A Practical Use Case: Lesson Learned From Social Science Research Data Centers

Stefan Bender, Jannick Blaschke[1], and Christian Hirsch (Deutsche Bundesbank)

Bender, S., Blaschke, J., & Hirsch, C. (2024). A Practical Use Case: Lesson Learned From Social Science Research Data Centers. Harvard Data Science Review, (Special Issue 4). https://doi.org/10.1162/99608f92.8a2f4507

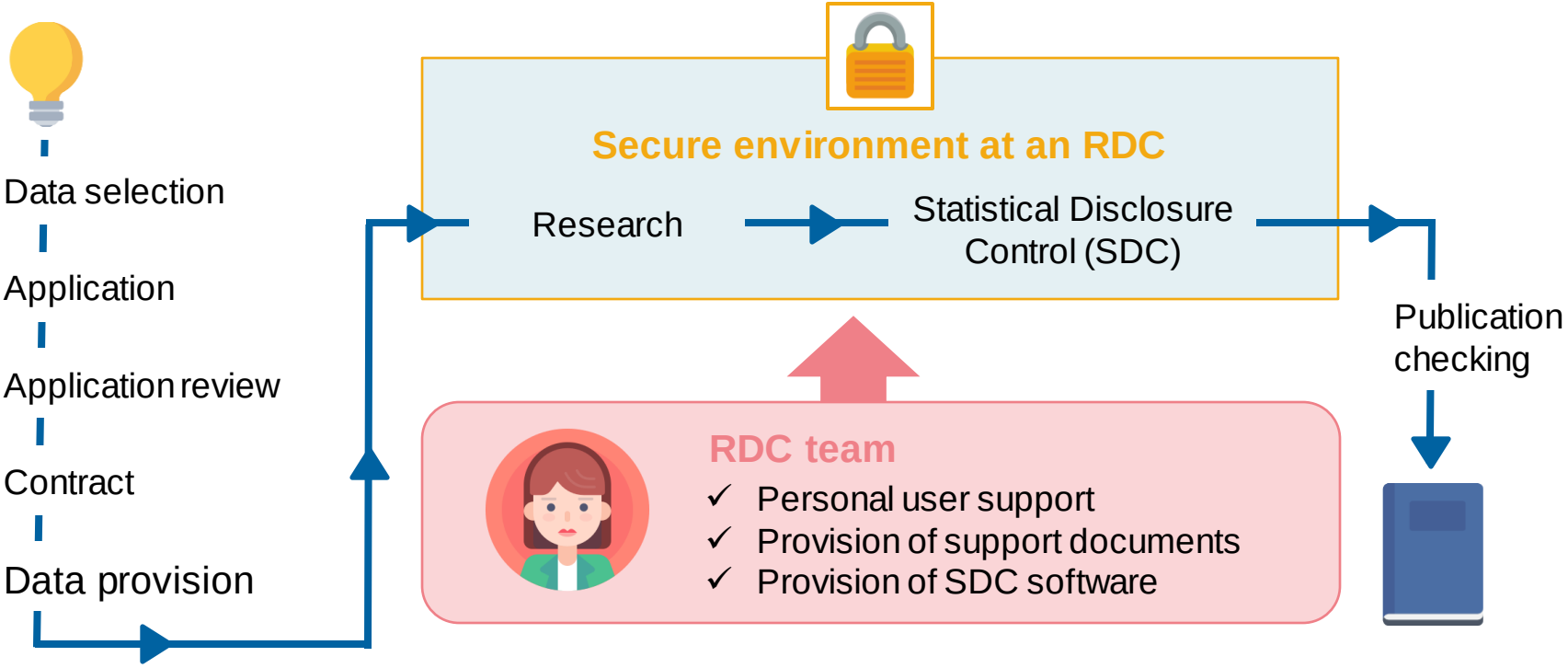
[1]The paper was completed while Jannick Blaschke was at the Deutsche Bundesbank.

Motivation

- Access to timely and high-quality granular data is increasingly becoming a key factor for research and evidence-based policy-making.
- For accessing confidential administrative data, the introduction of research data centers (RDCs) has been a success story.
- Successful data sharing approaches need to strike a balance between costs and benefits for all stakeholders. Trust is needed from all stakeholders, too.

A brief introduction to the work of Research Data Centres (1|2)

RDCs provide secure on-site access to confidential micro data for scientific research



Research Data and Service Centre

A brief introduction to the work of Research Data Centres (2|2)
Deutsche Bundesbank's RDC

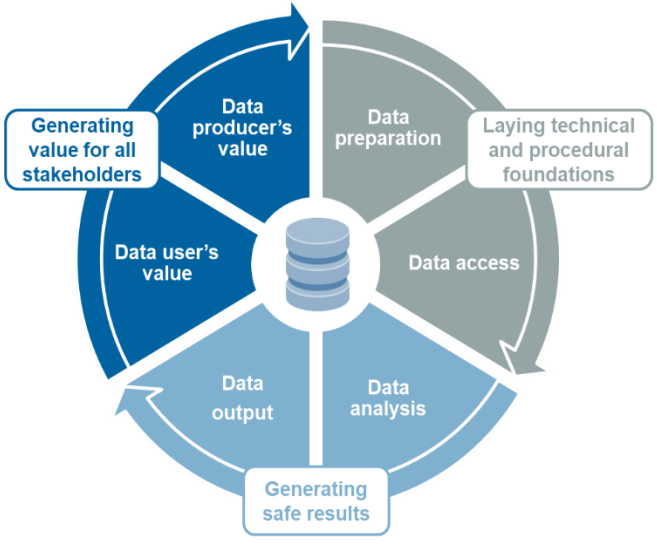
The **Research Data and Service Centre** (RDSC) of the Deutsche Bundesbank offers **free** access for **non-commercial** research to (highly sensitive) **micro data** of the Bundesbank.

Microdata for banks, companies, securities and households are available:

- Generate (standardized) (linked) micro data
- Offer advisory service on data selection and data access
- Provide data access and data protection
- Document data and methodological aspects of the data
- Work on own research projects
- Organize conferences and workshops

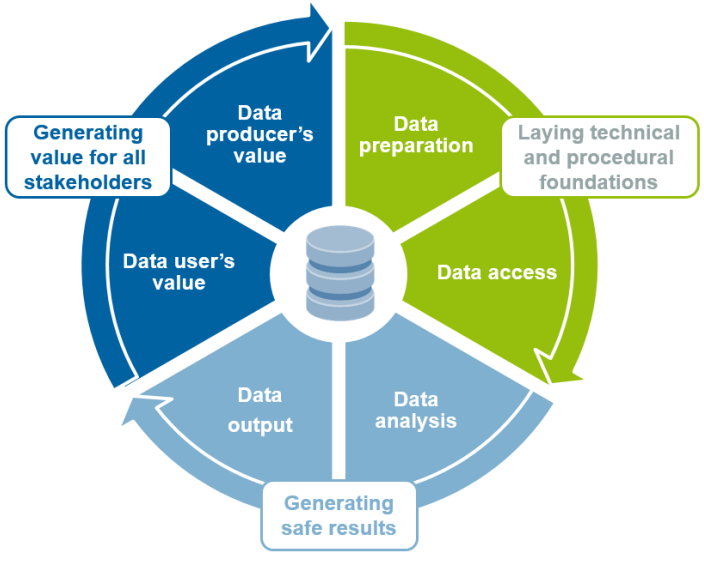
BUBMIC model: Building blocks to design workflows enabling access to micro data - Motivation

Trust must go in both directions:
➔the data producer needs to trust that the data user is not doing harm to the data, and
➔the data user needs to trust that the data producer is not doing harm to the analysis.



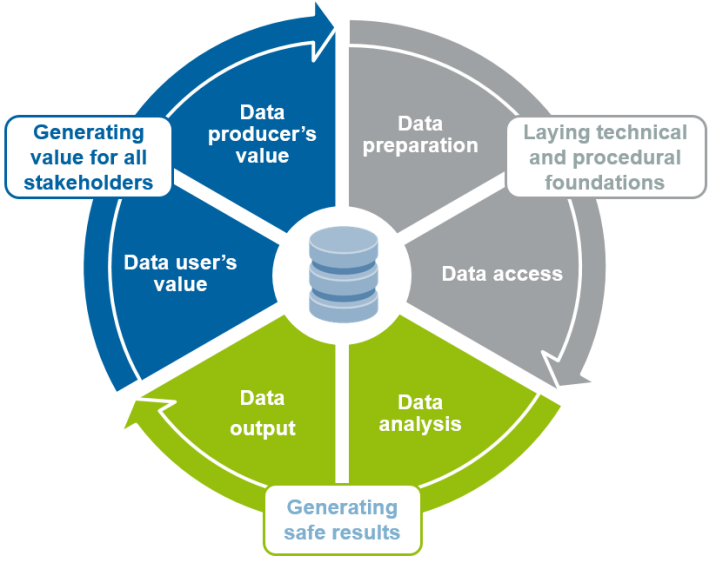
1. Many papers discuss the **costs** and **risks** incurred by **data producers** when providing access to data users/researchers (Five Safes).
2. However, in this kind of access model, **data users incur costs and risks**, too:
   a. real costs (like traveling),
   b. potential risk of censorship of undesirable topics by the data producer,
   c. undefined insufficient data descriptions or data quality,
   d. incorrect output checking, and
   e. misuse of data users' potential analysis ideas by the data producer.

Building block 1: Laying the technical and procedural foundations



1. Data producers bear most of the costs of data access.
    a. costs for making data ready for analysis
    b. providing meaningful descriptions.
2. Data provider also must decide on the appropriate level of detail for the data (5 Safes)
3. Data users must determine whether the content and level of detail of the data are sufficient for their planned analyses.
4. Data providers have to implement technical and organizational measures to safeguard the data, while also allowing researchers to analyze the data in an efficient way.
5. They need to develop procedures to manage applications and provide guidance to users, if needed.
6. Data users are often required to complete a significant amount of paperwork (5 Safes) to get access to the data.

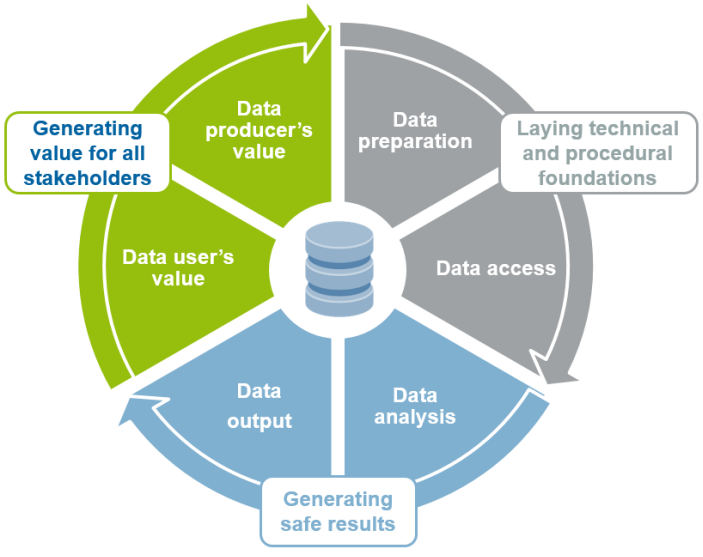
Building block 2: Generating safe results



- Users often must travel to the data producer's safe environment.
- Users must familiarize themselves with the applicable rules and comply with any additional regulation (programming or documentation).
- Data users and data producers must incur costs for output checking (as only safe results may leave the safe environment and be published).
- Data users must incur costs for programming.

- As Lane (2020) observed, the lack of precise information about the research outcomes leads to a situation where public data providers are not able to communicate societal value of their service.

Building block 3: Generating value for the stakeholder



- The concept of 'value' uses objective criteria and can therefore be measured independently of the data-providing institution.
- How much of the value of a publication can be attributed to the data? In a RDC context, there is no established approach to identifying the counterfactual data.
- Measuring the value of research becomes more challenging the further away we move from the research analysis.
- The closest in time to the research analysis is publication (i.e., 'immediate outcomes'), followed by 'intermediate outcomes.' which comprise the dissemination to and use of research in policy and practice.
- Blaschke and Hirsch (2023) take a different and more traditional approach and adopt the 'payback' framework (Buxton & Hanney, 1996; Rollins et al., 2020) to evaluate the benefits of RDCs.
- Knowledge production: Counting the number of projects that resulted in publication, projects.
- Capacity building: identify master or PhD students from their applications.

Conclusion

- It is important to capture the full life cycle, as the costs and benefits are not distributed equally among all stakeholders or across all phases of the life cycle (BUBMIC model). Trust is needed.

- There is a (strong) need to quantify the contribution of data to research outcomes.
- At the same time there (sometimes) is a lack of knowledge to extract the full value of research, for example, through policy debates.
- Establish new frameworks to help capture the full value of research outcomes.

Contact

Thank you!

Stefan Bender (stefan.bender@bundesbank.de)
Christian Hirsch (christian.hirsch@bundesbank.de)
Jannick Blaschke (jannick.blaschke@web.de)

Website: www.bundesbank.de/rdsc