DEUTSCHE
BUNDESBANK
EUROSYSTEM

# Data Production in a Digitised Age: The need to establish successful workflows for micro data access

Technical Report 2022-02

The views expressed in this technical report are personal views of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

This technical report will be published in the forthcoming book '*Advances in Business Statistics, Methods and Data Collection*.'

Deutsche Bundesbank, Research Data and Service Centre

Stefan Bender
Jannick Blaschke
Christian Hirsch

Research Data and
Service Centre

## Abstract

Access to timely and high-quality granular data is increasingly becoming a key success factor for the work of analysts and researchers as well as for political decision-making. However, a surprisingly large amount of such data remains hidden in tightly regulated silos, which means they are still underexploited. One reason lies in the nature of the data themselves, which oftentimes allow the disclosure of information about e.g. an individual person's health or a company's business model. In this technical report we present the BUBMIC framework to enable access to granular data. The framework, which stems from the authors work in the field, is organised around the three building blocks (1) Laying the technical and procedural foundations, (2) Generating safe results, and (3) Generating value for all stakeholders.

**Keywords:** statistical data production, research data centre, rdc, rdsc, micro data, value

**Version:** 1.0

**Citation:** Bender, S., J. Blaschke and C. Hirsch (2022). Data Production in a Digitised Age: The need to establish successful workflows for micro data access, Technical Report 2022-02 – Version 1.0. Deutsche Bundesbank, Research Data and Service Centre.[1]

**1** The authors would like to thank our colleagues at the Research Data and Service Centre (BBk-RDSC) for their valuable suggestions and feedback. We are particularly grateful to Tobias Schönberg for his detailed description of the BBk-RDSC's workflows, published in Schönberg (2019), which laid the foundations for Section 4 of this article. We also thank Julia Lane and her group as well as the INEXDA members for fruitful discussions that helped us sharpen our arguments and ideas. All remaining errors are our own.

# Contents

# 1 Introduction

Nowadays, empirical researchers and statisticians find themselves in a curious situation. On the one hand, data are everywhere and come from an ever-increasing number of different sources. Researchers use more data they no longer directly collect themselves (e.g. via surveys). Instead, they often analyse organic data (Groves (2011)) collected for other purposes that are now being reused, e.g. via ETL[2] or an adapted version of the Total Survey Error Approach (Biemer et al. (2017); Amaya, Biemer, & Kinyon (2020)). Many chapters in this book are evidence of how fundamentally the emergence of these new data sources has transformed the practice of social science research.

On the other hand, a surprisingly large amount of relevant data remains hidden in tightly regulated silos, which means they are underexploited by empirical research and statistics (e.g. Sachverständigenrat zur Begutachtung der Entwicklung im Gesundheitswesen (2021)). One reason lies in the nature of the data themselves, which oftentimes allow the disclosure of information about an individual person's health or a company's business model. Initiatives like the FAIR data principles (Wilkinson et al. (2016); Collins et al. (2018)) or the reproducibility standards of the AEA (Vilhuber (2019); Vilhuber (2021)) are a direct reaction to this situation and an attempt to build bridges to these silos by capturing best practices for research data use.

At the same time, there are strong movements for granting access to official micro data for the public good. For example, in the US, the Foundations for Evidence-Based Policymaking Act requires each statistical agency to produce and disseminate data while ensuring that data are used only for statistical purposes and that the confidentiality of the data is protected (Lane (2020)). Germany is heading in a similar direction, as evidenced by the work of the German Data Forum[3], the activities around the National Research Data Infrastructure,[4] or the recent announcement about opening new institutes to access new administrative data sources such as tax data.

However, providing access to micro data is quite a complex endeavour. The challenge comes from protecting the identity of the reporting entities while simultaneously allowing data users to study distinct features of these entities, e.g. effects on an individual person's health or a company's business model. In order to master this balancing act between increased confidentiality and the high analytical value of micro data, data providers need to implement statistical, organisational or technical measures to prevent the disclosure of sensitive information. However, how can this be done without placing too much of a constraint on the analytical value? In addition, how does the choice of measure interact with other aspects of data access?

To resolve this tension, we present a simple model in Section 2 of this paper that provides a guideline for evaluating proposals on how to successfully enable granular data access. The first building block of our model considers the technical and procedural requirements, while the second looks at safe output. Lastly, the third building block represents the value for stakeholders from providing access to granular data. The first two building blocks of our framework mirror existing models of knowledge generation and statistical production (Blanc, Radermacher, & Körner (2002); Radermacher (2020)). However, unlike these models we introduce "generating value" as an additional and, in our opinion, crucial condition for successful approaches to data access.

---

**2** ETR stands for Extract, Transform, Reload.
**3** For more information on the German Data Forum, see https://www.konsortswd.de/en/ratswd/
**4** For more information on the National Research Data Infrastructure, see https://www.nfdi.de/en-gb

This view fits well into recent developments (Ritchie (2016); Lane (2020)). Historically, the emphasis has been on avoiding the risk of identification. The "five safes" approach (Desai, Ritchie, & Welpton (2016); Ritchie (2017)), which provides an excellent framework for assessing and managing risk, bears testimony to this. Although this approach is widely used, especially in the public sector, the focus has slowly started shifting towards a more balanced approach in recent years[5)]. Risk is increasingly perceived as a binding constraint in an optimisation problem that aims at maximising stakeholder value[6)].

According to this view, data providers balance stakeholder value and identification risk by choosing the level of risk that they are willing to tolerate. In this context, a key success story in accessing administrative granular data has been the introduction of research data centres (RDCs) (Ritchie (2021)). RDCs are restricted-access facilities, often at the premises of the data owner, that provide accredited researchers with access to sensitive granular data.

This paper is organised as follows. The next section introduces the three building blocks for successful workflows enabling access to micro data (BUBMIC model), these being: 1. laying the technical and procedural foundations, 2. generating safe results, and 3. generating value for all stakeholders. Section 3 briefly discusses the concept of FAIR data. Section 4 applies the BUBMIC model to RDCs and shows why they are so successful. Section 5 concludes our paper.

**5** See Ritchie (2017)) for a discussion on the reasons for this.
**6** This view should not be mistaken as taking risks lightly. Readers with a background in economics will confirm that binding constraints determine the level of value attainable in the optimisation problem.

# 2 Building blocks for successful workflows enabling access to micro data

In this section, we present a simple model for the main BUilding Blocks for enabling MICro data access (BUBMIC model). Our framework is based on existing models of knowledge generation and statistical production (Blanc, Radermacher, & Körner (2002); Radermacher (2020)). Consistent with this literature, we organise the BUBMIC model's six key components in a circular process chart.

Figure 1 depicts these components grouped into the three broader building blocks:1. laying the technical and procedural foundations, 2. generating safe results, and 3. generating value for all stakeholders. Our contribution is to combine existing models from the previous literature (building blocks 1 and 2) with the idea of generating stakeholder value (building block 3). We will now discuss each category in greater detail.

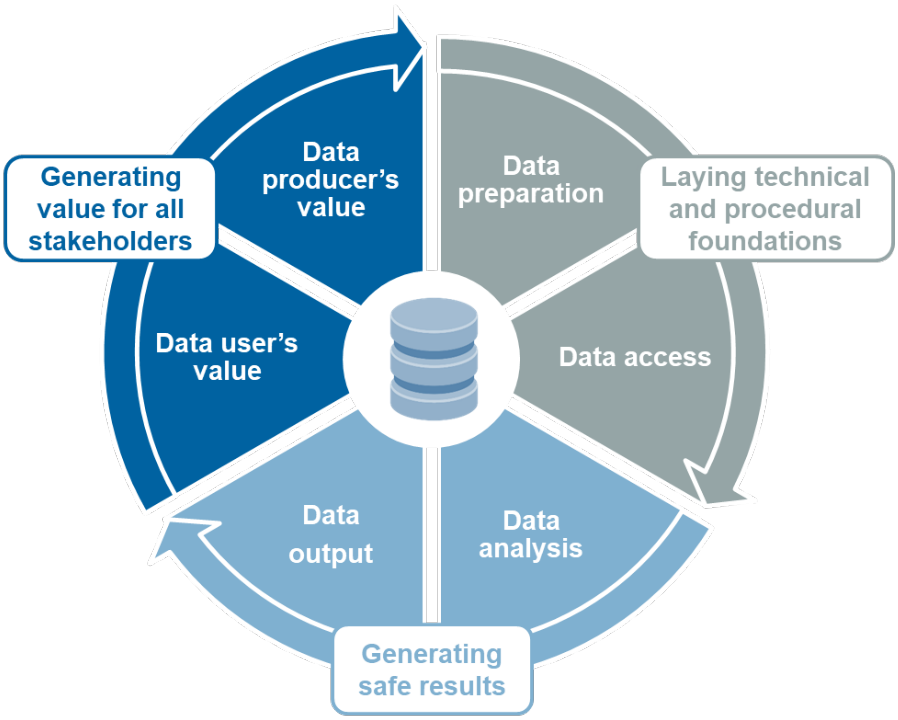


Figure 1: Building blocks to design workflows enabling access to micro data in the BUBMIC model

## 2.1 Building block 1: Laying the technical and procedural foundations

The first building block concerns the foundation of micro data access and combines data preparation and data access. We find that preparing data for usage usually involves four main decisions. First, the data provider needs to specify how the respective data can be transferred from the production system to the analytical environment in which the user can then access the data. Here, a specific emphasis needs to be placed on the reproducibility of later outputs (Vilhuber (2020)).

Second, since micro data are prone to disclosure risk, access is usually only granted to an anonymised version. The decision on the degree of anonymisation (e.g. full, formal or factual[7]) depends on different factors, such as the design of the analytical environment (e.g. remote access vs. secure on-site access[8]) and the risk preferences of the data-providing institution. For establishment data, anonymisation may need to go beyond removing direct identifiers such as names and addresses, as other variables in the data may also allow firms to be identified (Lane (2020)). For example, simply sorting data by firm size (e.g. total assets) may identify the largest firms in a given sector or region.

Third, unlike macro data, micro data allow information to be linked at the level of the individual entity. As linking data greatly enhances their analytical value for users, data providers should prepare their micro data in a way that facilitates their linkage at a later stage. Note that if linking different micro datasets is allowed, applying further anonymisation steps might be needed to ensure that no entity is identified on the basis of the enriched information.

So far, all decisions have involved manipulating the actual data in one way or another. However, preparing data for usage also involves describing the data appropriately – after all, "Data without metadata is just stuff. Nobody needs more stuff today." (Recker (2014)). Therefore, as a fourth decision, data providers need to assess which metadata standard best fits their data. In general, metadata standards[9] are often domain-specific or targeted at specific applications, and constitute very granular and semantically rich statements that provide important context to the data. Special emphasis should be placed on an unambiguously and machine-readable taxonomy, e.g. by utilising controlled vocabulary such as a thesaurus or keywords.

In addition, data providers need to pay particular attention to aspects that are specific to micro data, such as a clear description of entities and identifiers to support data linkage (INEXDA (2020)). This involves recording any available identifiers (e.g. country tax number, LEI, ISIN) that help identify the individual firm, bank or household, for example. In the case of establishment data, this may include drafting a uniform language on how to differentiate individual firms from company groups.

Designing a metadata schema for micro data also requires the documentation of structural breaks in the data that occur over time, e.g. due to changes in underlying rules for data reporting. Consequently, all metadata items should be time-dependent. For example, consider that in 2012 a legislative change altered the reporting population in the data from A to B. Therefore, metadata items providing information on the data's reporting population should cover both time periods: before 2012 the appropriate value is A, after the structural break in 2012 the value is B. Other examples of structural breaks include changes to the set of variables collected and changes in sampling.[10]

Traditional metadata largely describe data from the viewpoint of a data-providing institution

**7** We use the following definitions from INEXDA (2020): Non-anonymised: no anonymisation (i.e. raw data). Formal anonymisation: identifiers, names, and addresses deleted. No direct identification possible (i.e. scientific use files). Factual anonymisation: data perturbation. Identification possible only with significant effort (i.e. secure use files). Full anonymisation: no identification possible (i.e. public use files).

**8** We use the following definitions from INEXDA (2020): Secure on-site access: provision at the premises of the institution in a dedicated secure environment. Remote access: researchers can access data remotely from their own institution.

**9** Examples of common used metadata standards are the Data Documentation Initiative (DDI), the Dublin Core Metadata Initiative (DCMI), the DataCite Metadata Schema, the da|ra Metadata Schema, the Metadata Encoding and Transmission Standard (METS), Preservation Metadata Maintenance Activity (PREMIS) and the Statistical Data and Metadata Exchange (SDMX).

**10** For a full example of how to account for this, see the INEXDA Metadata Schema (Bender, Hausstein, & Hirsch (2019)).

(e.g. sampling and population). However, it is also recommended to add metadata on previous data usages as a way of helping users discover new data that might be appropriate for their research or analysis. For example, users could learn how and by whom the data have been used previously, and what other data sources were used in those research or analytical projects. One means of obtaining structured information on data usage could be recommendation systems that take empirical research papers as a source of information (Lane (2020)).

In addition, data providers should decide whether to provide standardised datasets to data users. Standardised datasets are characterised by a pre-selection of variables as well as refined data quality checks and comprehensive documentation. While standardised datasets might limit data users' flexibility when it comes to variable selection, they bring many advantages that cannot be achieved otherwise.

First, standardised datasets allow a perfect match between the actual data and the corresponding documentation, e.g. in the form of metadata on the variables and codes included (e.g. sector, country or currency codes), best practices for safe usage, or as descriptions of applicable access requirements.

Second, the standardisation of datasets enables their unique identification in the final research results. With the help of unique identifiers (e.g. digital object identifiers (DOIs)[11)]), data users can reference the exact data cuts, and readers of the research results can find them later and potentially reproduce the results. For this, standard datasets should be fixed snapshots that do not change over time. This implies that in the event of a data update, e.g. when data become available for a new month, a new standard dataset should be created containing the same information as the previous one plus the new data and potentially some revisions of old data points.

Third, standardisation eases data linkages. On the one hand, this refers to the technical feasibility of the linkage, as data providers know exactly which IDs are included in the datasets and can therefore provide consistent mapping tables. On the other hand, data providers can link the corresponding access rules directly to both the individual datasets and the linkage. For instance, it might be allowed to use two datasets individually, while a combination might cause a breach of confidentiality due to an identification of the underlying entities.

Finally, the decision to standardise data really is a decision on the degree of automation in work-flows to access micro data. For instance, access procedures could be largely automated if the respective access requirements can be clearly assigned to the requested dataset. We will discuss a specific use case of this in Section 4.4.5, where we present the Annodata schema.

Once the data-providing institution has decided on data preparation, it should start designing suitable data access pathways, which brings us to the second part of the first building block in Figure 1. This requires decisions on a number of aspects, including paperwork (e.g. applications and contracts), the governance of the approval process, the mode of data access (e.g. secure on-site or remote access), and the degree of data anonymisation available to each type of data user (e.g. internal, external). Rules and procedures related to data access should ideally be linked directly to the standard datasets and be "publicly available, transparent, and universally applicable as part of the metadata" (Cabrera et al. (2020)).

**11** For more information on DOIs, see https://www.da-ra.de/get-a-doi

For example, one generally observes a trade-off between the degree of anonymisation and the mode of data access, in that the greater the risk of identifying individual entities in the data, the more restrictive the access mode will be. In Section 4, we discuss the example of a research data centre (RDC) and its theoretical underpinning of the "five safes" approach (Desai, Ritchie, & Welpton (2016); Ritchie (2017)) in greater detail.

## 2.2 Building block 2: Generating safe results

Once access is granted, data users will work with the data and metadata and produce their own results. If data access has only been possible under specific technical or organisational conditions, the data provider will most likely also regulate how results are obtained. For example, if data access has been on-site in a secure environment at the data provider's premises, it may be required that all results leaving the secure environment must be fully anonymised[12)]. Usually, data providers equip users with a detailed description of rules (e.g. adherence to minimum sample size) and ways to check compliance of their results[13)]. In addition, data providers might conduct an output control for each result that users want to take out of the controlled analytical environment.

We recommend that readers interested in the discussion surrounding the assessment and management of risk use the "five safes" framework like many before them. We note that implementing approaches to micro data access naturally require these components to be discussed from a variety of different perspectives (e.g. legal, technical and organisational). Readers interested in any of these perspectives are encouraged to refer to Desai, Ritchie, & Welpton (2016), Ritchie (2017) or Schönberg (2019).

## 2.3 Building block 3: Generating value for all stakeholders

The four components we have presented so far – data production, access, analysis and output control – all describe the knowledge-generating process of micro data (e.g. Blanc, Radermacher, & Körner (2002); Radermacher (2020)). However, the third and final building block adds another very important prerequisite for designing micro data access workflows that goes beyond those classic approaches: the generation of value for all stakeholders involved. For simplification, we will assume that no intermediary such as a data trustee is involved in the provision and process of accessing micro data.

Generating value for data users is generally rather easy to motivate as it correlates strongly with the data's analytical value. Data users generate results in the form of scientific publications or analytical reports that are valuable for themselves, their community, or their employer. For example, the results might provide fundamental insights for policy decisions that in turn will serve the public interest. In the private sector, start-ups might use existing data sources to train algorithms without having to incur the cost of building up a dataset of their own. In the latter case, analytical value also corresponds to business value.

**12** Fully anonymised results can be identified neither directly (e.g. based on a name or address or an officially issued identification code) nor indirectly through deduction, accounting for all the means that might reasonably be used by a third party.
**13** See Research Data and Service Centre (2021) for an example of principles and rules applicable to visiting researchers.

We can conclude that once the value to this data user exceeds their costs of using the data, e.g. the effort involved in applying and actually implementing the data access process, data users will benefit from the opportunity to access the micro data. Also, note how the data provider's decisions described above can affect both the data users' value and cost. For instance, adequate metadata and uniquely identified data reduce the costs of data usage, while requiring (physical) on-site access at the data provider's premises increases them significantly.

Generating value for data providers is far more complex. The usual approach for private sector data provision follows the business model of commercial data vendors (e.g. Bloomberg), where the data user has to purchase access to the data. However, it is questionable whether this business model will also work if the data are of business value to both the data provider and the data user, e.g. a start-up operating in the same sector. If the data provider's data are of sufficient interest that data users are prepared to pay for them, it is likely that they are associated with a direct competitor. Hence, it is unclear whether the data producer would share the data even if they were remunerated for doing so.

In the case of public data providers, value is not usually generated through remuneration but purely through the generated knowledge itself. However, as it remains extremely difficult to measure the additional knowledge generated by data sharing (e.g. in an RDC), a more systematic approach needs to be explored. Figure 2 attempts to present different possible outcomes from data sharing and shows the corresponding channels that are organised around the first two building blocks of our model (i.e. knowledge generation or statistical production).

First, providing access to micro data may help enhance the allocation of resources to improve data that are of most value to scientific research and statistics. Second, describing data by their use instead of how they are produced help identify appropriate data for an analytical or research project. Finally, the impact of micro data usage by researchers or statistical agencies on policy decisions should be measured. Note that this latter topic can only be addressed if solutions have already been found for the previous two aspects.

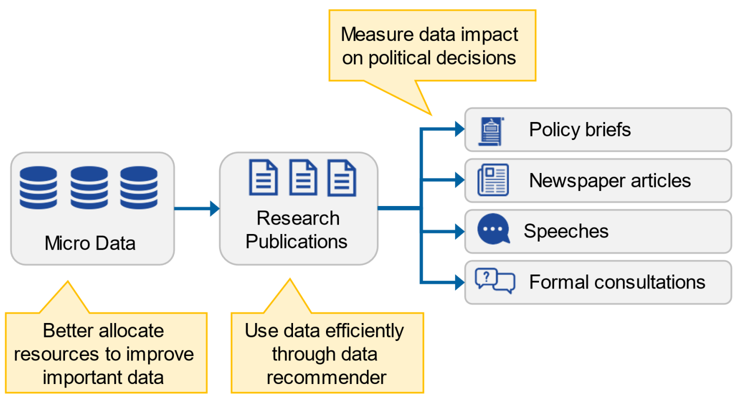


Figure 2: The ideal use of (sensitive) micro data for evidence-based policy making

Moreover, the results of data users' work could provide valuable insights into the analytical potential of the data or possible data quality issues. However, although most public data providers grant data access for scientific research purposes, data users seldom report their results back to them. As Lane (2020) observed, this lack of precise information about research outcomes leads to a situation where public data providers are not able to communicate the societal value of their service.

As a possible way to overcome this, she proposes a data platform on which data providers search for their data and find research results based on them. Specifically, Lane (2020) suggests a knowledge graph that connects results, data and data users so that one could identify data experts and recommend bundles of datasets that are often used together.

Prerequisites for this would include a sufficiently large body of research articles, a common understanding of the datasets included, and user engagement to provide feedback and the validation of suggestions. Importantly, the public, too, would benefit from this approach as they would learn what has been done with the data, increasing public awareness of their true value. Eventually, they could make an informed decision about whether the investment in the data has been worth the cost.

In conclusion, data providers who are not legally obliged to grant access to their data only do so voluntarily if their value exceeds the cost of data provision (data preparation, data access management and potentially output control) and the expected costs of a breach of confidentiality. As before, all variables in this maximisation problem are affected by the choices the data provider made in the other two building blocks in Figure 1. For example, the expected costs of a confidentiality breach decrease when stricter data protection measures as described in the "five safes" framework are implemented (e.g. only allowing secure on-site access to anonymised data followed by an output control). Similarly, the progressive automation of tasks related to data provision, such as tools to streamline output control or the generation of contracts, reduces the data provider's costs.

Here, it is important to understand that the cost functions of the data provider and data user are inverse, at least for the most part. For example, stricter access rules and the application of anonymisation and data perturbation techniques significantly reduce both the risk of identification as well as the data's analytical value.

# 3 An alternative approach to measuring value: FAIR data

One existing and widely used approach to measuring the value of data objects (e.g. data, metadata, code and the derived results) are the FAIR guiding principles (Wilkinson et al. (2016); Collins et al. (2018))[14]. Findable, Accessible, Interoperable and Reusable (FAIR) data objects facilitate data management and benefit the entire academic community, supporting knowledge discovery and innovation (Cabrera et al. (2020)). We argue that sound implementation of our three building blocks will generate data objects that generally adhere to all four FAIR principles.

The BUBMIC model ensures findability through accompanying and broadly available metadata as well as the use of unique identifiers (e.g. DOIs). This allows not only humans but also machines to find and clearly identify data objects. Moreover, we recommend linking different metadata objects to further support their findability. For example, metadata on a standard dataset could contain relations to previous versions of the same standard dataset or direct links to publications that are based on it (Bender, Hausstein, & Hirsch (2019)).

The second part of the BUBMIC model's first building block describes the importance of defining and communicating clear rules and prerequisites for data access. How this is implemented in detail very much depends on the data requested, the access modes available and the data user (e.g. internal or external data user). However, one thing that all approaches have in common is that "roles, rights, and responsibilities of all agents involved in data processing and the data access process have to be documented and communicated transparently" (Cabrera et al. (2020)).

The third FAIR dimension refers to the interoperability of different data. Enhancing analytical value by linking datasets is becoming more and more crucial, as it is also one of the main reasons to apply for micro data access rather than use what are often publicly available macro data. In Section 2.1 we recommend that data providers prepare their data in a way that supports semantic interoperability (e.g. via common identifiers or the provision of matching tables).

Lastly, our module discusses steps to ensure the reusability of data, particularly the importance of standard datasets and comprehensive metadata describing the content of each standard dataset. This is especially important for data that were originally collected for a specific use case (e.g. conducting a survey for a research project) and, therefore, might not have been properly documented in the first place. In addition, we have highlighted the importance of unique identifiers. Used properly, these allow data users to link their results to the respective standard dataset as well as clear and transparent data access rules, which should be linked directly to the relevant standard datasets. This would enable future data users to replicate the results, as they can now comprehend the exact data content as well as possible ways to access the data.

In conclusion, data objects that adhere to the FAIR principles generate significant value for both data users and data providers. While data users can work more easily with the data, data providers benefit above all from increased trust in the data they offer and the results calculated from those data. Stakeholders on both sides can capitalise on a significant transfer of knowledge and on a more efficient data access procedure.

**14** Many initiatives already focus on the FAIR guiding principles. For more details, see (Cabrera et al. (2020)) or https://www.go-fair.org/

# 4 Applying the BUBMIC model to research data centres

In this section, we explain how the three building blocks from the BUBMIC model have been applied in the context of RDCs. RDCs are a well-established means of facilitating access to administrative micro data for scientific research purposes in the public sector.[15] The existing literature typically uses the dimensions from the "five safes" framework (safe projects, safe people, safe settings, safe data and safe output) to describe an RDC's approach to safe data access. This framework has come to be widely recognised as the standard model for implementing data access in the public sector (Ritchie (2021)).

The "five safes" framework has a clear emphasis on assessing and managing the risk of confidentiality breaches. However, as described in the previous section, data providers balance the costs of expected confidentiality breaches and the effort involved in providing data against their expected value. We therefore introduce the data provider's value as a second counterweight alongside the risk-minimising measures described in the "five safes" framework, which we will refer to often on account of its paramount importance.

We use the Research Data and Service Centre of the Deutsche Bundesbank (BBk-RDSC) as a concrete real-world example. The BBk-RDSC manages access to confidential economic and financial micro datasets for scientific researchers. It also serves as an intermediary between the Bundesbank's various departments collecting and producing different micro datasets and the data users. It advises researchers on data selection, data access, data content and appropriate analytical approaches for their micro datasets.

For the purpose of this paper, we focus on the BBk-RDSC's approach to generating stakeholder value while minimising their costs. Readers interested in a more detailed description of the BBk-RDSC are warmly encouraged to refer to "Data Access to Micro Data of the Deutsche Bundesbank" (Schönberg (2019)).

## 4.1 Building block 1: Laying the technical and procedural foundations

The first part of the BUBMIC model is "data preparation," which is aligned with the "safe data" dimension of the "five safes" framework. The BBk-RDSC applies different anonymisation techniques to its data, ranging from removing direct identifiers such as names and addresses (formal anonymisation) to full anonymisation that renders it impossible to identify entities. In general, original data are most valuable to data users, followed by the de-identified version of those data Lane (2020).

Where applicable, the BBk-RDSC refines data to standardised research datasets containing preselected variables and comprehensive documentation[16], including detailed metadata. In comparison to the original data from the Bundesbank's data-producing business units, these data contain additional modifications to improve their usability for research purposes and ensure consistency with other datasets in the BBk-RDSC's portfolio. Each standard research dataset is uniquely identi-

**15** See Ritchie (2021) for more information on RDCs.
**16** The BBk-RDSC's data report series can be publicly accessed on its website: https://www.bundesbank.de/en/bundesbank/research/rdsc/publications/data-reports

fiable with a DOI[17], which must be used in all resulting publications. Note how this helps ensure compliance with the FAIR principles.

Moving on to the second part of the first building block, which is "data access," we evaluate the BBk-RDSC along the dimensions "safe people" and "safe projects." Access is only provided to external[18] researchers affiliated with a recognised scientific research institute who plan to use the data in the context of independent and non-commercial research projects.

In order to ensure that this is the case, the BBk-RDSC has established three measures. First, all interested researchers have to submit an application form with a detailed project description explaining why Bundesbank micro data are necessary and how they will be used. Data access may only be granted on a need-to-know basis, which means that if the research question can be answered without the micro data requested, access may not be granted.

Second, researchers must submit a recent version of their CV to demonstrate their affiliation with a recognised research institute and their non-commercial interest in the data. The BBk-RDSC further requires a minimum educational level of a bachelor's degree to ensure that applicants have appropriate statistical expertise to handle micro data.

Lastly, researchers will only be granted authorisation to access Bundesbank micro data after being briefed on their legal obligations during and after the project, and signing a contract. The contract contains a detailed description of all access requirements, obligations and possible penalties in case of misconduct.

## 4.2 Building block 2: Generating safe results

Regarding the "safe settings," the BBk-RDSC generally only provides formally anonymised data to external researchers which furthermore can only be accessed in the secure environment at the BBk-RDSC's premises. For experienced researchers, the BBk-RDSC offers controlled remote execution (CRE), where researchers can send their code to the BBk-RDSC, which runs it on the original data and returns the generated and controlled results[19]. In addition, some survey datasets can also be downloaded as scientific use files (SUFs), while fully anonymised data are available in the form of public use files (PUFs) on the Bundesbank's website.

Although researchers may have access to formally anonymised data for their research project, they can only take fully anonymised data out of the secure environment of the BBk-RDSC. In order to ensure that their calculation results (e.g. descriptive statistics or regression tables) do not

**17** For an overview of all DOIs registered by the BBk-RDSC, see https://www.da-ra.de/dara/search/search_result?v=1&q=defaultSearch%3A%28deutsche+bundesbank%29&fq=publicationAgent_facet%3ADeutsche%5C+Bundesbank&&rtList=&mdlang=en&lang=en&personal=false&widget=&widgetclient

**18** The BBk-RDSC also manages data access for in-house Bundesbank researchers and analysts, but this falls outside the scope of this paper.

**19** On request, the BBk-RDSC will send the researcher a data structure file which structurally resembles the original data but contains no real values. Together with a prototype program code provided by the BBk-RDSC, this can be used to write the researcher's own program code for performing the desired analyses. This code is then sent to the BBk-RDSC via email and used by staff to generate analytical results. These results are subject to statistical disclosure control. After they are checked, they are sent to the researcher by email. At no point does the researcher have access to the (original) micro dataset. CRE is open for researchers who have demonstrated, in a current or previous research project, that they have sufficient experience in using the micro dataset in question.

contain any directly or indirectly identifiable information on individual entities, researchers have to demonstrate that each observation in their calculation results is based on at least five entities from the underlying Bundesbank dataset and that the largest two entities do not account for more than 85% (dominance rule).[20)]

Researchers who believe that their calculation results meet these criteria submit them for clearance to BBk-RDSC staff members, who check and verify adherence to the rules once again (statistical disclosure control (SDC) or output control). With the exception of SUFs and PUFs, all results generated by researchers are subject to an SDC. The BBk-RDSC only releases calculation results that are fully anonymised and therefore considered "safe output" within the meaning of the "five safes" framework.

Once researchers have processed their calculation results for publication (e.g. as a scientific research paper, report or presentation), they are obliged to submit the publication to the BBk-RDSC for checking whether it only contains calculation results that underwent an SDC and to ensure that any Bundesbank micro datasets used are correctly cited. Results may not be published until the BBk-RDSC checker has given written approval, which may be withheld if the results intended for publication do not comply with the criteria for "safe output" outlined above. The BBk-RDSC does not, however, check the substantive plausibility of the calculation results or the final publication.

## 4.3 Building block 3: Generating value for all stakeholders

RDCs are widely viewed as a success model for data access in the public sector (Bender, Blaschke, Doll, Hirsch, & Ritchie (Mimeo)), as they create value for all stakeholders involved. First, they provide researchers with access to high-quality administrative micro data. Such data often contain information unavailable in data from commercial data vendors. Second, besides the data's sheer content, RDCs have also established measures to enhance analytical value by providing e.g. comprehensive metadata.

As a national central bank, the Bundesbank provides official statistics serving the public good. Making data accessible generates additional value for the Bundesbank in two ways. First, scientific results such as journal articles generate knowledge that might benefit both evidence-based policy decisions in the ESCB and the scientific community within the Bundesbank. Second, researchers might provide feedback on their experience using the data. This could help improve data quality or lead to a better understanding of how the data can be used.

However, granting external data users access to confidential micro data also comes with the potential risk of data leakage. Despite the BBk-RDSC's precautions ("five safes") and strict penalties in users' contracts, confidential information on individual entities might theoretically still reach the public. The second cost factor that should be carefully weighed against the value generated are the RDC's fixed and variable costs (e.g. staff costs).

For researchers, the value equation remains as described in Section 2.3. Their value mainly comes from the results they produce (e.g. publication of a novel article in a prestigious journal), while

**20** For more information, see Research Data and Service Centre (2021).

their costs depend on the amount of effort they need to spend in order to be granted data access (e.g. application process, signing a contract) and to get their research results through the SDC.

In order to make the most of its objective function, i.e. to maximise stakeholder value while minimising costs, the BBk-RDSC has implemented a number of measures that are targeted at different parts of the first two building blocks in Figure 1. The remainder of this section presents some selected examples[21].

## 4.4 Examples of generating value for all stakeholders

### 4.4.1 Rules for visiting researchers at the RDSC[22]

Access to confidential Bundesbank micro data via the BBk-RDSC is subject to strict requirements, in particular concerning output and publication control. To help researchers comply more easily with these requirements, the BBk-RDSC drew up the document "Rules for visiting researchers at the RDSC," which summarises all the applicable rules regarding guest research visits at the Bundesbank, the release of calculation results, and the approval of publications based on these results.

Inexperienced researchers in particular benefit from the clear and transparent communication of rules, which helps them acclimatise more quickly to working with confidential micro data at the BBk-RDSC. In addition, the document boosts confidence among reporting agents, such as banks and enterprises, and data producers, such as Bundesbank business units, that both researchers and the BBk-RDSC are handling the data properly. This reduces the expected costs in the event of a breach of confidentiality.

### 4.4.2 SDC packages in Stata[23] and R[24]

To help users make sure their output complies with the applicable rules, the BBk-RDSC provides software packages for Stata and R. Users can use the commands from the packages after descriptive or regression tables to assess whether publishing tables of that kind would cause a disclosure problem (e.g. in case the results are based on fewer than five different entities). Note that the commands are only semi-automatic, as users still have to use them correctly.

The two packages significantly reduce the workload for both researchers and the BBk-RDSC staff members who have to conduct the SDC. It also helps illuminate why calculation results might be rejected by the BBk-RDSC and reduces the number of interactions that are usually needed with inexperienced researchers wishing to submit their calculation results for SDC in the correct form.

---

**21** Examples of other value-enhancing measures include sample codes, detailed workflow documentation, email templates and a concept and R package to archive finished research projects.
**22** For more information, see Research Data and Service Centre (2021)
**23** The Stata ado files ("nobsdes5," "nobsreg5" and "maxrdsc") are available on the BBk-RDSC's website: https://www.bundesbank.de/en/bundesbank/research/rdsc/data-access
**24** The R package "sdcLog" is available on GitHub https://github.com/matthiasgomolka/sdcLog/issues

### 4.4.3 dobby, the BBk-RDSC's high-performance and streamlined data production pipeline[25)]

The BBk-RDSC recently started developing "dobby," an R package that helps streamline the production of standard research datasets, a process that includes cleaning, quality-testing and anonymising micro datasets. An acronym of "Data Orchestration Blueprint Based on YAML," dobby builds on the two well-tested and production-ready R packages drake (Landau (2018)) and validate (van der Loo & de Jonge (2021)).

dobby is the BBk-RDSC's attempt to unify and further automate the production codes for standard research datasets while maintaining flexibility to account for dataset-specific properties. It increases transparency on how datasets are produced and smooths the transfer of knowledge within the BBk-RDSC. Ultimately, the resulting efficiency gains will reduce the BBk-RDSC's costs for data production and thus boost the value it generates overall.

### 4.4.4 RDSC Contract Generator[26)]

Before any research project can begin, researchers have to sign a contract. The BBk-RDSC distinguishes between different contract types and asks researchers to sign a formal undertaking on the first day of their guest researcher visit at the BBk-RDSC. Due to dataset- and researcher-specific legal requirements, the different contract types vary strongly in terms of the information required and the overall level of complexity.

To reduce the time and effort spent on drafting contracts, the BBk-RDSC has developed an application that automates the creation of contracts and helps draft them in significantly less time. This RDSC Contract Generator, as it is known, was programmed in Python and contains a graphical user interface that guides BBk-RDSC staff members through the preparation of each individual contract in the most user-friendly manner possible.

### 4.4.5 Annodata schema[27)]

One of the BBk-RDSC's main responsibilities is to manage access to confidential Bundesbank micro data for scientific research purposes. To do this, it has established procedures designed to ensure ongoing compliance with all legal, technical, organisational and administrative requirements. These procedures and rules are well documented and aligned with all the affected stakeholders within the Bundesbank, such as the legal unit, IT security or the data-producing business units.

Whether or not a researcher is ultimately granted data access very much depends on (i) the data requested (e.g. degree of anonymisation or planned linkage), (ii) the characteristics of the researcher (e.g. potential commercial interest or educational background), and (iii) the selected mode of access (e.g. secure on-site access or download). Moreover, the applicable access protocols that are

**25** See Gomolka, Blaschke, Brîncoveanu, Hirsch, & Yalcin (2021) for more information on dobby.
**26** See Blaschke, Hering, & Huth (2020) for more information on the RDSC Contract Generator.
**27** See Bender, Blaschke, Doll, Hirsch, & Ritchie (Mimeo) for more information on the Annodata schema, and INEXDA (2020) for a specific Annodata use case.

required before the start of a research project (e.g. the contract type a researcher has to sign) may likewise vary, depending on these three dimensions.

As the number of datasets in the BBk-RDSC's portfolio steadily increases, along with the average number of datasets used in a single research project, it becomes more and more complex and cumbersome to keep track of all the rules and how they interact. Therefore, the BBk-RDSC started investigating how it could automate tasks and decisions associated with data access management, which eventually led to the development of the Annodata framework. Annodata (a contraction of "annotation to data") are structured metadata information needed for data access management that previously only existed in an unstructured form, e.g. in legal texts or as tangible expert knowledge. The aim of the Annodata framework is "to complement extant metadata schemas not to supersede them" (Bender, Blaschke, Doll, Hirsch, & Ritchie (Mimeo)).

The BBk-RDSC expects the use of the Annodata framework to yield four main benefits. First, the Annodata framework adheres to the FAIR principles. It creates transparency about where data can be found and how they can be accessed, which in turn facilitates both reusability and the reproducibility of previous outcomes. In addition, the Annodata framework covers linkage possibilities and potential restrictions.

Second, the availability of structured and standardised information on the requested data, the applicable access rules as well as the requesting researcher and their project will ease the automation of data access management at the BBk-RDSC. The Annodata framework places great emphasis on unambiguous, machine-readable information on the basis of which decisions can be derived deterministically.

Third, researchers will benefit from the increase in transparency surrounding the BBk-RDSC's data access management procedures, giving them a better understanding of the applicable access rules, restrictions and protocols. This improved grasp will speed up the application process and eventually lead to a higher level of trust in the objectivity of the BBk-RDSC.

Fourth, the Annodata schema provides a common taxonomy underpinning the exchange of knowledge between different data-providing institutions. In the context of the international INEXDA[28] network, representatives from different central banks, statistical institutes and international organisations discussed their approaches to data access for research purposes. As differences in the terminology and legal foundations used complicated the discussion, the INEXDA Working Group on Data Access developed the INEXDA Annodata schema[29] containing mostly controlled keywords with a clear definition. In a next step, the INEXDA Annodata schema will be used for harmonisation activities.

Overall, the Annodata framework has the potential to significantly increase the efficiency of data access management procedures and thus to reduce costs for both the BBk-RDSC as well as applying researchers. Moreover, by supporting the FAIR principles, it also boosts the value generated for stakeholders.

**28** The International Network for Exchanging Experience on Statistical Handling of Granular Data (INEXDA) supports the G20 process by providing a platform for exchanging experiences on statistical handling of granular data for central banks, statistical institutes and international organisations. For more information, see INEXDA (2018).
**29** See INEXDA (2020) for the INEXDA Annodata schema.

# 5 Conclusion

In this paper, we introduced the BUBMIC model, an approach that groups the workflow of providing access to, and the analysis of, sensitive micro data into three broader building blocks: 1. laying the technical and procedural foundations, 2. generating safe results, and 3. generating value for all stakeholders. By doing this, we are not reinventing the wheel, since building blocks 1 and 2 can be found in existing models from the literature. Our innovation is to combine these existing components with the idea of generating stakeholder value (building block 3), which in our mind is a crucial and often overlooked success factor in micro data access proposals.

To provide tangible examples for building block 3 of the BUBMIC model, we introduced developments that help data providers and data users minimise the cost function and increase the benefit of using sensitive micro data in Section 4. One example worth highlighting is the Annodata framework, which introduces structured, machine-readable metadata information that can support the automation of, and communication about, data access management (Bender, Blaschke, Doll, Hirsch, & Ritchie (Mimeo)). However, we also emphasised the need for more comprehensive discussions about all the channels through which RDCs generate value for stakeholders.

By introducing the BUBMIC model, we have clearly not resolved the peculiar situation of being flooded with (too much) data while at the same time still seeing relevant data hidden away in tightly regulated silos where they are underexploited by empirical research and statistics. We do, however, view the BUBMIC model as an important step towards building bridges to these silos to allow sensitive data to be analysed for the public good. In our mind, the key building block to these bridges is to identify the value that can be unlocked for all the stakeholders involved.

## References

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119. Oxford University Press.

Bender, S., Blaschke, J., Doll, H. C., Hirsch, C., & Ritchie, F. (Mimeo). *Can sensitive data be FAIR? – The annodata framework, Technical Report*. Deutsche Bundesbank, Research Data; Service Centre.

Bender, S., Hausstein, B., & Hirsch, C. (2019). *An introduction to INEXDA's metadata schema*. -Proceedings from the Ninth IFC Conference in Basel, Switzerland, 30-31 August 2018. Retrieved from https://www.bis.org/ifc/publ/ifcb49_31.pdf

Biemer, P. P., Leeuw, E. D. de, Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C., et al. (2017). *Total survey error in practice*. Wiley Series in Survey Methodology.

Blanc, M., Radermacher, W., & Körner, T. (2002). Quality and users. *Quality in the European Statistical System–The way forward, Luxembourg*, 45–60.

Blaschke, J., Hering, F., & Huth, L. (2020). *Document automation at the Research Data and Service Centre using the RDSC Contract Generator, Technical Report 2020-02*. Deutsche Bundesbank, Research Data; Service Centre. Retrieved from https://www.bundesbank.de/resource/blob/826288/d9d65905de96946a53a8254154fb6174/mL/2020-02-contract-generator-data.pdf

Cabrera, N. B., 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 Series, No 274/2020. Retrieved from https://doi.org/10.17620/02671.60

Collins, S., Genova, F., Harrower, N., Hodson, S., Jones, S., Laaksonen, L., Mietchen, D., et al. (2018). Turning FAIR into reality: Final report and action plan from the european commission expert group on FAIR data. Luxembourg: Publications Office of the European Union. Retrieved from https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_1.pdf

Desai, T., Ritchie, F., & Welpton, R. (2016). Five safes: Designing data access for research. Economics Working Paper Series No 1601, UWE Bristol.

Gomolka, M., Blaschke, J., Brîncoveanu, C., Hirsch, C., & Yalcin, E. (2021). *Data Orchestration Blueprint Based on YAML (dobby): Research data pipelines in R, Technical Report 2021-03*. Deutsche Bundesbank, Research Data; Service Centre. Retrieved from https://www.bundesbank.de/resource/blob/867820/9c653622c15a0c872cca613bbfba0950/mL/2021-03-dobby-data.pdf

Groves, R. M. (2011). Three eras of survey research. *Public opinion quarterly*, *75*(5), 861–871. Oxford University Press. Retrieved from https://doi.org/10.1093/poq/nfr057

INEXDA. (2018). *INEXDA – the Granular Data Network*. -Proceedings from the Ninth IFC Conference in Basel, Switzerland, 30-31 August 2018. Retrieved from https://www.bis.org/ifc/publ/ifcb49_31.pdf

INEXDA. (2020). *Final Report of the INEXDA Working Group on Data Access, Version 1.0*. International Network for Exchanging Experience on Statistical Handling of Granular Data. Retrieved from https://www.inexda.org/wp-content/uploads/2020/05/FINAL_INEXDA_WG_DA_Final_Report.pdf

Landau, W. M. (2018). The drake r package: A pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *3*(21), 550. Retrieved from https://doi.org/10.21105/joss.00550

Lane, J. (2020). *Democratizing our data: A manifesto*. MIT Press.

Radermacher, W. J. (2020). *Official statistics 4.0: Verified facts for people in the 21st century*.

Springer. Retrieved from https://doi.org/10.1007/978-3-030-31492-7

Recker, A. (2014). *Was sind Daten, Forschungsdaten, Metadaten? Forschungsdatenmanagement in den Geowissenschaften*. Retrieved from https://www.slideshare.net/GESISarchivtraining/was-sind-daten-forschungsdaten-metadaten

Research Data and Service Centre. (2021). *Rules for visiting researchers at the RDSC, Technical Report 2021-02 - Version 1-0*. Deutsche Bundesbank, Research Data; Service Centre. Retrieved from https://www.bundesbank.de/resource/blob/826176/ffc6337a19ea27359b06f2a8abe0ca7d/mL/2021-02-gastforschung-data.pdf

Ritchie, F. (2016). Can a change in attitudes improve effective access to administrative data for research? University of the West of England, Economics Working Paper Series No 1607.

Ritchie, F. (2017). The 'five safes': A framework for planning, designing and evaluating data access solutions. Data for Policy 2017: Government by Algorithm? (Data for Policy). Zenodo.

Ritchie, F. (2021). Microdata access and privacy: What have we learned over twenty years? *Journal of Privacy and Confidentiality*, *11*(1). Retrieved from https://doi.org/10.29012/jpc.766

Sachverständigenrat zur Begutachtung der Entwicklung im Gesundheitswesen. (2021). *Digitalisierung für Gesundheit: Ziele und Rahmenbedingungen eines dynamisch lernenden Gesundheitssystems*. Retrieved from https://www.svr-gesundheit.de/fileadmin/Gutachten/Gutachten_2021/SVR_Gutachten_2021.pdf

Schönberg, T. (2019). *Data Access to Micro Data of the Deutsche Bundesbank, Technical Report 2019-02*. Deutsche Bundesbank, Research Data; Service Centre. Retrieved from https://www.bundesbank.de/resource/blob/801044/6484d9e4aa2be3610b7378a48a1916de/mL/2019-02-data-access-data.pdf

van der Loo, M. P. J., & de Jonge, E. (2021). *Data validation infrastructure for R*. *Journal of Statistical Software* (Vol. 97, pp. 1–31).

Vilhuber, L. (2019). Report by the AEA data editor. *AEA papers and proceedings* (Vol. 109, pp. 718–29).

Vilhuber, L. (2020). Reproducibility and replicability in economics. *Harvard Data Science Review*, *2*(4).

Vilhuber, L. (2021). Report by the AEA data editor. *AEA papers and proceedings* (Vol. 111, pp. 808–17).

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1–9.