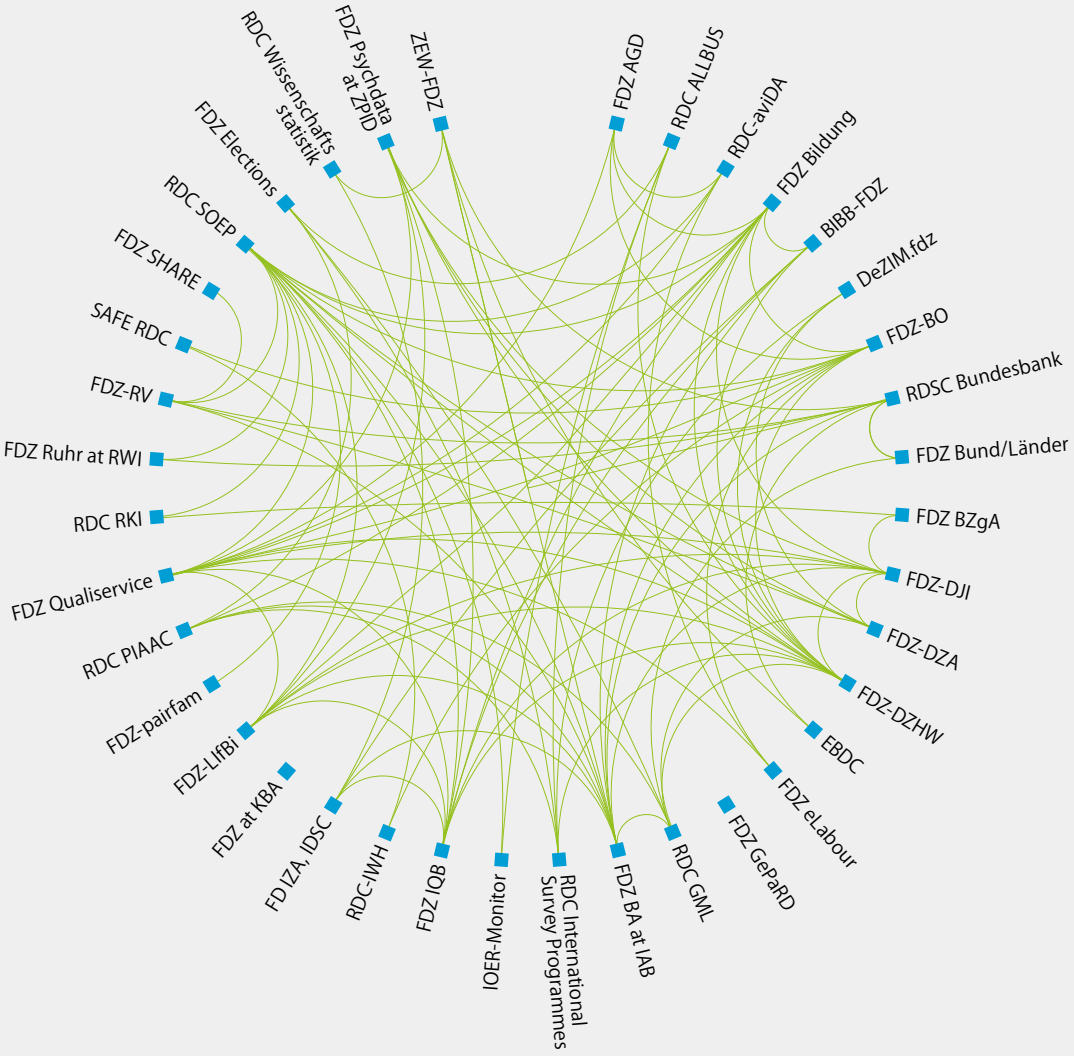
Cooperation and research activities

The number of RatSWD-accredited RDCs that maintain institutionalised cooperative relationships with other domestic RDCs has increased compared to the previous year.

Thirty-two RDCs now report maintaining such relationships, while the number of RDCs that had previously stated that they were not involved and did not plan to become involved in such institutionalised relationships has fallen to six (see Chapter 6 for international research partnerships). The first consortia, funded as part of NFDI, begun their work in this reporting year. Numerous RatSWD-accredited RDC sit on these consortia. It is likely that this has further facilitated institutionalised cooperative relationships. Lastly, KonsortSWD has significantly deepened cooperation between the RDCs. The network of RDCs forms the backbone of KonsortSWD through their experience in operating user-oriented research data infrastructures. Beyond the activities within the FDI Committee, the RDCs all collaborate with each other closely and continuously, as is shown in ▸ Fig. 6.

Cooperation within and beyond the FDI Committee has created a multi-layered, interdisciplinary network of RDCs

Fig. 6: Cooperation between accredited RDCs in 2020



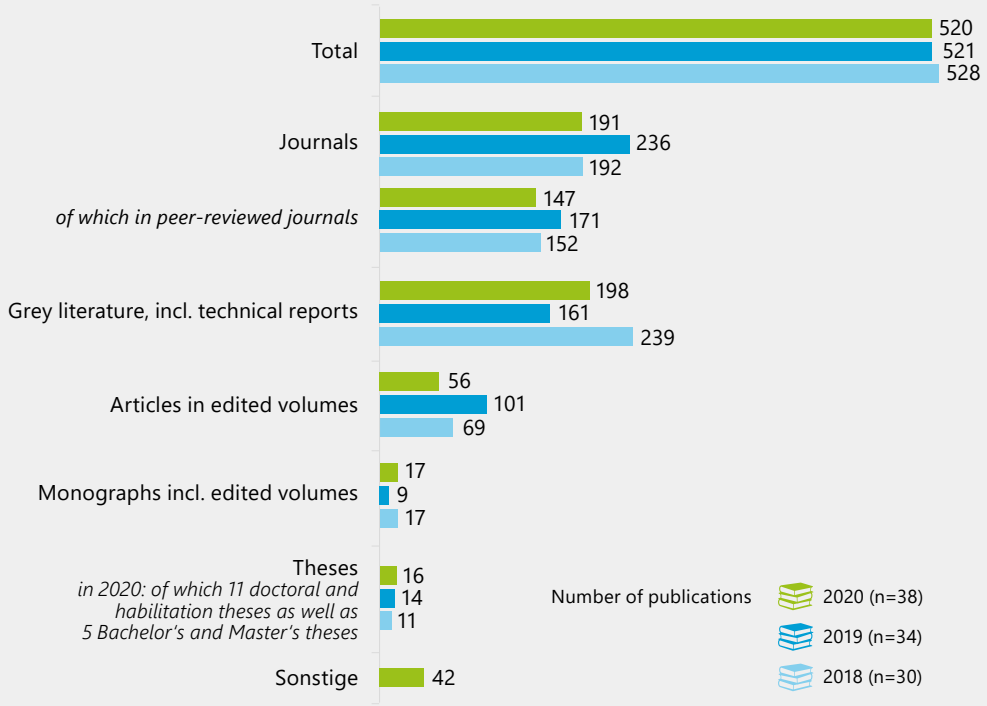
© RatSWD 2022

Academic publications

As already shown above, the majority of full-time equivalents at RDCs consists of academic staff who typically conduct their own research. The activities of RDCs are not focused solely on services and data provision. It also includes qualified support on the potential of the data they provide. This is not necessarily limited to questions of content but includes advice on possible applications and the restrictions of statistical methods. In-house research on and with the offered data helps to ensure that activities related to data, service, and support are carried out at an appropriate scientific level. Moreover, it is documented in academic publications. In total, publication output stayed roughly the same compared to the previous year, as is shown in ▸ Fig. 7. RDC staff issued 520 publications during the reporting period.

Fig. 7: Scientific publications of RDC staff

Please indicate the number of scientific publications produced by your RDC's staff, regardless of the type of data and whether the publication was prepared during RDC working hours.



© RatSWD 2022

RDC staff publishes many articles in peer-reviewed journals

478 publications can be categorised into the most important scientific publication formats. As in previous years, most contributions were published as grey literature or in journals. RDC staff were again able to publish many articles in peer-reviewed journals, which can be seen as indicative of the high quality of their content. The number of academic theses also saw a slight increase. Much like the results of previous years, these were mainly higher academic qualifications (dissertation, habilitation). Bachelor's and master's thesis also belong in this category but are not reported by the RDCs, which is why it is safe to assume substantial reporting here. Lastly, 42 publications fall under the category "Other", which includes conference proceedings and variable reports.

3 Rotating topic: Securing data quality

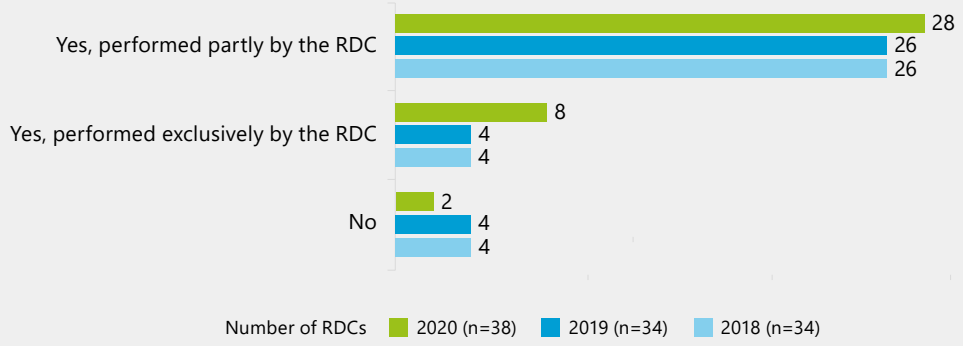
The tasks of the RDCs include the long-term securing of research data, their documentation for the re-use by third parties, securing data protection, and, at most RDCs at least, prior data quality checks. While all three areas were surveyed for the annual Activities Report in the past, the areas will be distributed across three survey years starting with the 2020 Activities Report at hand.[10] In 2020, the RDCs were surveyed about quality assurance procedures for datasets.

How involved the RDCs are in data quality assurance depends, among other things, on whether and to what extent the RDC is connected to a (parent) institution. Typically, the manner in which the RDCs are involved in data quality assurance is stable over time. Therefore, there were hardly any changes in the distribution of tasks in the past reporting years. In the 2020 reporting year, however, a shift towards a more active involvement of the RDCs in data quality assurance has been observable, which is not explained solely by the addition of newly accredited RDCs. To be sure, all five new RDCs are principally involved in quality assurance, either partially or exclusively. However, compared to the previous year, two of the already existing RDCs are now at least partially involved in data quality assurance and two others have expanded their responsibilities for this task.

Overall, this results in the following picture in the 2020 reporting year (▸ Fig. 8): Eight RDCs are fully responsible for performing data checks, while 28 RDCs and therefore over 70 % of RDCs are at least partly responsible for performing data checks. Two RDCs do not perform data checks at all.

Fig. 8: Data checks at RDCs

Is data checking (checking the quality of shared data) a task of your RDC?



© RatSWD 2022

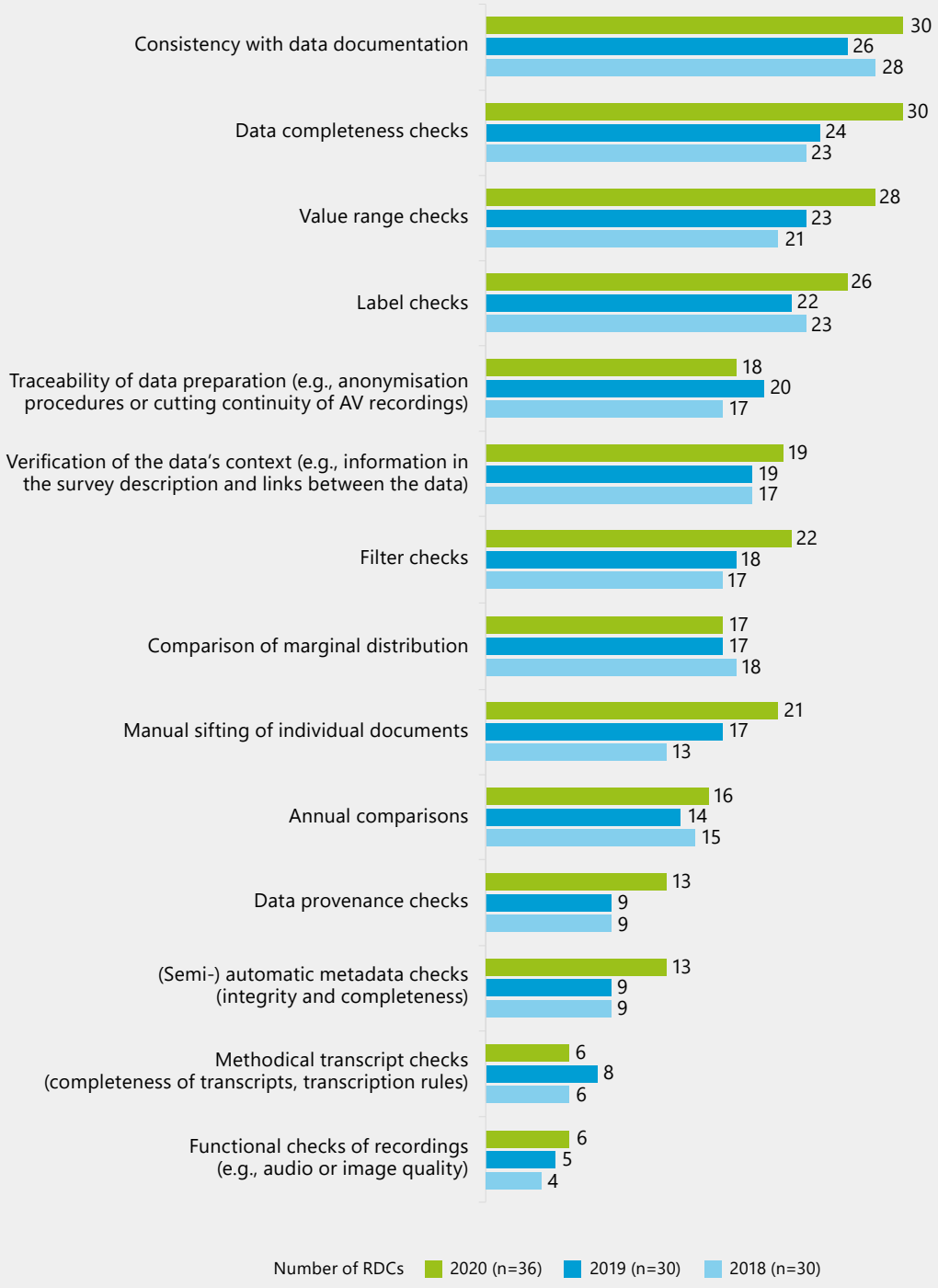
The majority of RDCs is involved in data checking

The most frequent data checks at RDCs include consistency with data documentation, data completeness, and correctness of value ranges and labels (▸ Fig. 9). The number of RDCs that report performing data provenance checks or automatic metadata checks is much lower. However, this could be because data provenance was already checked before the data were handed over to the RDCs. When certain data checks are not carried out within the RDC, it does not mean that they are not carried out at all but rather that they are anchored at another point within the data life cycle. Methodical checking of transcripts and technical checks of recordings are carried out at only few RDCs. This is mainly due to the fact that these checks are typically carried out in qualitative empirical research, which only a small portion of accredited RDCs is specialised on.

10 This new design is motivated by the fact that the RDCs have established, robust structures in the relevant work areas.

Fig. 9: Types of data checks

Which types of data checks are performed at your RDC?
(Multiple answers possible)



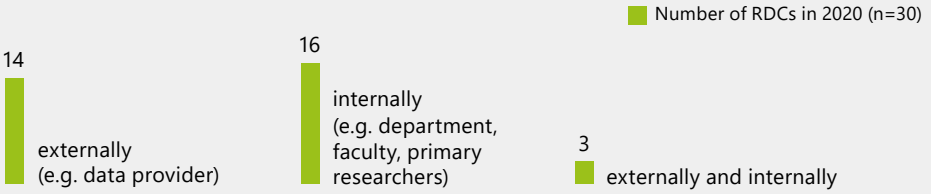
The RDCs perform appropriate data checks for each dataset; they depend particularly on the survey method.

© RatSWD 2022

RDCs who are not exclusively responsible for data checks most frequently name primary researchers and data producers (who can also be third-party institutes) as additional actors in quality assurance (▸ Fig. 10). Twenty-three RDCs have guidelines for data checks.

Fig. 10: Who (or who else) performs data checks?

(Evaluation of an open question, multiple answers possible)



© RatSWD 2022

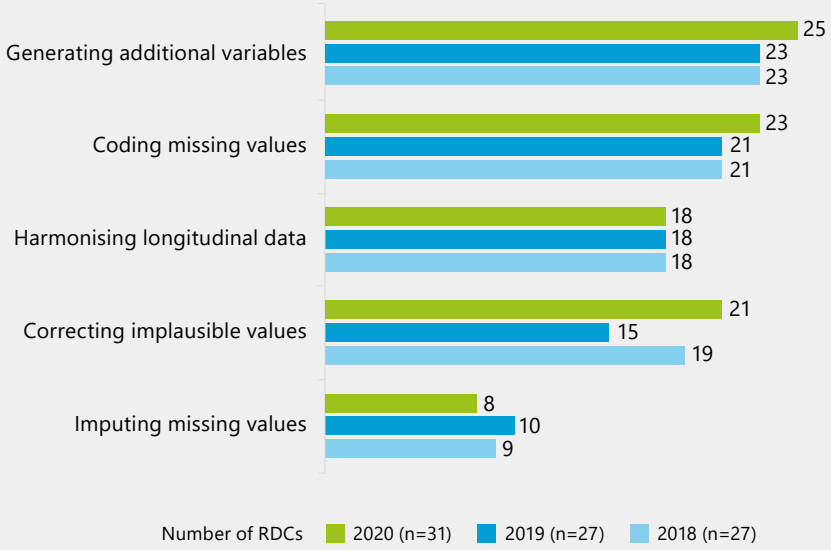
Data checks are performed both internally and externally

Data checks are an important step in data quality assurance. Another component of data quality assurance is the correction of data errors, including data generation, to increase user-friendliness, for example. The distribution of tasks is similar to that of data checks: Three RDCs are solely responsible for correcting the data; 28 do so only partially. Seven RDCs do not perform data correction measures at all.

In 2020, the three most common tasks in this area included the generation of additional variables, highlighting missing values through coding, and correcting implausible values. Harmonisation of longitudinal data is part of quality assurance measures at 18 RDCs (▸ Fig 11). There are clear guidelines for data correction at 18 out of 31 RDCs. Data corrections are not always made transparent to users at the individual level, e.g., when checks were performed for data protection or anonymisation reasons.

Fig. 11: Types of data correction at RDCs

Which types of data generation or data correction measures are performed at your RDC?
(Multiple answers possible)



© RatSWD 2022

Data corrections increase the quality and re-usability of the data

4 Available data and data use

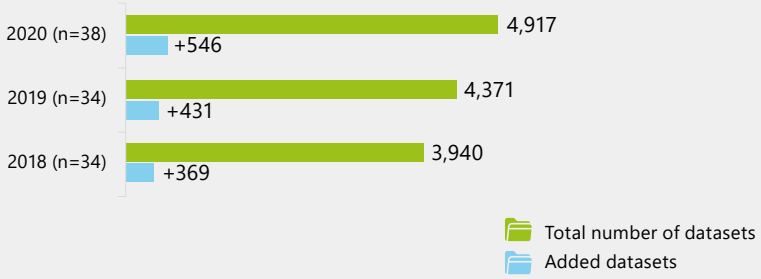
In this chapter, we will look more closely at the data made available by RDCs, data use, and research output. The last section will give a differentiated overview of how many times data were downloaded, how many data use agreements were signed, and how many new data users were added in 2020.

Range of available data

The increase in available data of previous years continues: On the cut-off date for this publication (31/12/2020), the RDCs made available 4,917 datasets. In the 2020 reporting year, the RDCs added 546 datasets, which were either assigned a digital object identifier (DOI) or were suitable for DOI registration in principle (see ▸ Fig. 12). The number of newly added datasets varies strongly across RDCs, with some adding datasets in the single digits and one RDC adding up to 130 new datasets. Since a dataset can contain several individual studies, the number of available studies is significantly higher.

Fig. 12: Datasets made available by the RDCs

Please indicate the number of new surveys and datasets that were added during the reporting year.
Datasets are regarded as new, for example, if they were assigned a DOI or were suitable for DOI registration.



© RatSWD 2022

There has been a continuous increase in available datasets in the past years

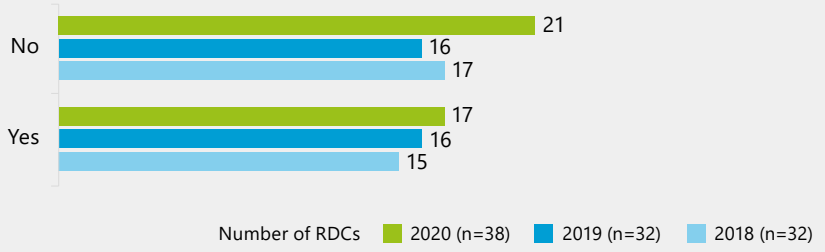
Thirty-three RDCs assign persistent identifiers (PIDs) to their available datasets, e.g., DOIs or Uniform Resource Names (URN), to ensure their long-term findability and citability. All RDCs not yet using PIDs, like DOIs, are currently planning to implement this practice.

Time of availability of datasets and fees

A central goal of RDCs is to facilitate low-threshold and timely access to data. In order to ensure equal opportunities, research data should be made available to all qualified users at the same time. High fees and long closure periods for datasets stand in the way of this goal. At the same time, however, good reasons exist for imposing a closure period on transmitting certain data. Most RDCs argue that closure periods ensure that primary researchers retain the opportunity and right to be the first to utilise the data. Furthermore, closure periods may also protect academic theses currently being written.

Fig. 13: Closure periods for datasets at RDCs

Do you have closure periods for some datasets?



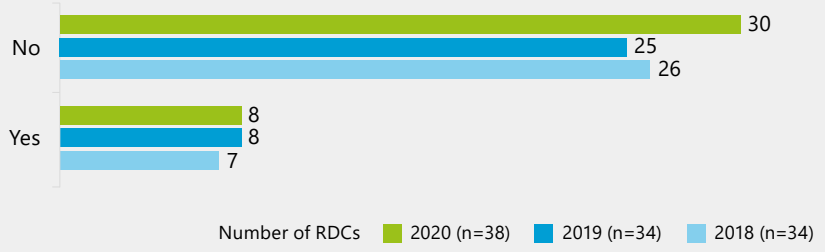
© RatSWD 2022

More than half of RDCs make datasets available without closure periods

Twenty-one out of 38 RDCs do not have closure periods, i.e., the data are made accessible immediately after they were received and prepared (▸ Fig. 13). Seventeen RDCs reported imposing closure periods on at least parts of their data offering in the 2020 reporting year. Five RDCs report fixed waiting periods ranging from six months to no more than two years. Closure periods at all other RDCs are not generally defined but hinge on certain requirements, for example, the end date of a research project, specifications made by funding organisations, or specifications by the data givers themselves. Overall, there has been a certain dynamic in the use of closure periods; the absolute number of RDCs that implement closure periods on datasets saw a slight increase due to new accreditations.

Fig. 14: Fees for data access at RDCs

Do you charge fees for data access for research purposes?



© RatSWD 2022

A large majority of RDCs does not charge fees for data provision

In addition to closure periods, user fees can be an obstacle to data use. A majority of RDCs does not charge any fees at all: Out of 38 RDCs, 30 reported that they did not charge any fees in 2020 (▸ Fig. 14). Compared to the previous year, the differences were very slight. The fees reported by eight RDCs were mostly in the two-digit or lower three-digit euro range. Many RDCs offer discounts for academic theses.[11] RDCs vary in the way they charge fees: some charge per dataset, or data access path, while some charge for data use per survey year. Some charge fees to account for the effort that goes into certain data preparation measures, such as specific anonymisation measures or additional support services. It is safe to assume that very low fees contribute little to covering the RDCs' expenses. However, in the sense of a token fee, they help to ensure that the data are used only for their designated purposes.

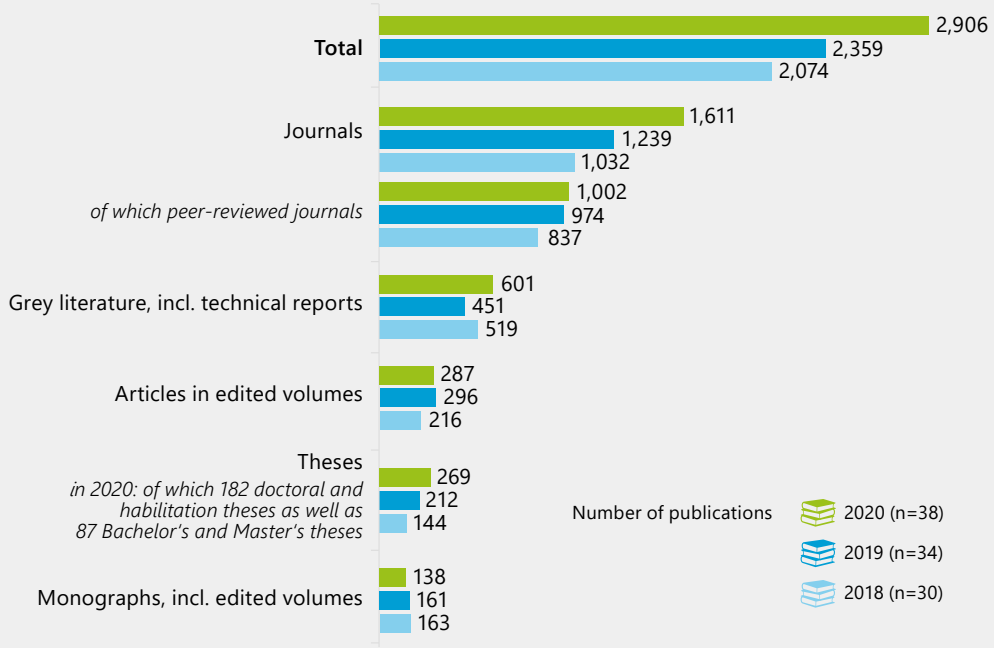
11 The Federal Statistical Office has been waiving the already reduced costs for data use by young researchers since June 2020. In the entire official statistics community, data access for bachelor's, master's and PhD theses is now free-of-charge.

Research output based on RDC research data

Use of available datasets by the research community is primarily reflected by research output. The number of publications, in turn, is a key indicator of that output. In 2020, 38 RDCs reported a total of 2,906 publications (▸ Fig. 15), based on the research data they make available. It must be noted that many researchers neglect to notify the RDCs about publications or to send a copy to them. Moreover, despite widespread use of persistent identifiers for research data by RDCs, many researchers do not yet make use of data citation in their publications. It is therefore safe to assume substantial underreporting here.

Fig. 15: Number of publications based on research data provided by the RDCs

Please indicate the number of scientific publications that appeared in the reporting year 2020 based on the research data provided.



© RatSWD 2022

There is a continuous increase in the number of publications

Even though determining the number of publications based on the data made available by RDCs is incomplete, there has been a total increase of recorded publications in long-term comparison. Articles in peer-reviewed journals continued to be the publication type most reported by data users in 2020. It is important to note the qualitative differences between publication types. Articles in peer-reviewed journals are to be viewed as superior to journal articles without peer review. The share of articles in peer-reviewed journals was particularly high in 2020.

However, the amount of grey literature including technical reports and academic theses based on RDC data has also seen a significant increase.[12] While the number of articles in edited volumes has decreased slightly, the number of monographs has declined.

12 Considerable underreporting is also to be expected for academic theses, since particularly bachelor's and master's theses are hard to find and usually only recorded when data users make an effort to do so. In addition, authors of PhD theses tend to work within larger projects and tend not to register data use directly with the RDC. The number of theses written using RDC data is likely much higher.

Scope of data use

The scope of data use is another key indicator for the relevance of RDCs in the research landscape. One central variable is the number of datasets retrieved from the RDCs. Additionally, the number of researchers that benefitted from RDC services is also instructive.

Since some RDCs are highly integrated into their parent institutions, while others have a strong service infrastructure and external orientation, putting the usage figures into context is not trivial.

One of the core tasks of the RDCs is to provide researchers with comprehensive and flexible data access and to continuously expand and improve upon it. Owing to differences in data protection regulation and other legal provisions, researchers are offered a range of different data access paths. Due to this flexibility in access paths and differences in how data on contracts (data use agreements), projects, and data users are gathered by RDCs, it is difficult to determine precise numbers regarding the full scope of data use. In previous years, it was therefore not possible to rule out double counting of contracts, projects, or individual access paths.

Getting a clear idea of this diversity has proven a complex task. For this reason, four partial indicators are being differentiated since the 2019 Activities Report: number of downloads, external data users, data use agreements, and the number of surveyed users and datasets.

68,752
Free downloads
n=17

Downloads

Overall, 17 RDCs make datasets available as free downloads. At the majority of RDCs, downloading requires prior registration. Some datasets are freely downloadable without prior registration – making it more difficult to identify users. The number of users can therefore not be documented for all data retrievals. Furthermore, not all data access paths have the technical means required for gleaning usage statistics. Another complicating factor was that due to technical problems at the beginning of 2020, data access was initially not possible at a large FDZ. Subsequently the number of downloads could not be determined precisely due to technical changes. Overall, the number of downloads in 2020 is even more undercounted than in previous years. The eleven RDCs who are able to provide information on user numbers reported 68,752 downloads of open datasets in the 2020 reporting year.

The standard case, however, is that RDC-held research data are made available only after users have registered or signed an agreement.

43,703
External data users
n=28

+ 7,662
New
external data users
n=28

External data users

The RDCs counted a total of 43,703 data users in this reporting year. Users are considered external when they are not affiliated with an RDC or an RDC's parent institution. Twenty-eight RDCs were able to provide data on the number of external data users.

In 2020, the number of external data users increased by 7,662 persons. Twenty-eight RDCs were able to provide data on new data users in 2020, while two RDCs experienced technical difficulties counting new data users. This means that the number of new data users in 2020 tends to be undercounted and is likely much higher.

Data use agreements

For data protection reasons, agreements on data access and usage contain explicit references to research projects (purpose limitation), i.e., a separate agreement must be drawn up for every research project using the data. However, there are no formal templates for such agreements. How RDCs design their agreements is governed by the freedom of contract, and the contractual depth is determined by legal provisions and requirements. Access to official statistics data, for example, is legally restricted by a string of laws and regulations. Access to survey data is also subject to data protection regulation. This applies particularly to sensitive personal data. Other data, including regionalisation and land use data, are openly accessible for some purposes and subject to licensing for others. This diversity is also reflected by the agreement design. This is true for the data themselves as well as the signatory parties: Agreements can be with individuals, projects, or entire research facilities. They can cover entire data troves, collections of studies, or individual datasets.

In 2020, 37 RDCs reported 38,219 existing data use agreements in the RatSWD's research data infrastructure. Thirty-three RDCs concluded 6,321 new data use agreements in 2020.

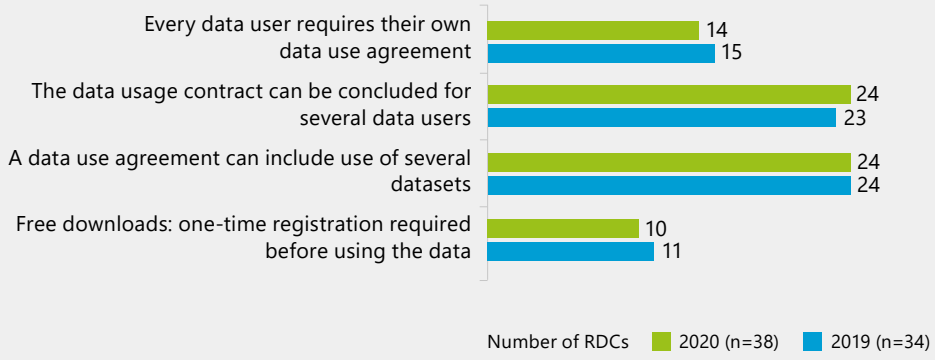
Fourteen RDCs reported that each dataset required an individual data use agreement. Twenty-four RDCs allow for agreements covering several users. At 24 RDCs, data use agreements govern access to several datasets (▸ Fig. 16).

38,219
Current data use agreements n=37

+6,321
Newly concluded data use agreements in 2020 n=33

Fig. 16: Contract design

Which of the following applies to your RDC? *(Multiple answers possible)*



Contract design varies strongly: Data use agreements can cover one or several persons or datasets

© RatSWD 2022

When agreements are signed with several users, 19 RDCs permit use only for specifically stated persons. On average, an agreement gave access rights to 2.8 persons. Agreements of 16 RDCs are signed at the project level; ten RDCs draw up institution-level agreements.

When agreements govern access to several datasets, only one designated person may use them at 12 RDCs; 15 RDCs extend access to several persons. On average, agreements covered slightly more than two persons. Project-level agreements were used by nine RDCs. With only one RDC, institution-level agreements are an exception.

5 Established data access paths and service concepts

In this chapter, the data access paths established at RDCs as well as the broad spectrum of service concepts to support data users are individually examined in greater detail. Additionally, the first section highlights the various instruments used by RDCs to advertise their own data offerings.

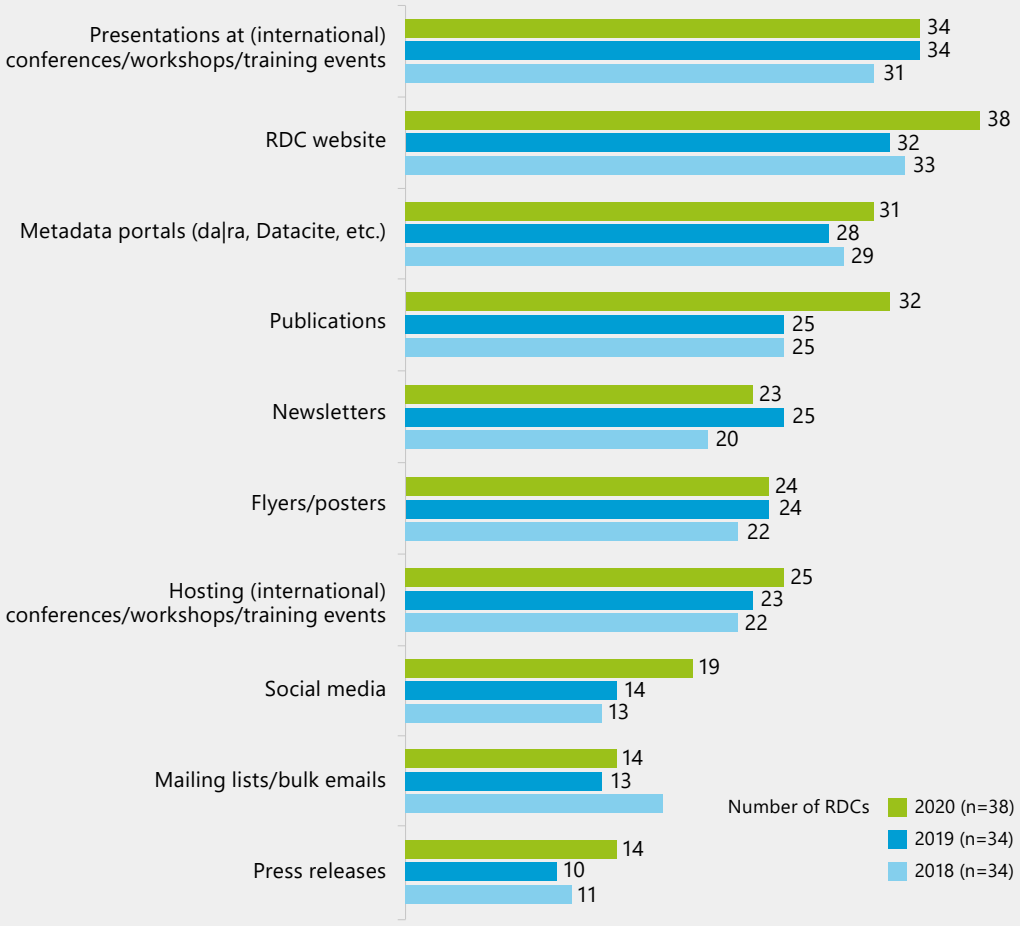
Advertising the data offering

The RDCs use a wide range of communication channels to advertise their data offering within the scientific community and to highlight its potential for answering scientific questions. The dominant instrument for self-promotion is an own website (n=38), presentations at (inter)national conferences, trainings, and workshops (n=34), and the use of metadata portals like da|ra or Datacite (n=31). Additionally, publications as a means of self-promotion saw an increase from 25 RDCs in the previous years to 32 RDCs in the 2020 reporting year.

Publications and social media platforms are gaining relevance for communicating data offerings

Fig. 17: Information channels used by RDCs to communicate their data offering

How and through which channels are the offered data made known to the scientific community?
(Multiple answers possible)



© RatSWD 2022

Likewise, there is a visible increase in acceptance of social media platforms for self-promotion. While only 14 RDCs reported using social media platforms to publicise their data offering in the previous year, 19 RDCs did so in the 2020 reporting year (▸ Fig. 17, p. 26). It must be noted that these increases in using publications and social media platforms as communication channels cannot be explained solely by new RDC accreditations. Several more "experienced" RDCs have also discovered these channels for themselves.

Access paths in data provision

As shown in Chapter 4, the RDCs offered a total of 4,917 datasets on the cut-off date 31/12/2020. Of course, these microdata differ not only thematically but also in their degree of anonymisation. The latter decisively determines the possible data access paths.

Info box 2

Data access paths

Guest researcher workstations

Data are made available to researchers at specially secured workstations at the RDCs. It is common for guest researcher workstations not to have unregulated internet access and to disable local saving of files. Resulting files are checked for adherence to data protection regulation before they are transmitted (output control).

Controlled remote data processing

Researchers can analyse data using remote access without being on location at the RDC. Data storage and processing occurs exclusively on RDC servers and result files are only transmitted to researchers after having been checked by RDC staff (output control). Two basic procedures can be distinguished here:

- Remote execution:
  Data access path that does not enable researchers to directly view the data or intermediate results. Researchers submit analysis scripts to RDC staff, which they write at their respective workplaces (sometimes using a structured dataset to test their code), and submit these to the RDC, which applies them to the original data.
- Remote desktop:
  Data access path that transmits the RDC server's user interface to the screen of a researcher's local client using remote access. Appropriate configuration is used to prevent local saving of data. The local access device is only used to communicate with the data server. Researchers use the analysis software stored on the server and can work with the data as if they were stored locally.

Other data access paths include data transmission to users via download, email, or regular mail, where data can be analysed directly on a local computer. These access paths differ regarding their level of anonymisation and purpose:

Scientific Use Files (SUFs)

SUFs are research datasets that are de-facto anonymous datasets but still have considerable analytic potential.

Campus Files (CFs)

Even more anonymised research datasets compared to scientific use files, created for university teaching purposes.

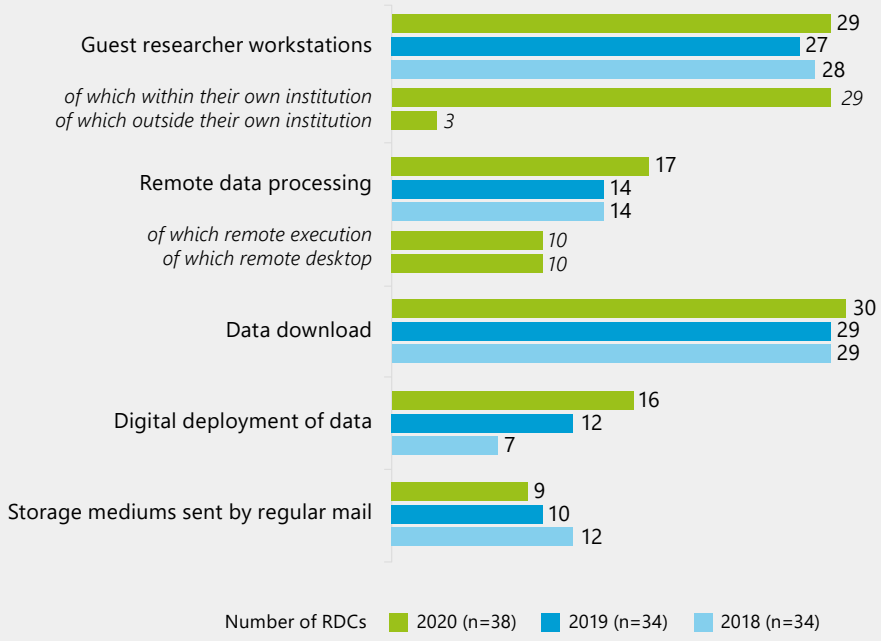
Public Use Files (PUFs)

Anonymised research datasets without use restrictions that can be shared even for non-academic purposes.

The EU's General Data Protection Regulation (EU-GDPR) disposed of the hitherto used terms formal and de-facto anonymisation.[13] It now only differentiates between pseudonymisation and anonymisation. However, the old terminology is still used here to better distinguish the levels of anonymisation of the data access paths. Formally anonymised microdata can only be shared while applying high technical and organisational measures. Guest researcher workstations and the modes of remote access have become the established data access paths here. De-facto or absolutely anonymised microdata like scientific use files (SUFs), campus files (CFs; sometimes called campus use files, CUFs), or public use files (PUFs) are made available via download by default but can also be transmitted via email or regular mail. Info box 2 on page 27 presents the data access paths in greater detail.

Fig. 18: Data access paths offered by RDCs

Which data access paths does your RDC offer?
(Multiple answers possible)



© RatSWD 2022

Digital deployment of data is becoming more relevant

More than three quarters of RDCs (n=29) make available microdata via guest researcher workstations, while three[14] of these enable researchers to access their data offering at workstations that are not based at their own institution. Seventeen RDCs make their data offering available using remote data processing, either as remote execution or remote desktop[15] (▸ Fig. 18). Only 14 RDCs offered this access path in the previous year. This is not on only an increase in absolute but also in real terms since only one out of five newly accredited RDCs offers remote desktop access.

Slightly more than three quarters (n=30) of all RDCs facilitate protected downloading of microdata. Sending physical storage mediums via regular mail saw a slight decrease from ten RDCs in 2019 to nine in 2020.

13 Recital 26 of the EU-GDPR refers to the concept of anonymisation. Additionally, Article 89 EU-GDPR deals with the "research privilege" of working with personal data while adhering to the anonymisation provision for certain categories of personal data.

14 In the previous year, the RDCs of the federal and state statistical offices were surveyed separately. The decrease from four RDCS in the 2019 reporting year to now three RDCs is solely due to the fact that those RDCs were jointly surveyed this reporting year.

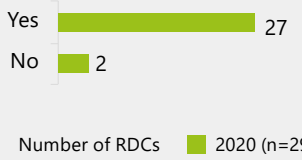
15 Some RDCs use controlled remote data execution methods that do not permit looking into the data but make possible previews of the results.

The trend is going towards digital data deployment. While seven RDCs in 2018 and twelve RDCs in 2019 deployed their data offering through digital means, 16 now do so in the 2020 reporting year. This increase is not purely caused by newly accredited RDCs.

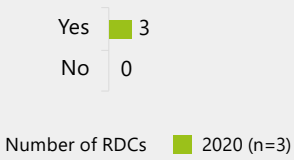
The COVID-19 pandemic, which has been affecting European countries since the 2020 reporting year and involved strict lockdown and social distancing measures, also had a negative impact on operating guest researcher workstations (▸ Fig. 19). Twenty-seven RDCs reported that they temporarily shut down guest researcher workstations in their institutions due to restrictions for curbing the COVID-19 pandemic. These limitations also affected RDCs operating guest researcher workstations outside of their own institution. Moreover, restrictions of guest researcher workstations continued after their reopening. Especially capacities were still limited due to the restrictions of the Infection Protection Act.

Fig. 19: Effects of the COVID-19 pandemic in 2020

19a: Did your RDC have to close any guest research positions within its own institution due to the COVID 19 pandemic in 2020?



19b: Did your RDC have to close any guest scientist positions outside your institution due to the COVID 19 pandemic in 2020?



© RatSWD 2022

Closing of guest researcher workstations due to the pandemic

Processing time after a signed agreement was received

This section examines in greater detail the average amount of time needed by an RDC to approve the release of the applied microdata after receiving a signed data use agreement. The actual time of data provision is not the issue here, as data users can cause delays in the actual provision of data, in which case RDCs do not hold responsabiltiy. Data users can, for example, retrieve microdata much later after they have been deployed for downloading. With microdata offered through a guest researcher workstation, the date of the actual data provision depends on the date of the guest's stay.

The duration of processing at RDCs depends, for example, on whether datasets must be produced especially or existing datasets must be at least adjusted, or whether existing datasets can be made available to data users without any necessary preparation. Lastly, various bureaucratic processes and the level of automation within an institution also play a role.

Processing time upon receipt of a signed agreement varies between RDCs. Slightly more than half of all RDCs (n=20) is able to make the applied datasets available within a week after the signed agreement has been received. Slightly more than a quarter of all RDCs (n=11) typically makes datasets available within four weeks. Provision takes longer than four weeks at four RDCs.

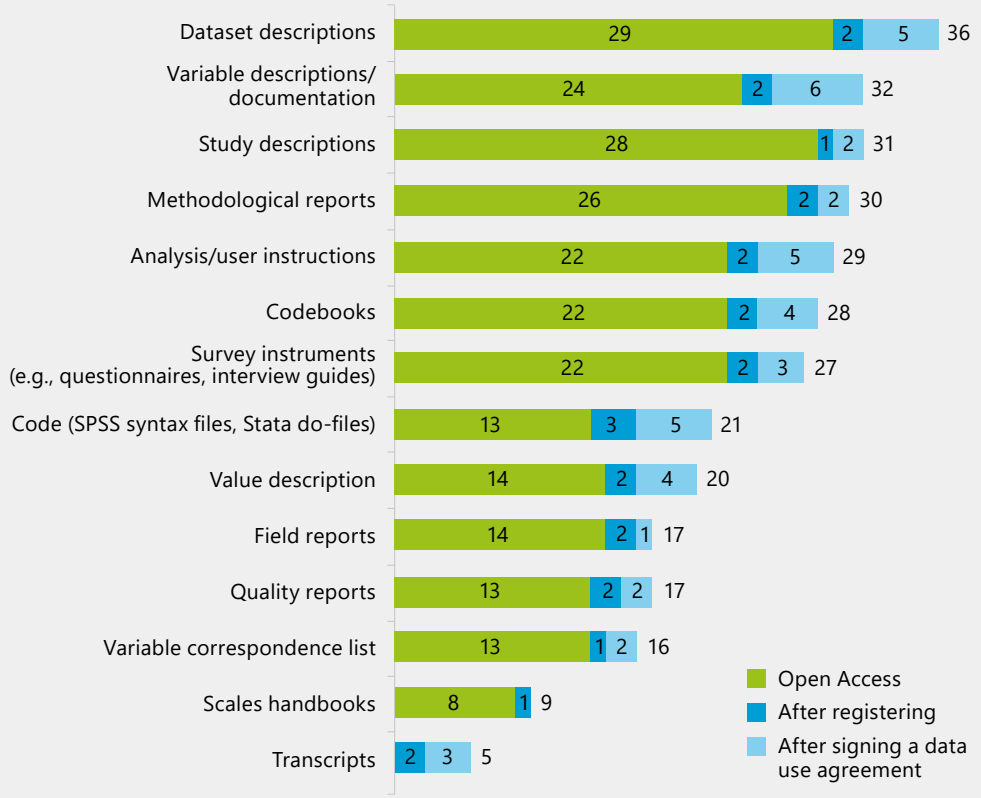
Provision of manuals and tools

Dataset-specific tools are useful for data users to get an insight into the potential of a dataset for answering certain research questions. Openly accessible, dataset-specific tools are an advantage for both data users and RDCs: Comprehensive dataset descriptions or codebooks, for example, can help data users decide whether a dataset is suitable for a research project ahead of applying for it, which avoids drawing up unnecessary contracts and reduces the workload. The following graph (▸ Fig. 20) gives an overview of which tools are offered and how they are accessed.

RDCs offer a comprehensive and diverse range of tools

Fig. 20: Provision of manuals and tools in 2020

Which tools and materials do you offer for which data access path? (n=38)
(Multiple answers possible)



© RatSWD 2022

As shown in the figure above, almost all RDCs make additional information material available with every dataset they offer. The majority are made available in an open access format[16] through the RDC websites, enabling data users to gather information ahead of their application. Independent of the data access path, the main tools provided by RDCs include dataset, variable, and survey descriptions as well as methodological reports.

Individual RDCs offer more specific tools beyond that. In qualitative research, this includes additional information on board notes, seating plans, or interview protocols, for example. RDCs also use new forms of communication to support data users, including video tutorials or interactive metadata portals.

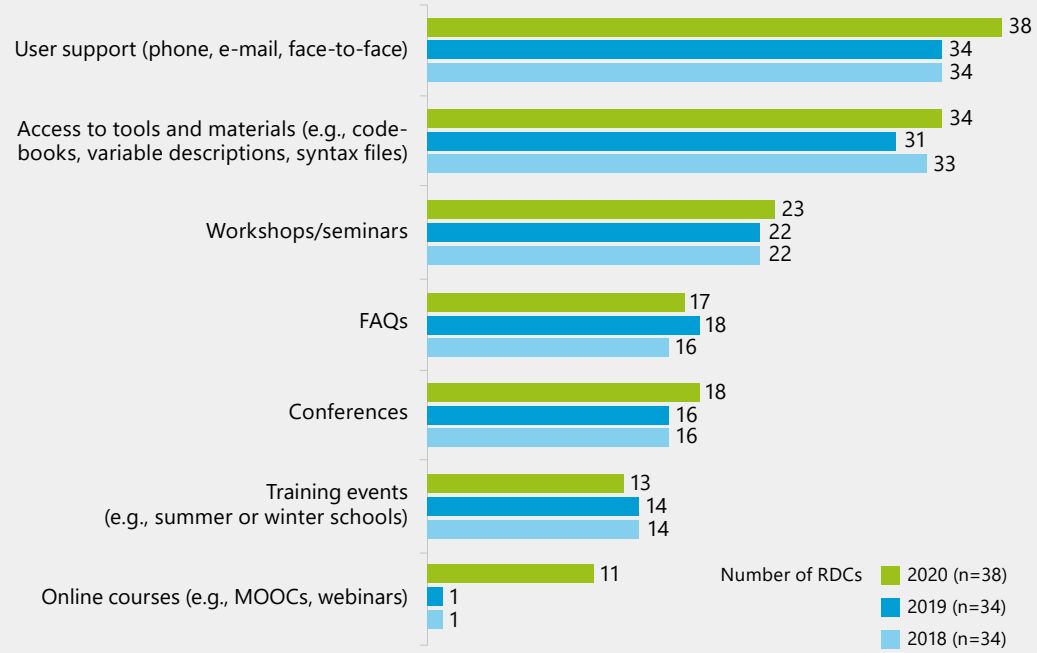
16 The aim of open access is to make scholarly literature and material freely available to users; free of charge as well as free of technical and legal barriers.

Services for data users

In addition to preparing microdata and deploying tools, supporting, and advising data users is an important pillar of the work done by RDCs. As shown in Chapter 2, RDC staff can invest a portion of its work hours into active research. This experience from using their own data offering to answer scientific research questions is particularly useful when supporting external data users because RDC staff are uniquely aware of the potential and the pitfalls of datasets and can communicate them well. The following section will look at RDC user services in greater detail.

Fig. 21: User services at RDCs

Which user services does your RDC provide?
(Multiple answers possible)



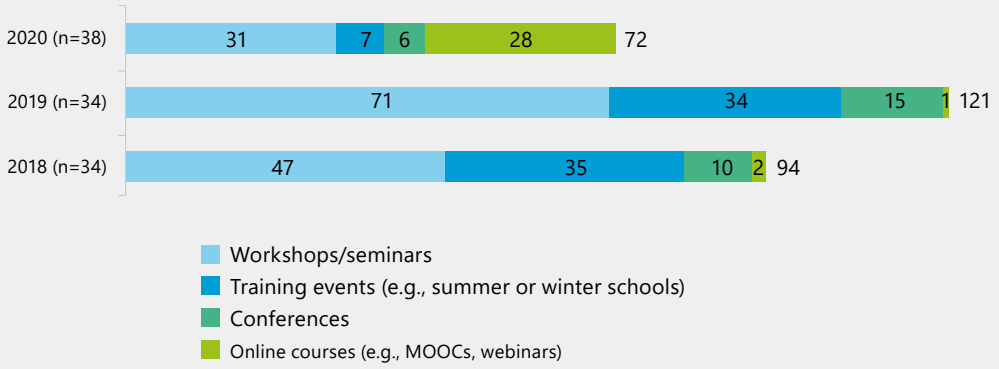
© RatSWD 2022

All RDCs focus on tailored user support

All RDCs provide individual data user support via telephone, email, and on site. The number of RDCs offering workshops, seminars, conferences, and trainings has remained almost constant. Compared to the previous year, the large increase in RDCs offering online courses like Massive Open Online Courses (MOOCs) and webinars is particularly interesting. While only one RDC offered online courses in 2018 and 2019, eleven RDCs did so in the current reporting year. Additionally, some RDCs offered virtual support such as video tutorials and discussion forums (▸ Fig. 21). This extraordinary increase can largely be attributed to a more general shift into the digital realm as a result of the COVID-19 pandemic and the efforts of RDCs to offer additional online services to compensate for pandemic-related restrictions on offline services. This trend is also observable in skills development measures in the 2020 reporting year (▸ Fig. 22, p. 32).

Fig. 22: Skills development

How many of the following skills development measures does your RDC offer?
(Multiple answers possible)



© RatSWD 2022

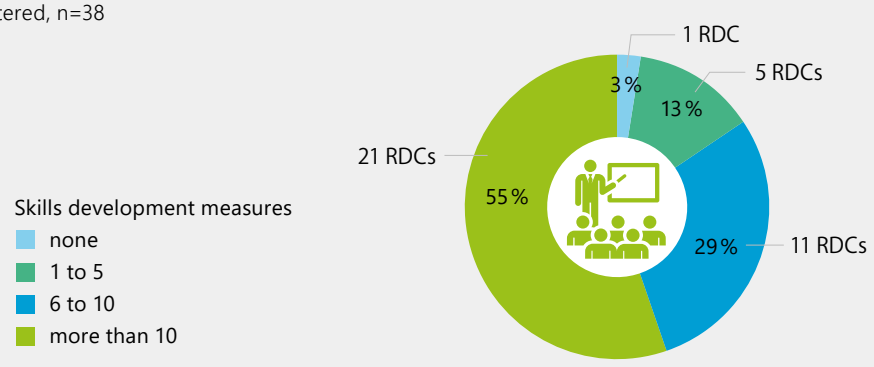
Increase in online courses with the start of the pandemic

While 2019, the last year before the COVID-19 pandemic, saw an increase to 71 workshops and seminars, only 31 workshops and seminars were hosted overall in the 2020 reporting year. The same negative trend was observable regarding trainings, with a decrease from 34 trainings in 2019 to seven in 2020, and conferences with 15 events in 2019 and only six events in 2020. The number of online courses soared, however. While only one RDC hosted online courses in 2019, nine RDCs hosted a total of 28 events in the 2020 reporting year. This increase can be attributed to three RDCs in particular, who considerably expanded the number of online courses they offered. Compared to the previous year, the number of skills development measures sunk from 121 to 72 events.

The skills development measures offered in the 2020 reporting year are spread unevenly across RDCs (▸ Fig. 23). More than half of RDCs (n=21) did not offer any skills development measures, while eleven RDCs offered between one and five skills development measures. It should be noted that one RDC was capable of offering more than 10 skills development measures, the majority of which were online courses.

Fig. 23: Number of skills development measures at RDCs

clustered, n=38



© RatSWD 2022

The number of skills development measures per RDC varies strongly

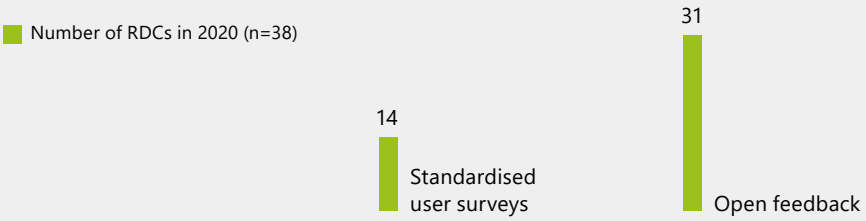
Service quality assurance

Most RDCs use process-integrated measures to assure the quality of their services (▸ Fig. 24). Thirty-one RDCs use open user surveys for quality assurance, while a total of 14 RDCs use standardised user surveys (of which eleven did so additionally to open feedback). The frequency with which user surveys are conducted varies significantly. Six RDCs report conducting these surveys continuously, while five RDCs survey data users less than once a year.

Some RDCs receive additional feedback through feedback questionnaires following training events or through internal evaluation by a scientific advisory board.

Fig. 24: Quality assurance through user surveys

How are services evaluated, and their quality secured?
(Multiple answers possible)



© RatSWD 2022

Evaluating open feedback from users is the most common form of quality assurance

RDCs discuss, evaluate, and, where possible, implement the feedback they receive. This constitutes an important impulse for the continual development of the infrastructure. Additionally, anonymised feedback is sometimes published on RDC websites. To summarise, it can be shown that feedback, which the RDCs receive through various channels, is a relevant basis for them to optimise their service offerings as well as to improve internal work processes.

6 Current development of the research data infrastructure in the social, behavioural, and economic sciences

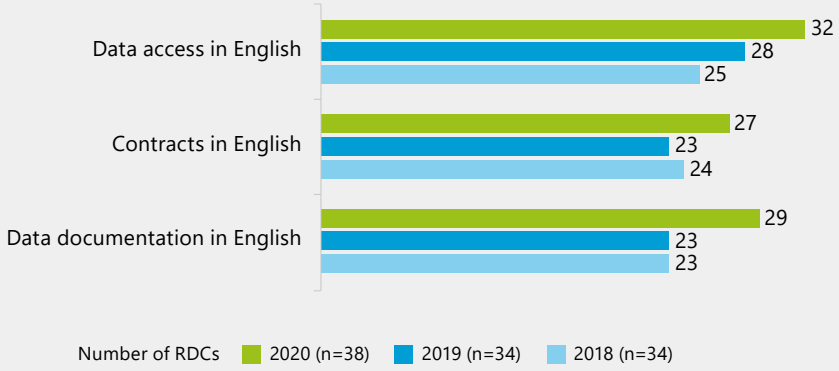
Current developments include the international orientation of RDCs with respect to data offering as well as cooperation. Moreover, this section details innovation and development within the research data infrastructure.

Internationalisation

The knowledge society of the 21st century, where global interdependence is rapidly accelerated by the digital transformation, is continuously giving rise to new fields of research, which can only be addressed in the context of international scientific cooperation. On the one hand, a prerequisite for this is that international researchers can access national-level data. On the other hand, there is also an increasing demand for international microdata, which facilitate better comparative analysis. International exchange and cooperation have long become everyday practice at RDCs and is increasingly shaping their work. For example, RDCs have created access paths and data documentation in English to cater to the international research community.

Fig. 25: Support for international researchers

How do you support international researchers?
(Multiple answers possible)



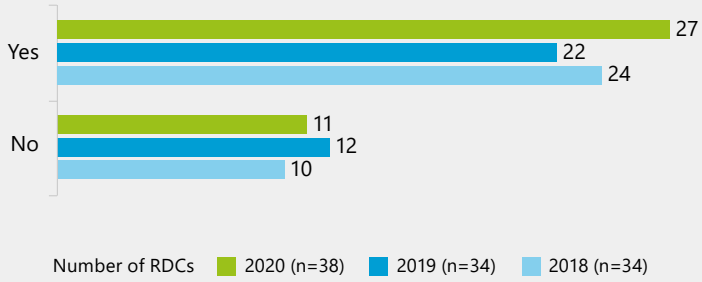
© RatSWD 2022

Four additional RDCs now offer access paths in English

Thirty-two RDCs, four more than the previous year, now offer data access in English. This increase can be explained in particular, but not exclusively, by newly accredited RDCs. In terms of contract management as well as data documentation, there is a clear trend towards expanding English-language offerings at RDCs. While 23 RDCs additionally provided data use agreements and data documentation in English the previous year, 29 and 27 RDCs, respectively, did so in the 2020 reporting year (▸ Fig. 25). Furthermore, international users are provided with tailored user support via email and telephone in English as well as English versions of websites and newsletters. Moreover, individual RDCs offer regular data trainings, workshops, and conferences in English as well as offering opportunities for researchers to go abroad.

Fig. 26: RDCs with international contacts

Does your RDC have contacts to international research facilities?



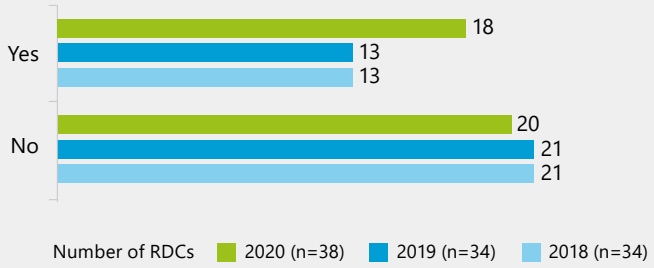
© RatSWD 2022

Contacts to research facilities at the international level are increasing

The international orientation of RDCs is also reflected by ongoing exchanges with research facilities at the international level and international research partnerships, serving not only to respond to research questions but also to better coordinate and harmonise regulation in administration and data protection. In total, 27 RDCs have contacts to international research facilities (▸ Fig. 26).

Fig. 27: RDCs with close international research partnerships

Does your RDC maintain international research partnerships? (i.e., international researchers working together on specific issues)



© RatSWD 2022

International research partnerships are increasing

Eighteen RDCs, which is more than half, maintain a close exchange with international institutions as part of a research partnership (▸ Fig. 27). The nature of these research partnerships ranges from projects to working groups with European and international facilities like universities, RDCs, data archives, and research institutes.

Innovation and improvement of the research data infrastructure

The 2020 reporting year again saw a broad thematic range of innovation and improvements of the research data infrastructure. This included the continuous expansion of the data offering by adding new data sets, which also comprised the provision of linked datasets of various RDCs. Moreover, several RDCs successfully expanded their existing data access paths by adding controlled remote execution to formally anonymised microdata. A trend that was already observable in the previous year.

Opportunities for international data users to use the RDCs' microdata to work on scientific research questions have been extended by opening up new locations for guest researchers. In addition, many RDCs have advanced the automation of internal work processes as well as offering trainings and workshops in an online format.

Further developing the research data infrastructure

All RDCs are committed to continuous improvement. They continuously expand their data offering as well as opportunities for data access, complement services with online formats, and intensify their international orientation. True to the motto "If you don't go forward, you go backwards," the RDCs continue to further what they have achieved in quantity and quality. In this context, it is important to note that the situation among RDCs is heterogenous. While some RDCs have been offering and applying certain services and internal processes for several years, other RDCs are still working on planning or implementation. For this reason, it does not come as a surprise that the issues in which the RDCs indicate a need for support or an interest in knowledge exchange show a large overlap with the issues of the previous years.

In the 2020 reporting year, data protection continues to be an issue that RDCs show considerable interest in, particularly in terms of knowledge exchange and support. This includes robust information on legal issues regarding data protection and the continuing development of anonymisation methods against the backdrop of EU-GDPR requirements.

In *research data management*, five key issues were identified by RDCs as particularly relevant for the future:

- In the field of data access, RDCs are calling for an exchange on issues of alternative and innovative data access paths, particularly on standards and the technological basis for remote execution and remote desktop. There is also a need for exchange on possible automation of statistical confidentiality checks (output control).
- In regard to user management, they seek to create a definition of the term scientific institution with a corresponding list of pros and cons. Moreover, there is an interest in an exchange on current methods and tools for inquiry and contract management, particularly on ways to automate administrative processes.
- The RDCs also expressed an interest in discussing methods and tools for data harmonisation and ways to facilitate information flows in the form of structured metadata – from the collection of data until their end use.
- They also mentioned developing criteria for data quality, measures to increase secondary data use, portfolios, and tools to support data preparation during the research process, and skills development for young researchers.
- The surveyed RDCs also mentioned the handling of new tasks, for example, the role of RDCs as data trustees.

7 Special topic: Metadata standards

The so-called **FAIR principles (short for Findability, Accessability, Interoperability, Re-usability)** have been established as guidelines for how to organise research data infrastructures in recent years. Nationally[17] and internationally[18], they are seen as a standard for data sharing. Ways to re-use data are often at the forefront of the debate. A prerequisite for successfully re-using data are high-quality and granular metadata. When data descriptions are incomplete or incomprehensible, even data that are open access are unusable.

The launch of KonsortSWD as part of the NFDI created a structure for further increasing the FAIRness of data from the social sciences by improving metadata quality and further harmonising the creation of metadata. One of the three pillars of KonsortSWD is strengthening the FAIRness of data and metadata.

The strong growth of the RDC network in the past few years necessitated generating an empirical basis to examine how metadata are created and managed at RDCs. For this reason, the FDI Committee decided to dedicate this year's special topic of the monitoring report to metadata standards.

Metadata practices will be presented based on the four pillars of the FAIR principles. Each section is concluded with a brief assessment of the current state of implementation at the RDCs.

17 Hartl, N., Wössner, E., & Sure-Vetter, Y. (2021). *Nationale Forschungsdateninfrastruktur (NFDI).* Informatik Spektrum, 44(5), 370-373. https://doi.org/10.1007/s00287-021-01392-6.

18 Streit, A., & Wezel, J. (2021). *Deutschland in der European Open Science Cloud.* In M. Putnings, H. Neuroth & J. Neumann (Hg.), Praxishandbuch Forschungsdatenmanagement (S. 31-52). De Gruyter Saur. https://doi.org/10.1515/9783110657807-003.

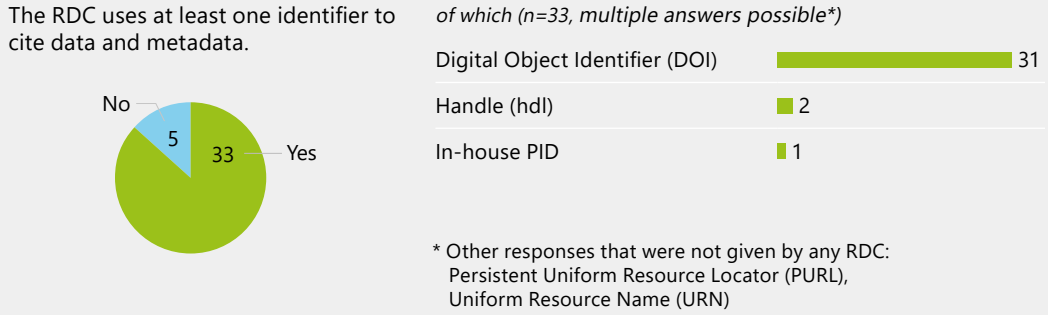
Findability

A core element of FAIR data infrastructures is systematic use of PID for data and metadata. These identifiers facilitate making datasets and their metadata findable and citable. Contrary to other serial identifiers (e.g., website URLs), PIDs are resolved not to an object's online location but to the object itself. If the location of a digital object, referenced by a PID, is changed, the identifier stays the same. When PIDs are used, they are typically connected to the referenced object with a minimum of generic metadata (for example, authors, title, survey period, etc.).

In the past few years, the introduction of PIDs was one of the main areas of development at RDCs. From 38 accredited RDCs, 33 are already using PID, of which 31 use the DOIs from DataCite[19] and two use so-called *handles*[20]. One RDC uses an in-house PID (▸ Fig. 28). Three other RDCs are currently working on introducing DOI for their data.

Fig. 28: Use of persistent identifiers

Which worldwide unique, permanent, and resolvable identifiers (referred to in the following as Persistent Identifier – PID) does your RDC use to cite data and metadata? n=38



© RatSWD 2022

PIDs are used by RDCs to reference data. However, concrete implementation varies. In 20 cases, PIDs are applied to all data of a survey (i.e., one PID is resolved for several data files that belong to a survey, as is the case for individual panel waves, for example). At 19 RDCs, PIDs are applied to individual data files (every panel wave has its own PID) and, at five RDCs, PIDs are also applied to individual data objects (interview transcripts, for example). The RDCs do not apply PIDs to individual objects or fragments within data files. However, at least eight RDCs assign them to context and study documents (including questionnaires, codebooks, methodology reports, working papers, and test instruments). At 30 RDCs, the PIDs are not directly resolved to the digital object but to a so-called landing page, which contains the metadata and documents as well as general information on how to access the data. Only at one RDC do PIDs resolve directly to the digital object.

As recommend by best practices on PID use, 29 RDCs explicitly list these in the metadata (data citation link). This helps users to identify the relationship between the PID and the data. At a technical level, PID and metadata are connected via the metadata schemes of DOI in da|ra[21] or DublinCore[22]. CMDI[23] is used in one case.

19 https://datacite.org/ (Last retrieved: 31/01/2022).
20 https://www.handle.net/index.html (Last retrieved: 31/01/2022).
21 Da|ra is the registration service for social science and economic data of GESIS - Leibniz Institute for the Social Sciences and ZBW - Leibniz Information Centre for Economics. https://www.da-ra.de (Last retrieved: 31/01/2022).
22 Dublin Core is a metadata standard used to describe electronic resources, as specified by the Dublin Core Metadata Initiative (DCMI). https://dublincore.org/ (Last retrieved: 31/01/2022).
23 Component Metadata Infrastructure (CMDI) is a component-based description of metadata. https://www.clarin.eu/content/component-metadata (Last retrieved: 31/01/2022).

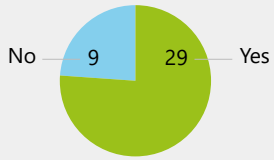
Using PIDs for entire datasets and studies is a tried-and-tested method at RDCs. However, the concept of *FAIR Digital Objects* (a model for the technical implementation of the FAIR principles in data management) provides that PID also be used for variables, fragments, or other attributes within data files.[24] High-granular referencing of data and metadata using PIDs is to facilitate automated access to (meta)data in the future. Indeed, six RDCs report that they plan to use additional PIDs for more granular referencing of data objects. Unlike PIDs assigned to data and studies, however, they plan to use Uniform Resource Names (URN) for attributes instead of DOI. Presumably, the reason is that assigning a very high number of DOI causes high costs as long as research data managers do not have access to a "national" DOI license.

Detailed metadata describing the data are an other pillar of improving data findability. To make these descriptions machine-readable and linkable, metadata should be generated using an established standard. ▸ Figure 29 shows that a large majority of RDCs already uses metadata standards for documentation and some also combine different standards, especially if they developed their own metadata scheme for describing data. Twenty RDCs use the da|ra metadata scheme, which is required for assigning DOIs. Eleven RDCs use their own metadata scheme for documentation, or the metadata standard of the Data Documentation Initiative (DDI), respectively. RDCs also adjust the DDI metadata scheme to meet their needs. Examples for this are DDI-LimDAS[25] by GESIS, or the metadata model of RDC-BO[26]. Other metadata schemes used by RDCs include DublinCore, Schema.org, PsychData, Inexda, CLARIN-Component MetaData Infrastructure, the metadata scheme of Verbund Forschungsdaten Bildung, and INSPIRE-OGC.

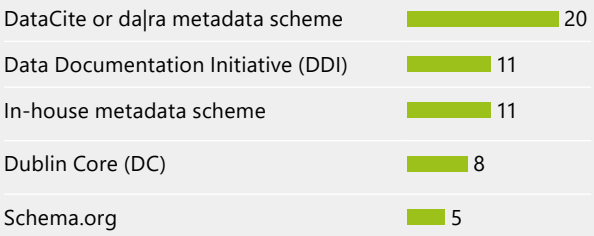
Fig. 29: Use of metadata standards and schemes

Which metadata standard or which metadata scheme does your RDC use to ensure data findability? n=38

The RDC uses at least one standard or metadata scheme to ensure data findability.



of which (n=29, multiple answers possible)*



* Other responses that were not given by any RDC: Statistical Data and Metadata Exchange (SDMX), Directory Interchange Format (DIF), Qualitative Data Exchange Format (QuDEx) or REFI-QDA, Data Catalog Vocabulary (DCAT)

© RatSWD 2022

Data findability is increased by sharing metadata with specialised search portals. The metadata from 28 RDCs, which use the da|ra service for documenting their data, are indexed by international data search portals such as DataCiteSearch, CESSDA Data Catalogue, the European search portal B2Find, the OpenAire search portal, and GoogleDatasetSearch. At least four of these RDCs prepare their metadata specifically for the search portal of Verbund Forschungsdaten Bildung, an infrastructure for education data. Of the RDCs that do not use da|ra, three use DataCiteSearch and three use other search portals. Only three RDCs do not (yet) feed their metadata into international search portals for research data (such as B2Find, OpenAire, CLARIN-VLO).

24 Betancort Cabrera, N., Bongartz, E. C., Dörrenbächer, N., Goebel, J., Kaluza, H., & Siegers, P. (2020). *White Paper on implementing the FAIR principles for data in the Social, Behavioural, and Economic Sciences*. RatSWD Working Paper 274/2020. Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.60Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.60.

25 http://dx.doi.org/10.21241/ssoar.65593 (Last retrieved: 31/01/2022).

26 https://www.diw.de/documents/publikationen/73/diw_01.c.620524.de/diw_datadoc_2019-099.pdf (Last retrieved: 31/01/2022).

Interim result – Findability

Persistent identifiers and metadata in accordance with established standards are cornerstones of a FAIR data infrastructure. The accredited RDCs have set a good example here. Few exceptions aside, users can find RDC data through subject-specific or generic search portals. Nine RDCs do not use a metadata scheme and ten RDCs use an in-house metadata scheme to describe their data. The heterogeneity of the way data are described also reflects the diversity of domains in which RDCs make data available. RDCs that do not yet use PIDs should plan their implementation soon in order to keep up with new developments in the NFDI or the European Open Science Cloud (EOSC) and to make their own data more visible. The few RDCs that do not yet share their metadata with search portals should seek a technical solution here to improve the findability of their data for users. In future, it should be evaluated how to achieve increased standardisation of the metadata practice at RDCs to optimise automated processing of metadata.

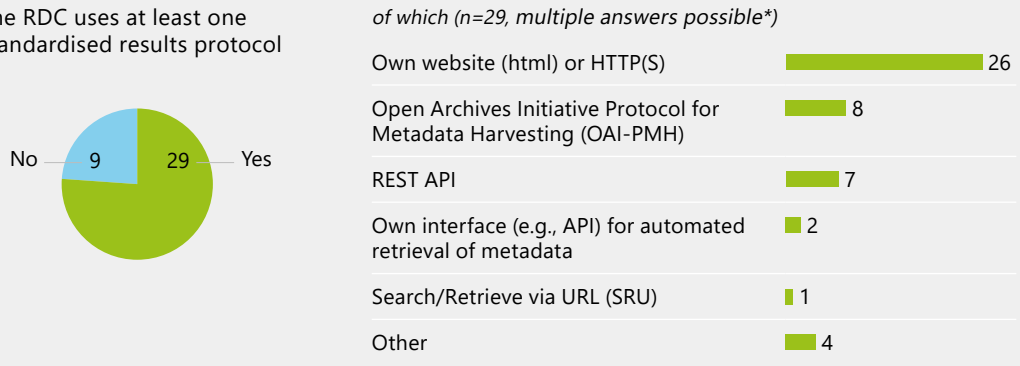
Accessability

Accessibility in accordance with the FAIR principles means that metadata are accessible through standardised communication protocols and are retrievable for humans and machines. Data and metadata should be preserved in the long term. The protocols that are used should be open, free, and universally implementable. This means that the use of communication protocols should not incur extra costs.

Since the websites of their institutions are the RDCs' main advertising instrument, websites based on the HTML or HTTPS protocol are the most widespread communication protocols of RDCs. Moreover, eight RDCs use the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which was specifically developed for exchanging metadata between archives, and seven RDCs use a Representational State Transfer-Application Programming Interface (REST API) – a generic interface for exchanging data in web-based applications (▸ Fig. 30). OAI-PMH and REST API are designed for automated data exchange and make sure that metadata can be processed by machines. OAI-PMH is also offered by da|ra so that metadata reported by RDCs to da|ra can be retrieved through the da|ra API.

Fig. 30: Standardised communication protocols for metadata

Which standardised communication protocol does your RDC use to retrieve metadata?
n=38 *(Multiple answers possible)*



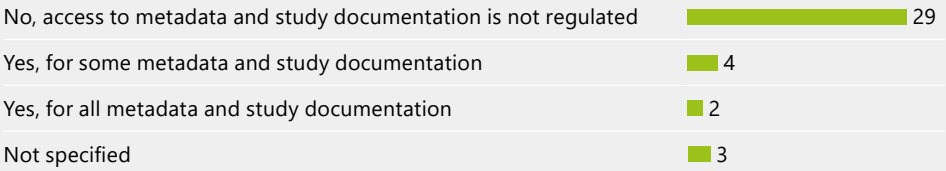
* Other responses that were not given by any RDC:
File Transfer Protocol (FTP) and Simple Mail Transfer Protocol (SMTP)

© RatSWD 2022

Authentication of users and administration of usage rights are crucial tasks for RDCs due to the various degrees of protection applied to data and metadata. The FAIR principles call for access to data to be as open as possible, as far as this is legally possible. The overwhelming majority of RDCs provide free access to metadata and study documentation. At six RCDs, users must authenticate themselves to gain access to metadata (▸ Fig. 31).

Fig. 31: Open access to metadata

Does your RDC use an authentication infrastructure (user profile) to regulate access to metadata and study documentation? n=38



© RatSWD 2022

Another dimension of accessibility is the long-term storage of metadata, ensuring that knowledge about existing data troves is preserved. Only seven RDCs do not have a strategy to preserve metadata beyond their own existence. All other RDCs hand over their metadata to infrastructure partners who ensure their long-term availability. Da|ra and the GESIS Data Archive are mentioned, although the latter does not yet have an official succession plan itself. Other metadata services include HEBIS[27], Pangea[28], PsychArchives[29], or university-based services.

Interim result – Accessibility

The accessibility of metadata at RDCs largely meets the requirements of the FAIR principles. They all provide information on available data and data access. Fifteen RDCs make metadata available in a machine-readable format through standardised programming interfaces. This allows search portals to incorporate the metadata into their own search indices and thus contributes to the long-term safety of the metadata. Expanding the use of programming interfaces to access metadata would further improve accessibility in future. Parallel to this – e.g., in the context of CoreTrustSeal certification – the RDCs could make reciprocal follow-up agreements for (meta)data amongst themselves to further reduce the risk of outages.

27 https://www.hebis.de/ (Last retrieved: 31/01/2022).
28 https://www.pangaea.de/ (Last retrieved: 31/01/2022).
29 https://www.psycharchives.org/ (Last retrieved: 31/01/2022).

Interoperability

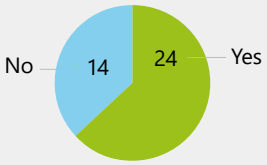
Interoperability is ultimately the most demanding of the FAIR principles because it is aimed at opening up data in a way that creates interfaces for linking data. Not only human data users are to be linked but also computer systems. A central prerequisite for this is the use of widespread standards that describe the content of data at a granular level. The aim is to enable computer systems to decide whether the content of datasets is comparable. Creating and applying such metadata requires controlled vocabularies, ontologies, and thesauri, and a clearly defined framework, e.g., in the sense of a Semantic web.

Such Semantic Web technologies are already used by RDCs to some extent (▸ Fig. 32). Seventeen RDCs make their metadata available in an Extensible Markup Language (XML) format. The advantage of this format is that it is interpretable for both humans and computer systems. Three RDCs use JavaScript Object Notation (JSON LD). This standard is used for exchanging data between different applications. One RDC uses the Resource Description Framework (RDF), which is primarily and extensively used in web development. Fourteen RDCs do not represent their metadata in any knowledge representation language. This limits possibilities for automated evaluation and linking of data.

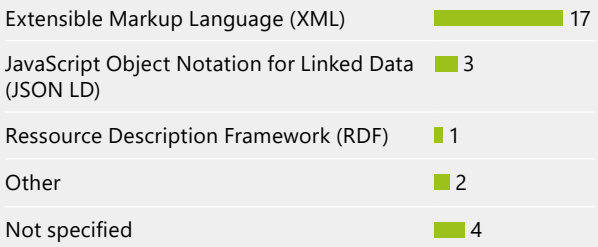
Fig. 32: Knowledge representation languages used to describe and display metadata

Which knowledge representation languages are used at your RDC to describe and display metadata? n=38

The RDC uses at least one knowledge representation language to describe and display metadata for facilitating machine-readability.



of which (n=24, multiple answers possible)*



* Other responses that were not given by any RDC: Ressource Description Framework Schema (RDFS), RDF(S) extensible knowledge representation model, Web Ontology Language (OWL), Agent Markup Language-Programm + Ontology Inference Layer (DAML+OIL)

© RatSWD 2022

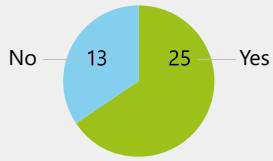
Standardised content indexing does not only require using open exchange formats but also semantically equivalent descriptions of the content. To do this, measurements of identical or similar constructs must be indexed using the same keywords in the data description. For this purpose, libraries and archives traditionally use thesauri and ontologies, ensuring uniform keywording. The use of similar keywords facilitates identifying linkable content. Even when different thesauri are used, mappings can help to identify comparable content of data.

Figure 33 shows the diversity of the thesauri in use. The controlled vocabularies of the Data Documentation Initiative (DDI CV) are most common, which were developed primarily for standard fields in the documentation of survey studies (e.g., survey method, sampling method). Some RDCs also use Thesaurus Sozialwissenschaften (TheSoz)[30], CESSDA Topic Classification[31], and the classification of the Journal of Economic Literature (JEL)[32]. In addition, subject-specific thesauri are widespread, including OGC-INSPIRE[33], PSYNDEX Terms[34], or the controlled vocabulary of Verbund Forschungsdaten Bildung[35]. The GESIS RDCs use the ISO 639-3 language codes for languages and 3166-1/2/3 for keywording of countries and regions. However, 13 RDCs do not use controlled vocabularies for content indexing of data, while nine RDCs have developed their own.

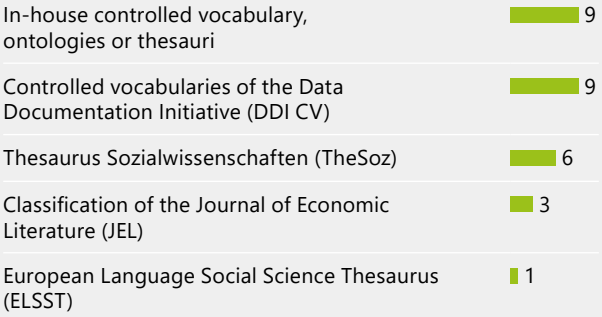
Fig. 33: Describing datasets with vocabularies, ontologies, and thesauri

Which controlled vocabularies, ontologies, and thesauri, used to describe datasets, do you use to annotate metadata documentation at your RDC? n=38

The RDC uses at least one controlled vocabulary, ontology, or thesaurus to describe datasets.



of which (n=25, multiple answers possible)*



* Other responses that were not given by any RDC:
Standard Thesaurus Wirtschaft (STW)

© RatSWD 2022

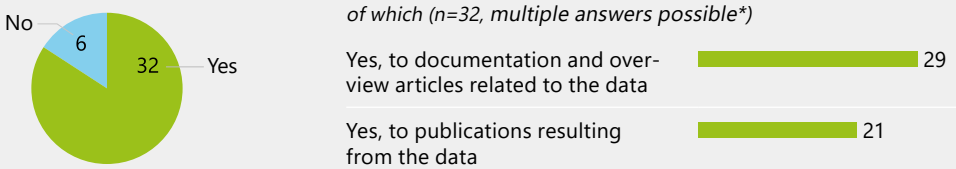
Creating standardised metadata requires appropriate tools that support relevant metadata standards, assignment of standardised keywords from controlled vocabularies, and access via the above-mentioned Semantic Web technologies. Here, the survey of RDCs shows that established editors are only used in a few exceptional cases. Most RDCs use their own metadata editors (Study level: n=16; variable level: n=9) or no editors at all (Study level: n=13; variable level: n=20). Commercial or open-source metadata editors are hardly used at all.[36]

30 https://lod.gesis.org/thesoz/de/ (Last retrieved: 31/01/2022).
31 https://vocabularies.cessda.eu/vocabulary/TopicClassification (Last retrieved: 31/01/2022).
32 https://www.aeaweb.org/jel/guide/jel.php (Last retrieved: 31/01/2022).
33 https://inspire.ec.europa.eu/ (Last retrieved: 31/01/2022).
34 https://www.psyndex.de/ueber/inhalte-aufbau/schlagwoerter-klassifikationen/ (Last retrieved: 31/01/2022).
35 https://www.forschungsdaten-bildung.de/files/fdbinfo_8_metadatenset_v1.0.pdf (Last retrieved: 31/01/2022).
36 The few exceptions mentioned include Colectica, Questasy, GeoMIS, DSpace-Web-UI, and da|ra.

On a positive note, data documentation includes references to publications associated with the data (▸ Fig. 34) in most cases. They include publications that provide additional information on the creation of the data or publications based on analysing the data.

Fig. 34: Referencing publications in the metadata

Do you refer to related publications in the metadata at your RDC? n=38



© RatSWD 2022

Interim result – Interoperability

To achieve high interoperability of research data, metadata must be highly standardised and be available in machine-readable exchange formats. The data descriptions must be semantically equivalent. These are high standards that require harmonised metadata practices. Although the data documentation of the RDCs is already based on common standards to a large extent (see above), the provision of metadata in modern exchange formats is still the exception. Harmonising metadata practice using common metadata editors could be a solution that facilitates linking of RDC (meta)data. Since operational open-source systems are currently not available, this goal will likely not be achieved in the short and medium term. However, more systematic use of controlled vocabularies within the standards implemented by the RDCs can be recommended to standardise semantic indexing of data.

Re-usability

Re-usability of data requires that users be able to trace how the data was created and which analyses are possible using them. The principle of re-usability differs from findability because it is up to the users' ability (machine or human) to decide whether certain data in their respective context are suitable for their intended use. Making this decision requires metadata that describe the context of data generation and make transparent who may use the data and under which conditions. This also applies to the use of metadata, which is what the following section is dealing with.

Terms of use for metadata can be defined, for example, using standard licensing models (for an overview of standard licensing models (see ▸ Tab. 1). ▸ Figure 35, p. 46 shows that 23 RDCs make their metadata available for re-use using a certain licensing model. Of those, eight RDCs use their own licenses and 15 use various iterations of the Creative Commons licenses. The most widespread is one permitting completely free use (CC0), which includes commercial use. This makes it possible for commercial search portals to harvest and index metadata. Some RDCs also use a CC BY license that requires giving appropriate credit. Only five RDCs use more restrictive licensing, among other things, to rule out commercial use of metadata.

Tab. 1: Creative Commons Lizenzen[37]

License	Permits:	Under the following terms:
CC BY	Reproducing, sharing, adapting the material, and reproducing and sharing adaptations for commercial and non-commercial purposes	Attribution: Credit the creator (if stated); Indicate the respective license type and link to the license text by URL/hyperlink; URL/hyperlink to the licensed material, as far as reasonably practicable; Copyright notice and disclaimer notice (both only where stated); where appropriate, indicate where changes were made to licensed material
CC BY-SA	see above	Attribution (see above); Share Alike: adapted material must be distributed under the same license
CC BY-ND	Reproducing, sharing, and adapting the material for commercial and non-commercial purposes; however, adaptations may not be reproduced or shared	Attribution (see above)
CC BY-NC	Reproducing, sharing, adapting the material, and reproducing and sharing adaptations but only for non-commercial purposes	Attribution (see above)
CC BY-NC-ND	Reproducing, sharing, adapting the material, and reproducing and sharing adaptations but only for non-commercial purposes; no reproduction/ sharing of adaptations	Attribution (see above)
CC0	Partial waiving of copyright; since this is not possible in German copyright law, this means the highest possible permissiveness when used	No attribution required
CC0 Plus (unofficial licence type, sometimes used by libraries*)	like CC0	like CC0, but with non-binding request to attribute work, as far as practicable

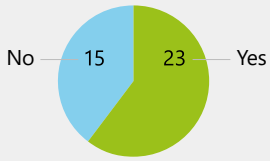
*See, for example, the British Library's Usage Guide for Catalogue datasets, https://www.bl.uk/about-us/terms-and-conditions/catalogue-datasets-in-rdf-and-csv.

37 This is an unedited reprint from: Lauber-Rönsberg, A. (2021). *Rechtliche Aspekte des Forschungsdatenmanagements*. In M. Putnings, H. Neuroth & J. Neumann (Hg.), Praxishandbuch Forschungsdatenmanagement. S. 89-114. De Gruyter Saur. https://doi.org/10.1515/9783110657807. This publication is licensed under the Creative Commons license (CC BY 4.0): https://creativecommons.org/licenses/by/4.0/ (Last retrieved: 31/01/2022).

Fig. 35: Licenses for metadata documentation

Which license does your RDC use for metadata documentation? n=38

The RDC has specified at least one license.



of which (n=23, multiple answers possible)*

* Other responses that were not given by any RDC: CC BY ND and CC BY NC ND

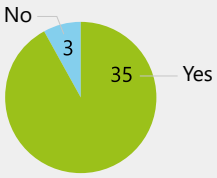
© RatSWD 2022

For users to be able to assess the usefulness of data, they need detailed information on the data-generating procedure and – ideally – on data processing procedures, especially when these have resulted in altering the data. This is referred to as provenance information in the terminology of the FAIR principles.

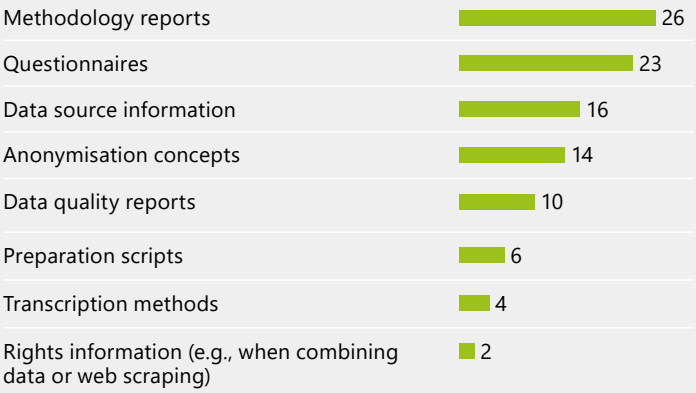
Fig. 36: Information on provenance

Which provenance information does your RDC publish as part of the study documentation? n=38

The RDC publishes at least one source information on provenance as part of the study documentation.



of which (n=35, multiple answers possible)*



© RatSWD 2022

Figure 36 illustrates the broad range of provenance information made available by the RDCs. Which information is relevant for re-use strongly depends on the data type. The majority of RDCs publish methodology reports or transcription methods, and questionnaires. The latter are only relevant for survey studies, of course. Data source information and anonymisation concepts are also among the tools commonly made available. Descriptions of data preparation processes are sometimes published as technical papers. Data quality reports are also provided as a service at some RDCs.

Interim result – Re-useability

The barriers to using RDC metadata for scientific purposes are low. In most cases, free use is possible or possible when giving appropriate credit. This is in accordance with recommendations on the implementation of the FAIR principles in the social and economic sciences.[38] It should be evaluated in future whether it is possible to increasingly use CC licenses for re-use of RDC metadata because it would increase transparency from the user perspective and make the terms of use comprehensible to machines.

RDCs ensure transparency in data-generating processes through a variety of different documents. What they have in common is that the information is not yet integrated into the metadata schemes in a standardised way but typically has to be extracted from the documents. To date, only 14 RDCs publish their data anonymisation concepts. Since this is a stage in data preparation in which the data are often significantly altered, a transparent anonymisation concept can help users to track the changes that were made. If there are no legal reasons against publishing the anonymisation concepts, such information should be made public by the RDCs.

CONCLUSION: FAIR (meta)data?

High-quality metadata are the key to a FAIR infrastructure for research data. With their long-standing expertise in data management, RDCs have all components of a FAIR infrastructure at their disposal. Use of persistent identifiers is widespread, data are indexed in large international search portals, most RDCs use established metadata schemes, facilitate re-use of metadata through open licenses, and some already use semantic technologies for knowledge representation. The further development of the data infrastructure can be built on these metadata practices and technologies. It is important to note that there is a high degree of professionalism in metadata management. Overall, however, the practice is still highly fragmented as indicated by the absence of a common technical solution for metadata management (neither have the RDCs developed one, nor are commercial products used).

Fragmentation is not necessarily a problem in itself since metadata should be oriented towards standards and the needs of the respective communities. Consequently, harmonisation of metadata is particularly useful within communities (much like, e.g., Verbund Forschungsdaten Bildung has been doing for education data). Harmonisation of metadata via commonly used tools (e.g., metadata editors) could help improve interoperability.

Based on the present survey, it is not possible to come to a generalised conclusion whether additional measures for improving FAIRness at RDCs are called for. What is certain is that larger investments into the implementation of programming interfaces, granularity of data documentation, and data content indexing are necessary. Such investment decisions can only be made against the background of concrete usage scenarios. There is still a lack of concrete requirements from the user communities, particularly for automated (meta)data access as well as linking data.

Therefore, the recommendations developed in this special chapter relate primarily to those aspects that enable RDCs to close relevant gaps in information on data with manageable effort.

38 Betancort Cabrera, N., Bongartz, E. C., Dörrenbächer, N., Goebel, J., Kaluza, H., & Siegers, P. (2020). *White Paper on implementing the FAIR principles for data in the Social, Behavioural, and Economic Sciences*. RatSWD Working Paper 274/2020. Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.60.

8 Complaints management

One of the key tasks of the RatSWD is to assure and improve the quality of RDC services. Since its inception, the RatSWD has acted as a dedicated point-of-contact for complaints relating to RDC data and services. In addition to overseeing the annual monitoring process, of which the present activities report is one outcome, the RatSWD's monitoring commission also handles complaints put forward by research data users.

The RatSWD set up a complaints office at the RatSWD business office to professionalise complaints management and make it more transparent. The complaints office ensures a swift and professional response to complaints and feeds the results back into RDC processes to further improve the data infrastructure.

If data users become aware of major shortcomings in the data services of an accredited RDC, it is recommended they first approach the RDC directly to try to find a solution. If the problem cannot be resolved, users may direct their concerns to the complaints office. The complaints office's mandate is limited to issues concerning compliance with the RatSWD accreditation criteria. The RatSWD is not responsible for delays during everyday procedures or for staff conduct at RDCs. Complaints of this nature should be directed to the RDC in question.

For more detailed information about the procedures, see the RatSWD Output 8 (5)[39] or the updated version of that output on the German Data Forum (RatSWD) website.[40]

Current complaints procedures in the 2020 reporting year

The RatSWD received no complaints during the 2020 reporting year.

39 RatSWD [German Data Forum] (2018): *The German Data Forum (RatSWD) and Research Data Infrastructure: Status Quo and Quality Management*. RatSWD Output 1 (6). Berlin: German Data Forum (RatSWD). https://doi.org/10.17620/02671.30.

40 https://www.konsortswd.de/en/datacentres/monitoring-and-complaints-management/complaints-office/ (Last retrieved: 31/01/2022).

Appendix

Appendix A: Development of the RatSWD's research data infrastructure and RDCs 50
Appendix B: Index and categories of data of the research data infrastructure of the RatSWD 54
Appendix C: The monitoring commission 60
Appendix D: Contributors to the 2020 Activities Report. 61

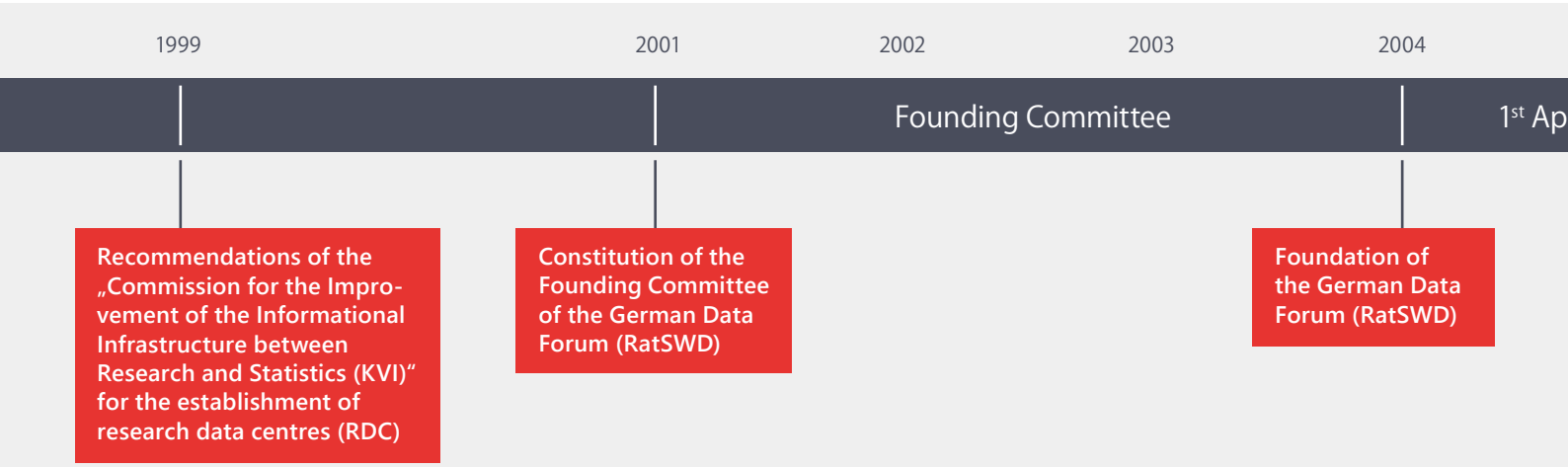
Appendix A

Development of the RatSWD's research data infrastructure and RDCs

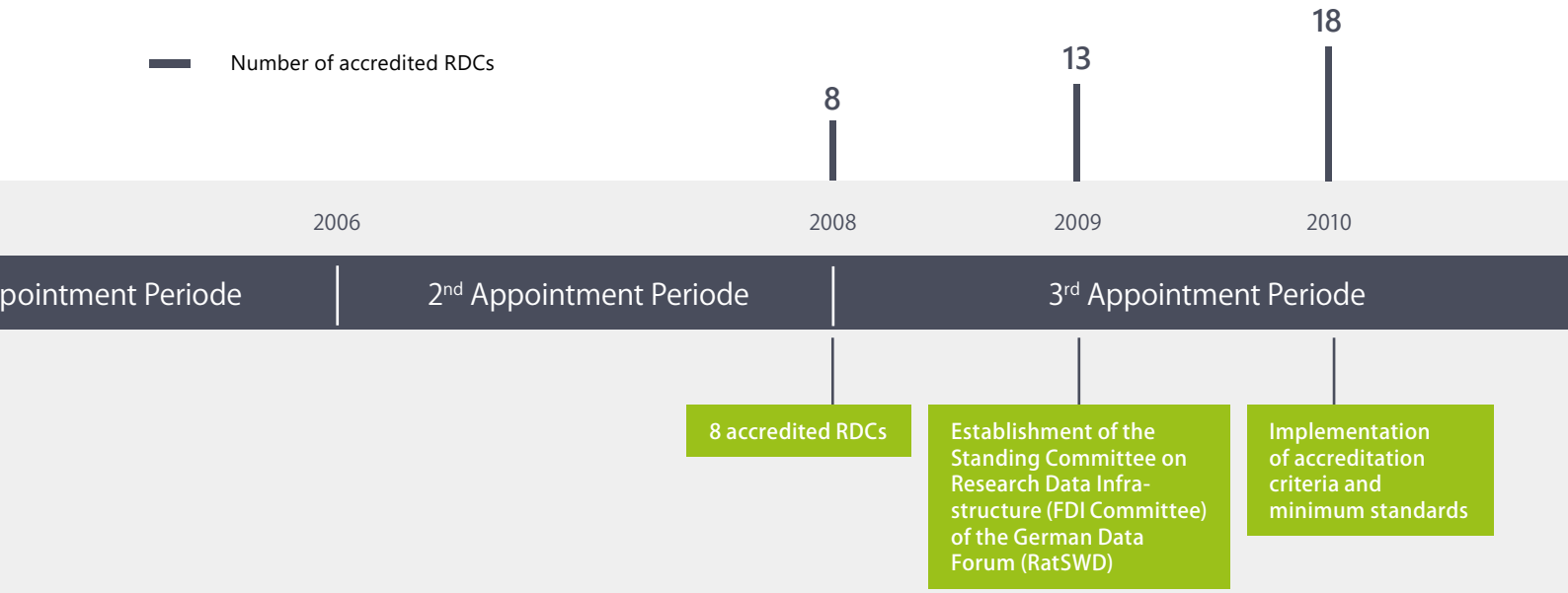
Last update: December 2021

The research data centres Federal Statistical Office, Statistical Offices of the Länder, GML, IZA, BA at IAB, and RV were established prior to the foundation of the German Data Forum (RatSWD) and became part of the research data infrastructure in 2004. In these cases, the year of the RDCs' foundation is listed. All other RDCs were accredited after 2004 by the German Data Forum (RatSWD). With these RDCs, the year provided is the year of their accreditation..

The following RDCs are founded:



The following RDCs are accredited:



The following RDCs are accredited:


DeZIM fdz Forschungs Daten Zentrum
GePaRD German Pharmacoepidemiological Research Database
Leibniz Institute of Ecological Urban and Regional Development
Qualiservice data sharing
LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe
baua: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin FORSCHUNGSDATENZENTRUM | FDZ
fdz.DZHW. Forschungsdatenzentrum für Hochschul- und Wissenschaftsforschung
eLabour
Kraftfahrt-Bundesamt
.aviDa share and reuse audio-visual data
Bundesamt für Migration und Flüchtlinge
31
32
34
39
41
2017
2018
2019
2020
2021
2022
6th Appointment Periode
7th Appointment Periode
Foundation of the KonsortSWD
41 accredited RDCs in the FDI Committee

Appendix B

Index and categories of data of the research data infrastructure of the RatSWD

Last update: December 2021

Available data: Social · Economic · Education · Health · Behavioural · Qualitative · Other

BAMF-FDZ
(provisional accreditation)

Research Data Centre of the Federal Office for Migration and Refugees

The Research Data Centre of the Federal Office for Migration and Refugees (BAMF-FDZ) will commence operations in summer 2021. Thereafter, access for research purposes to data from the Central Register of Foreign Nationals will be granted by the BAMF-FDZ. In addition, selected survey data from the BAMF Research Centre will be accessible prospectively.

https://www.bamf.de/EN/Themen/Forschung/Forschungsdatenzentrum/forschungsdatenzentrum-node.html

BIBB-FDZ

Research Data Centre of the Federal Institute for Vocational Education and Training

Firm-level and individual-level datasets of vocational education research dealing primarily with the attainment and use of vocational knowledge and skills.

https://www.bibb.de/en/53.php

DeZIM.fdz

Research Data Centre of the German Centre for Integration and Migration Research

The DeZIM.fdz organises access to research data collected at the German Centre for Integration and Migration Research. Additionally, the DeZIM.fdz offers comprehensive support on this data and on various methodological key issues.

https://dezim-institut.de/forschungsdatenzentrum-dezimfdz/

EBDC

LMU-ifo Economics & Business Data Center

Datasets of German companies, including survey data collected by the ifo Institute on firms' business status, innovativeness, and investment behaviour, as well as external data on corporate financing and governance structure. Merged panels of the aforementioned two data sources are also available.

https://www.ifo.de/en/EBDC

FDZ AGD

Research Data Center Archive for Spoken German at the Institute for the German Language

Data on spoken German in interactions (conversation corpora) and data on domestic and non-domestic varieties of German (variation corpora).

https://agd.ids-mannheim.de/index_en.shtml

FDZ-aviDa
(provisional accreditation)

Research Data Centre for audio-visual data of qualitative social research

aviDa is the research data centre (RDC) for audio-visual data of empirical qualitative social research at the Department of General Sociology at the Technische Universität Berlin, developed in cooperation between the Technische Universität Berlin and the University of Bayreuth. aviDa aims at opening and sharing videographic research data since 2018.

https://fdz-avida.tu-berlin.de/

FDZ BA at IAB

Research Data Centre of the German Federal Employment Agency at the Institute for Employment Research

Data on persons, households, and employers, as well as combined datasets consisting of survey data and administrative research data in the fields of social security and labour market, and employment research.

https://fdz.iab.de/en.aspx

FDZ-BAuA

Research Data Centre of the Federal Institute for Occupational Safety and Health

Data from cross-sectional and longitudinal studies on working and employment conditions and their effects on health, safety and well-being of workers in Germany.

https://www.baua.de/EN/Service/Research-Data/Research-Data_node.html

FDZ Bildung

Research Data Centre for Education at the DIPF | Leibniz Institute for Research and Information in Education

The hosted datasets include approaches of qualitative educational research such as video data, transcriptions, contextual materials from observations and interviews and survey tools of quantitative educational research such as questionnaires and assessment tests. The collected datasets refer mainly to the quality of instruction and to the quality of schools but also cover all levels of education throughout the entire span of life.

https://www.fdz-bildung.de/home?la=en

FDZ-BO

Research Data Centre for Business and Organizational Data

Quantitative and qualitative business, organizational data, linked employer and employee data, and data from employee and member surveys.

http://www.fdz-bo.diw.de

FDZ-Bund

Research Data Centre of the Federal Statistical Office

Germany-wide access to official statistics microdata from the following fields: population, education, health, business, agriculture, environment, administration of justice, finance, and taxes.

https://www.forschungsdatenzentrum.de/en

FDZ BZgA

Research Data Centre of the Federal Centre for Health Education

Data from nationally representative surveys, repeated at regular intervals, measuring the population's susceptibility to health education and prevention campaigns, as well as the knowledge, attitudes, and behaviour in the general population concerning the health issues addressed by BZgA.

https://www.bzga.de/home/bzga

FDZ-DJI

Research Data Centre of the German Youth Institute

Data from the surveys on children and young people growing up and the life situations of adults and families, conducted in regular intervals since 1988.

https://www.dji.de/abt2

FDZ-DZA

Research Data Centre of the German Centre of Gerontology

Data from the long-term German Ageing Survey (DEAS) on the changing life situations and ageing processes of people in mid- and older adulthood, and from the German Survey on Volunteering (FWS), a representative survey programme with a focus on voluntary activities and civic participation in Germany.

https://www.dza.de/en/research/fdz

Available data: Social | Economic | Education | Health | Behavioural | Qualitative | Other

FDZ-DZHW

Research Data Centre for Higher Education Research and Science Studies

Quantitative and qualitative research data from the field of higher education and science studies, especially the DZHW Panel Study of School Leavers with a Higher Education Entrance Qualification (Studienberechtigtenpanel), the DZHW Graduate Panel (Absolventenpanel), the DZHW Social Survey, and the DZHW Science Survey.

https://www.fdz.dzhw.eu/en

FDZ eLabour

Research Data Centre eLabour

Qualitative data from studies in industrial and occupational sociology with a focus on the changing nature of work, including open and semi-standardised interviews, observations, and expert interviews.

http://elabour.de

FDZ GePaRD
(provisional accreditation)

German Pharmacoepidemiological Research Database

The FDZ GePaRD is based on data provided by statutory health insurance providers in Germany since 2004. GePaRD can be used to investigate research questions on the utilization and safety of drugs and vaccines in routine care, provided the respective data use has been approved in accordance with § 75 SGB X.

https://www.bips-institut.de/en/research/research-infrastructures/gepard.html

FDZ IQB

Research Data Centre of the Institute for Educational Quality Improvement

German datasets from the major national and international school performance studies and national studies measuring educational standards.

https://www.iqb.hu-berlin.de/fdz

FDZ IZA, IDSC

International Data Service Centre at the Institute for the Study of Labour

National and international labour market datasets with standardised information (https://www.eddi-conferences.eu). Research with, methods and resources for using online data for labor economics and social science. Development of tools and methods for remote access (statsdirect.org) and remote processing (JoSuA).

https://www.iza.org/en/research/idsc

FDZ at KBA

Research Data Centre at Kraftfahrt-Bundesamt

With the registers kept by the KBA as the data basis, the data offered by the FDZ in the KBA currently includes quantitative microdata on access to the Register of Driver Fitness (Fahreignungsregister).

https://www.kba.de/DE/Statistik/Forschungsdatenzentrum/forschungsdatenzentrum_node.html

FDZ-Länder

Research Data Centre of the Statistical Offices of the Länder

Germany-wide access to official statistics microdata from the following fields: population, education, health, business, agriculture, environment, administration of justice, finance, and taxes.

https://www.forschungsdatenzentrum.de/en

FDZ-pairfam

Research Data Centre of the German Family Panel

Datasets from the "Panel Analysis of Intimate Relationships and Family Dynamics" (pairfam), a representative, interdisciplinary longitudinal study for the analysis of private living arrangements in Germany.

https://www.pairfam.de/en

FDZ PsychData at ZPID

Research Data Centre PsychData of the Leibniz Institute for Psychology Information

Pooled quantitative datasets from both basic research and applied psychology; data archiving with a focus on longitudinal studies, large-scale survey studies, and development testing.

https://rdc-psychology.org/

FDZ Qualiservice

Research Data Centre Qualiservice

Qualiservice focuses on archiving, curating and providing qualitative research data from a range of disciplines. Its secure, flexible, and research-oriented services include processing primary qualitative studies for secondary use, comprehensive user support, long-term preservation, and the provision of archived research data as well as relevant context information.

https://www.qualiservice.org/en/

FDZ Ruhr at RWI

Research Data Centre Ruhr at the RWI – Leibniz Institut for Economic Research

Specialisation on regional data: socioeconomic data measured by 1 square km grids. Aside from geo-referencing data on a scientific basis, the RDC provides various individual-level and employer-level data collected in RWI research projects.

https://www.rwi-essen.de/en/research-advice/further/research-data-center-ruhr-fdz

FDZ-RV

Research Data Centre of the German Pension Insurance

Data on the insurance accounts of individuals insured in the Federal Pension Insurance. The accounts contain data on the insured persons' insurance history and the pension and rehabilitation benefits they received.

https://www.eservice-drv.de/FdzPortalWeb/dispcontent.do?id=main_fdz_english

FDZ-SHARE

Research Data Centre of the Survey of Health, Ageing and Retirement in Europe

Data from the multidisciplinary and cross-national panel study "Survey of Health, Ageing and Retirement in Europe" (SHARE), which produces microdata on health, socio-economic conditions, and social and family networks of approximately 140,000 individuals in its seventh wave aged 50 or older in more than 20 European countries and Israel. The eighth wave of SHARE was collected in 2020.

http://www.share-project.org/data-access.html

IOER-Monitor

Monitor of Settlement and Open Space Development

The IOER Monitor is a service of the Leibniz Institute for Ecological Urban and Regional Development (IOER). It provides data and information on the sustainability of land cover and land use change and for the landscape quality for the whole of Germany.

https://www.ioer-monitor.de/en

RDC ALLBUS

Research Data Centre ALLBUS at GESIS

Data from the Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) and German General Social Survey (GGSS) in English, on the attitudes, behaviours, and social structure of the German population.

https://www.gesis.org/en/allbus/allbus-home

RDC Elections

Research Data Centre Elections at GESIS

Access to German national election surveys (federal elections and state elections), Politbarometer, Forsa-Bus, ARD Deutschlandtrend and Surveys for the Federal Government. The RDC's largest project at this point is the German Longitudinal Election Study (GLES).

https://www.gesis.org/en/institute/research-data-centers/rdc-elections

Available data: Social | Economic | Education | Health | Behavioural | Qualitative | Other

RDC GML

Research Data Centre German Microdata Lab at GESIS

Research based services for researchers working with microdata from European and German official statistics: tools for data management and data analysis. Metadata (MISSY): comprehensive data documentation for official microdata on a detailed level. Knowledge transfer: consulting, training, workshops and user conferences on methodological and substantive research questions in the analysis of official microdata. Established 1987.

https://www.gesis.org/en/institute/research-data-centers/rdc-german-micro-data-lab

RDC International Survey Programmes

Research Data Centre International Survey Programmes at GESIS

Internationally comparative survey data from more than 70 countries on nearly all social science topics: Comparative Study of Electoral Systems (CSES), European Values Study (EVS), Eurobarometer, European Election Studies (EES), International Social Survey Programme (ISSP).

https://www.gesis.org/en/institute/research-data-centers/rdc-international-survey-programs

RDC-IWH

Research Data Centre of the Halle Institute for Economic Research

Company data from panel studies and longitudinal studies on development trends in the manufacturing and construction sectors of Eastern Germany, on privatisation activities of the *Treuhandanstalt*, on the choice of location for multinational companies in Eastern and Central Europe, and productivity and competitiveness indicators of European countries.

https://www.iwh-halle.de/en/research/data-and-analysis/research-data-centre

RDC-LIfBi

Research Data Center of the Leibniz Institute for Educational Trajectories at the University of Bamberg

Longitudinal data from the National Educational Panel Study (NEPS), which was launched in 2010 with more than 60,000 panel participants in six starting cohorts to study skills formation, educational processes, educational decisions, and educational returns in formal, non-formal, and informal contexts across the lifespan.

https://www.lifbi.de/Institute/Organization/Research-Data-Center

RDC PIAAC

Research Data Center Programme for the International Assessment of Adult Competencies (PIAAC) at GESIS

German and international data of the Programme for the Assessment of Adult Competencies (PIAAC). For Germany, additional regional data and longitudinal data are available.

https://www.gesis.org/en/institute/research-data-centers/rdc-piaac

RDC RKI

Research Data Centre of the Robert Koch Institute

Data on the state of health and health-related behaviour of Germany's resident population, collected on the basis of nationally representative studies.

https://www.rki.de/puf

RDC SOEP

Research Data Center of the Socio-Economic Panel Study at DIW Berlin

Data from representative annual surveys of private households. The SOEP-CORE sample features topics such as income, employment, education, and health. In addition, there is the longitudinal innovative sample (SOEP-IS), which enables external researchers to contribute research projects of their own.

https://www.diw.de/en/diw_01.c.678568.en/research_data_center_soep.html

RDC Wissenschafts-statistik

Research Data Center Wissenschaftsstatistik of the Stifterverband

Data on the research and development activities of German companies, on the financial volume, structure, and regional distribution of research and development activities (R&D), and on R&D staff in the business sector

https://www.fdz-wissenschaftsstatistik.de

RDSC Bundesbank

Deutsche Bundesbank Research Data and Service Centre

Various datasets on banks, securities, investment funds and enterprises, as well as combinations of those; panel survey on household finances.

https://www.bundesbank.de/en/bundesbank/research/rdsc

SAFE RDC

(provisional accreditation)

Research Data Center of the Leibniz Institute for Financial Research SAFE

The lack of pan-European financial data means that researchers have to resort to US data and cannot easily transfer research results to the European area. The SAFE Research Data Center not only pools existing data, but also collects and creates new German and European data sets to strengthen the European perspective of empirical research.

https://safe-frankfurt.de/data-center.html

ZEW-FDZ

ZEW Research Data Centre for European Economic Research

The ZEW-FDZ provides microdata from ZEW firm surveys on innovation activities, the development of young firms, the use of information and communication technologies, and further topics. Data from individual and expert surveys are also accessible – for example, the ZEW Financial Market Survey.

https://kooperationen.zew.de/en/zew-fdz/home

Appendix C

The monitoring commission

For quality assurance purposes, the German Data Forum (RatSWD) agreed to establish a monitoring commission in July 2016. Its main task is to collect and assess the regular reports handed in by the RDCs. Moreover, the commission monitors compliance with the obligations arising from provisional accreditation. The FDI Committee elects the commission from its own membership for a three-year term concurrent with the German Data Forum (RatSWD) appointment period. The commission thus enjoys a special level of trust and legitimacy. It consists of four members of the FDI Committee and two deputy members (to replace elected members, if required) and the German Data Forum (RatSWD) chairpersons sit in as guests.

Members of the monitoring commission

Lea Eilers (until July 2021)
Research Data Centre Ruhr at the RWI – Leibniz Institute for Economic Research (FDZ Ruhr at RWI)

Dr. Benjamin Fuchs
Research Data Centre at Kraftfahrt-Bundesamt (FDZ at KBA)

Dr. Cornelia Lang
Halle Institute for Economic Research (RDC-IWH)

Dr. Laura Menze (since August 2021)
Research Data Centre of the Federal Institute for Occupational Safety and Health (FDZ-BAuA)

Dana Müller (Chair)
Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)

Holger Quellenberg
Research Data Centre of the German Youth Institute (FDZ-DJI)

Dr. Pascal Siegers
Research Data Centre ALLBUS at GESIS

Standing guests of the monitoring commission

Prof. Dr. Monika Jungbauer-Gans
Chair of the German Data Forum (RatSWD) (since July 2020)

Prof. Dr. Kerstin Schneider
Vice chair of the German Data Forum (RatSWD) (since July 2020)

Appendix D

Contributors to the 2020 Activities Report

Florence Baillet
RatSWD business office

Dr. Benjamin Fuchs
Research Data Centre at Kraftfahrt-Bundesamt (FDZ at KBA)

Dr. Anna Fräßdorf
RatSWD business office

Dr. Cornelia Lang
Halle Institute for Economic Research (RDC-IWH)

Dr. Laura Menze
Research Data Centre of the Federal Institute for Occupational Safety and Health (FDZ-BAuA)

Dana Müller
Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)

Holger Quellenberg
Research Data Centre of the German Youth Institute (FDZ-DJI)

Thomas Runge
RatSWD business office

Lea Salathé
RatSWD business office

Dr. Pascal Siegers
Research Data Centre ALLBUS at GESIS

Imprint

Publisher:
German Data Forum (RatSWD)
Business office
Am Friedrichshain 22
10407 Berlin
Germany
office@ratswd.de
https://www.ratswd.de

Editors:
Florence Baillet, Dr. Anna Fräßdorf, Lea Salathé

Layout:
Claudia Kreutz

Icons:
made by Freepik from https://www.flaticon.com

Berlin, March 2022

The RatSWD office is funded by the German Research Foundation (DFG) as part of KonsortSWD within the framework of the NFDI – project number: 442494171.

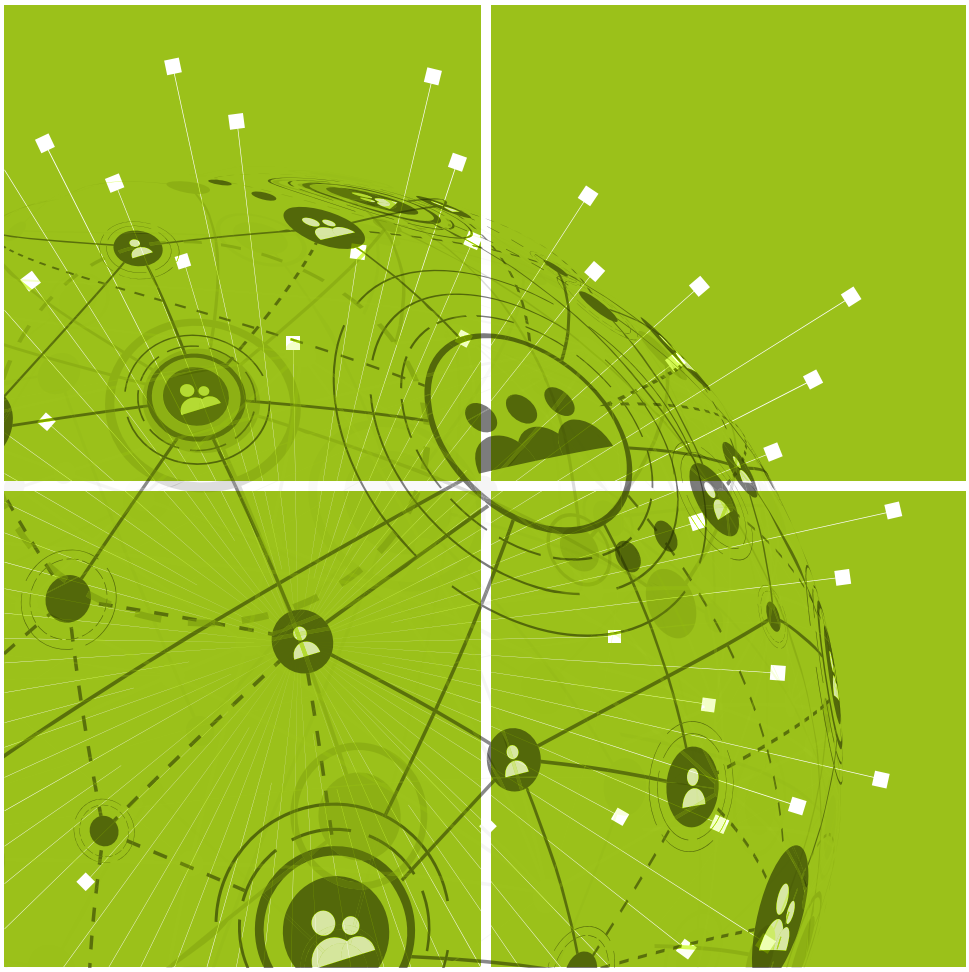
This publication is licensed under a Creative Commons Attribution 4.0 (CC-BY_4.0):
https://creativecommons.org/licenses/by/4.0/
Excluded from the above licence are parts, illustrations and other third-party material, if otherwise indicated, as well as the cover images on pages 1 and 64.

doi: 10.17620/02671.69

Citation suggestion:
RatSWD [German Data Forum] (2020): *Activities Report 2019 of the Research Data Centres (RDCs) accredited by the German Data Forum (RatSWD)*. Berlin, German Data Forum (RatSWD). https://doi.org/10.17620/02671.69.

The German Data Forum (RatSWD) advises the federal government and the governments in the federal states on expanding and improving the research data infrastructure for the empirical social, behavioural and economic sciences since 2004. The German Data Forum (RatSWD) is made up of ten elected representatives from the social, behavioural, and economic disciplines who work together with ten representatives from key data producers.

The German Data Forum (RatSWD) is part of the Consortium for Social, Behavioural, Educational, and Economic sciences (KonsortSWD) in the National Research Data Infrastructure (NFDI). It acts as an institutionalised forum for dialogue between science and data producers, as well as developing recommendations and opinions. It is committed to supporting an infrastructure that enables sciences to have broad, flexible, and secure data access. These data are provided by state, science-based, and private-sector actors. The German Data Forum (RatSWD) has currently accredited 41 research data centres (as of January 2022), and encourages their cooperation.



www.ratswd.de