International Network for Exchanging Experience on Statistical Handling of Granular Data (INEXDA)

Stefan Bender, Christian Hirsch (Deutsche Bundesbank)

Conference of European Statistics Stakeholders (CESS) in Paris

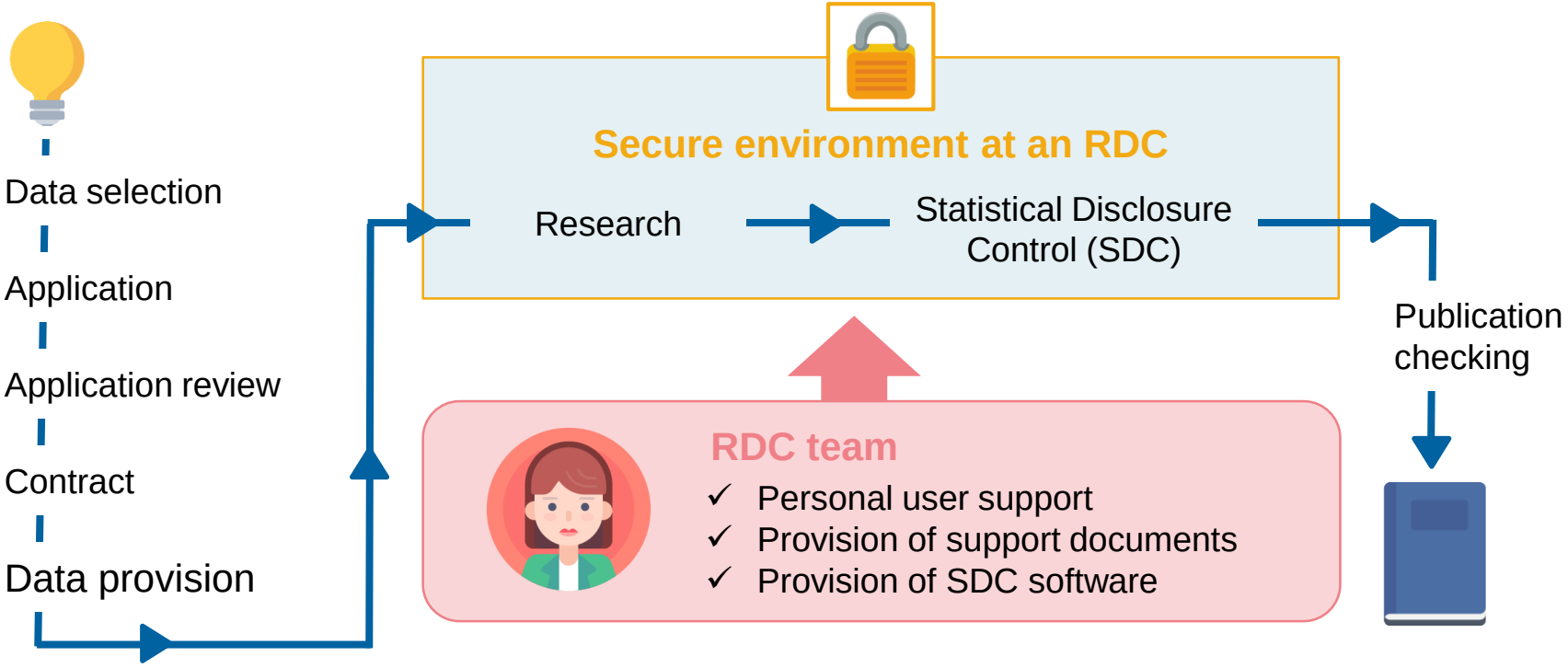
15-16 October 2024

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.
- Exchange of experiences and best practices are strongly needed: INEXDA

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



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

BANCO DE ESPAÑA
Eurosistema

MAIN OUTCOMES OF THE INEXDA WORKING GROUP ON STATISTICAL DISCLOSURE CONTROL (SDC)

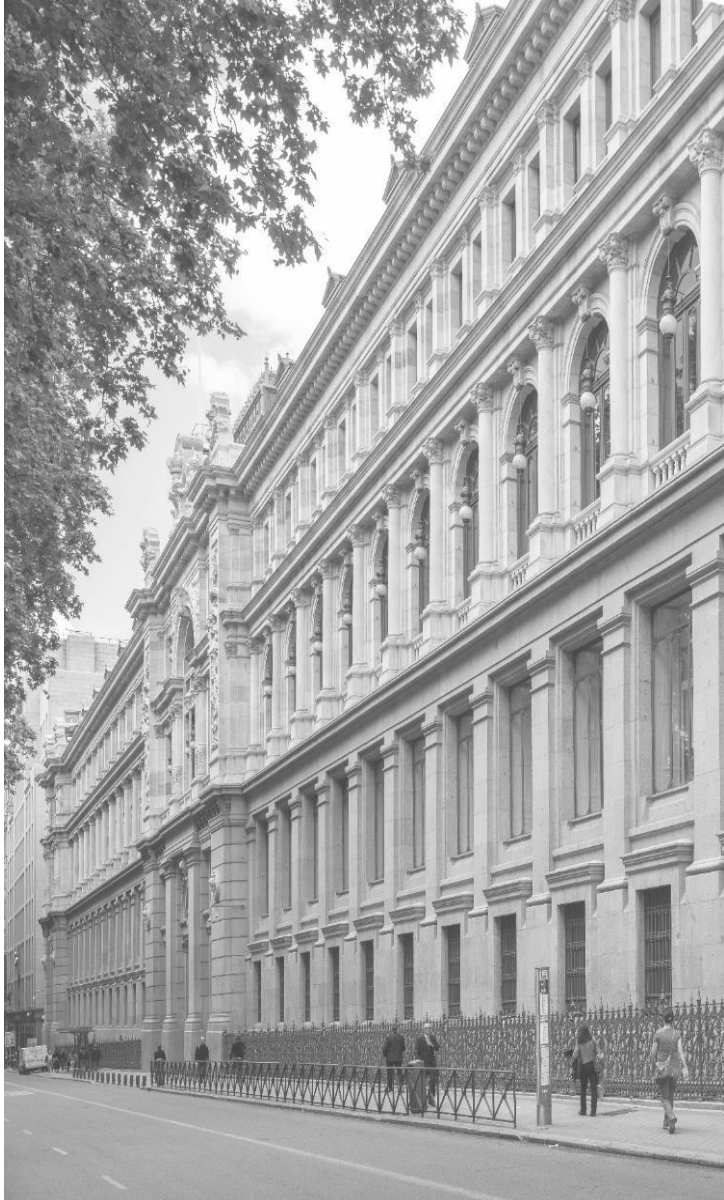
Ana Esteban

Head of Bank of Spain Data Laboratory (BELab) Unit

12TH BIENNIAL IFC CONFERENCE "STATISTICS AND BEYOND: NEW DATA FOR DECISION MAKING IN CENTRAL BANKS"

Basel

22 August 2024



INEXDA
The Granular Data Network

International Network of Exchanging Experiences on Statistical Handling of Granular Data

General misión

- Promote data sharing and data access
- Promote G20 Data Gaps Initiative II, in particular recommendation 20, addressing the accessibility of granular data.
- Acknowledging and supporting the work on data sharing of the Irving Fisher Committee on Central Bank Statistics

Organisation

- Current chair: Stefan Bender of the Deutsche Bundesbank.

Statistical disclosure control (SDC) is a major challenge faced by research data centers (RDC) when making sensitive microdata available to external researchers.

The goal of SDC methods is to prevent the disclosure of sensitive information of individual units (respondents). This can be achieved with: (1) microdata anonymization, and (2) output control.

Microdata anonymization aims to protect the original microdata before making it available to researchers.

Output control procedures ensure that individual agents cannot be re-identified in the results published outside the research data center

INEXDA WG ON STATISTICAL DISCLOSURE CONTROL

Participants

BANCO CENTRAL DE CHILE · DEUTSCHE BUNDESBANK EUROSYSTEM · BANQUE DE FRANCE EUROSYSTÈME · BANCO DE PORTUGAL EUROSISTEMA · TÜRKİYE CUMHURİYET MERKEZ BANKASI · BANCO DE ESPAÑA Eurosistema · BANCA D'ITALIA · BANCO DE MÉXICO · eurostat

Goals

- Identify the current SDC needs, procedures, and tools used among INEXDA members
- Foster harmonization in the area of SDC within INEXDA
- Identify open challenges in the area of SDC and define a plan to address them.

Tasks

- Conduct a survey among INEXDA members to identify current SDC needs,
- Conduct a technical session to review specific use cases and discuss implementation details.
- Define a plan to address open challenges.

.

What we have learned from the working group

Members of the working group presented contributions to RDCs in various areas:
- Algorithms used to anonymize very different microdata
- Tools that can support researchers and staff at statistical disclosure control
- Environments that enable the efficient provision and use of microdata

- You don't need to start from zero, as it's very likely that someone in another RDC has found a solution to a similar problem. That is why exchange is so important.

- But solutions are tailored to specific legal, technical, organisational, and administrative requirements. This is why harmonisation of approaches remains an important goal for the future.

Results of the survey and the final report may be accessed here:
https://www.inexda.org/activities/

What we have learned from the working group

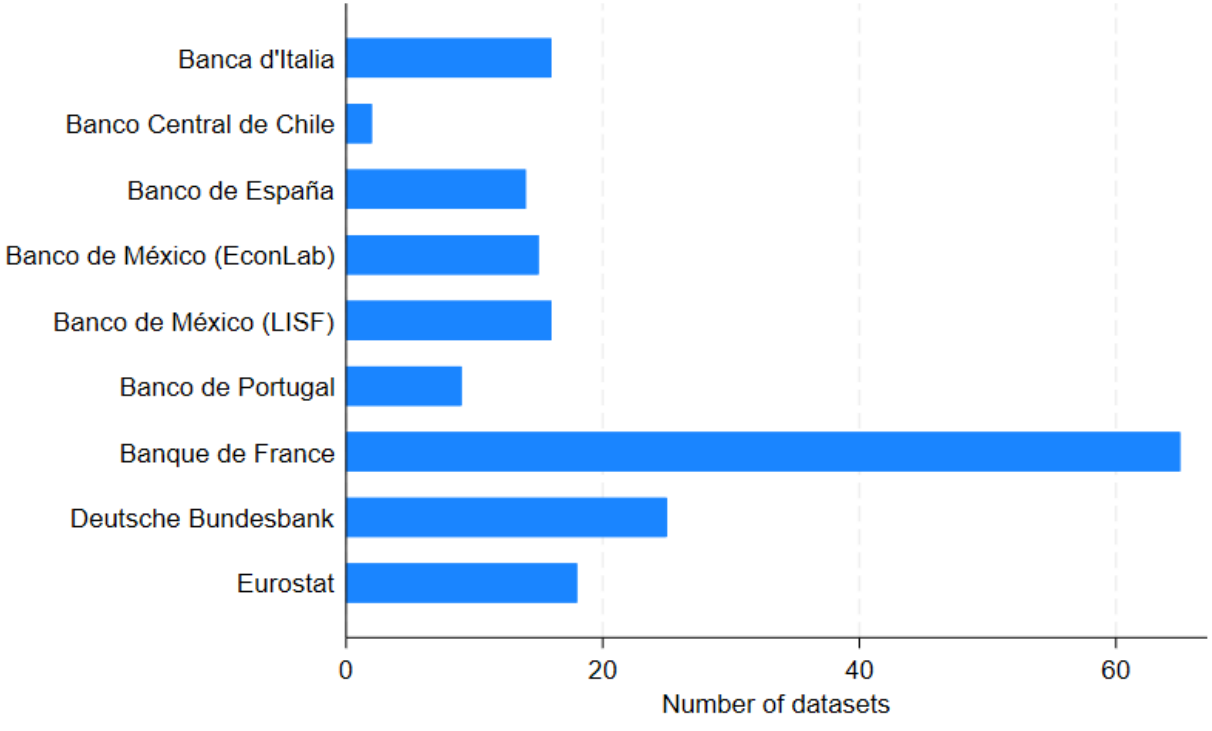
Members of the working group presented contributions to RDCs in various areas:

- Algorithms used to anonymize very different microdata
- Tools that can support researchers and staff at statistical disclosure control
- Environments that enable the efficient provision and use of microdata

- You don't need to start from zero, as it's very likely that someone in another RDC has found a solution to a similar problem. That is why exchange is so important.
- But solutions are tailored to specific legal, technical, organisational, and administrative requirements. This is why harmonisation of approaches remains an important goal for the future.

Results of the survey and the final report may be accessed here:
https://www.inexda.org/activities/

The INEXDA data catalogue



Idea behind the **data catalogue**

- The data catalogue is maintained by the Banque de France.
- Help data users discover and use datasets appropriate for research and analysis by using harmonised metadata.
- Provide framework to facilitate a possible harmonisation of datasets in the future

https://www.inexda.org/data-catalogue/

Research Data and Service Centre

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

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

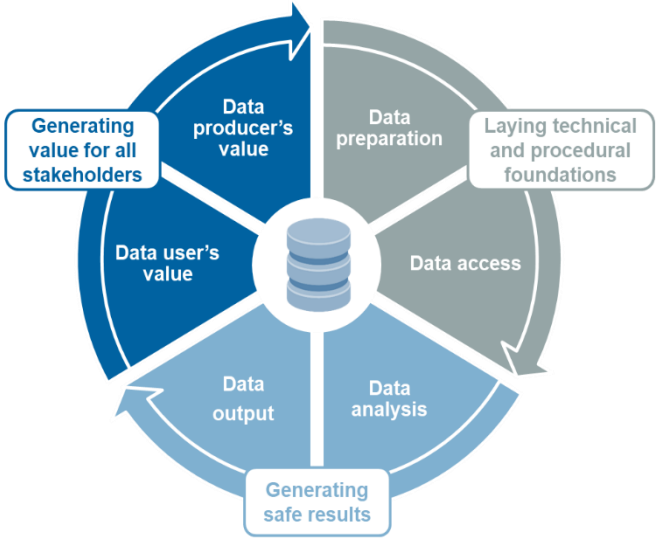
Bender, S., Blaschke, J., & Hirsch, C. (2024). A Practical Use Case: Lesson Learned From Social Science Research Data Centers. Harvard Data Science Review, (Special Issue 4). https://doi.org/10.1162/99608f92.8a2f4507

The views expressed here do not necessarily reflect the opinion of the Deutsche Bundesbank or the Eurosystem.

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

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

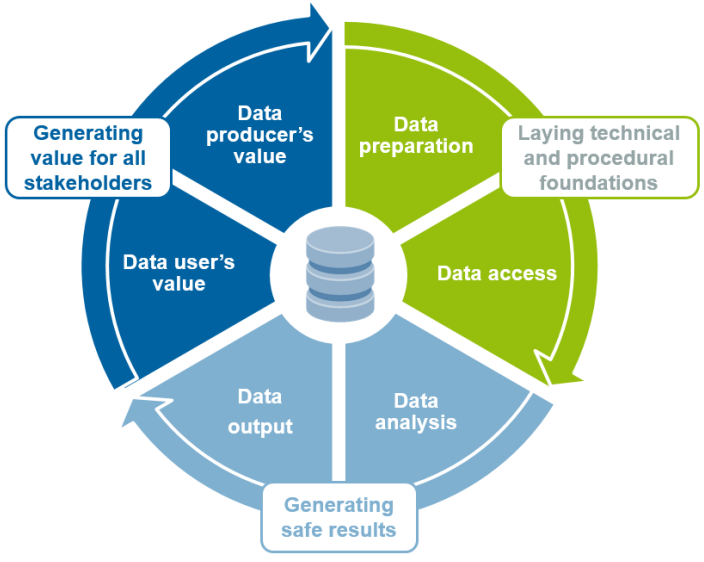
Trust must go in both directions:
➔the data producer needs to trust that the data user is not doing harm to the data, and
➔the data user needs to trust that the data producer is not doing harm to the analysis.



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.

Research Data and Service Centre

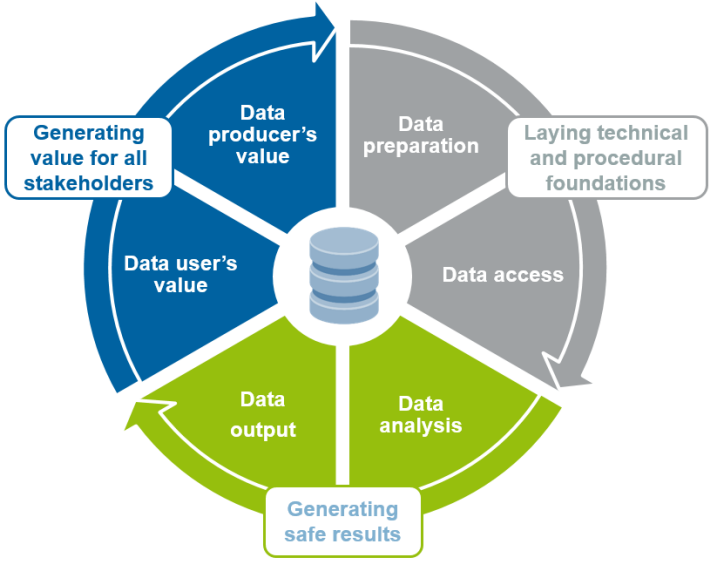
Building block 1: Laying the technical and procedural foundations



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.

Research Data and Service Centre

Building block 2: Generating safe results

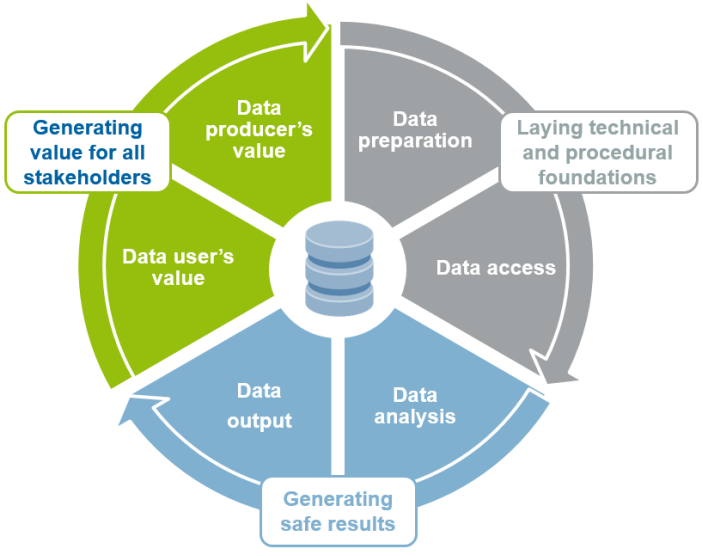


- Users often must travel to the data producer's safe environment.
- Users must familiarize themselves with the applicable rules and comply with any additional regulation (programming or documentation).
- Data users and data producers must incur costs for output checking (as only safe results may leave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

Research Data and Service Centre

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.

Why Democratizing Data Is Important (Julia Lane/Nancy Potok)

- Our world is awash in data. (…) much data is also produced by researchers or government agencies and could potentially be transformed into valuable assets for the public good.
- That potential is now beginning to be realized thanks to important technological advances and policy interventions.
- Our goal (…) is to draw attention to concrete successes, document some key lessons learned from a high-profile pilot, and sketch a future agenda to build on those successes.
- We hope that this special issue is not only a look at what has been done over the last 8 years but is also a way to peer into the future and galvanize the democratizing data community to expand the work. Our future depends on it

Special Issues of Harvard Data Science Research (4/24): Democratizing Data
https://hdsr.mitpress.mit.edu/specialissue4

Contact

Thank you!

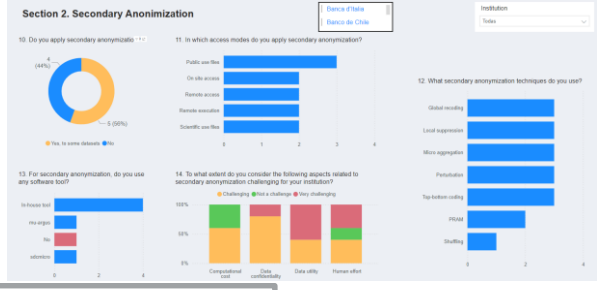
Stefan Bender (stefan.bender@bundesbank.de)
Christian Hirsch (christian.hirsch@bundesbank.de)
Jannick Blaschke (jannick.blaschke@web.de)

Website: www.bundesbank.de/rdsc

BACK UP

Primary Anonymization

Secondary Anonymization

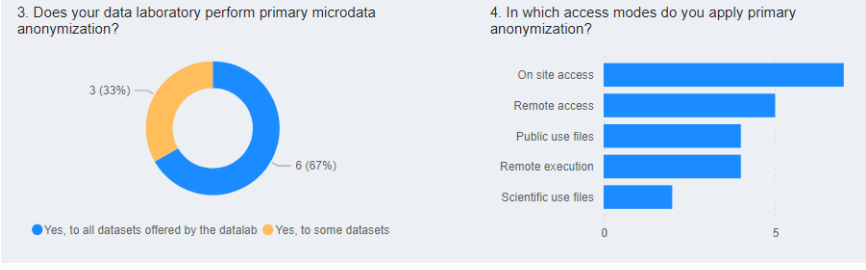


SURVEY CONTENT

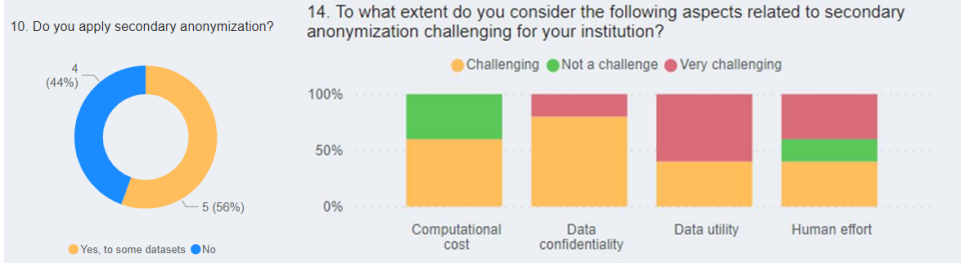
Output Control

General Questions

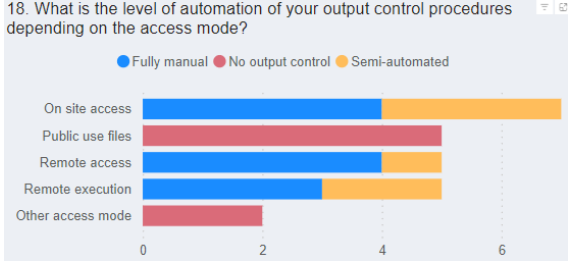
Primary Anonymization



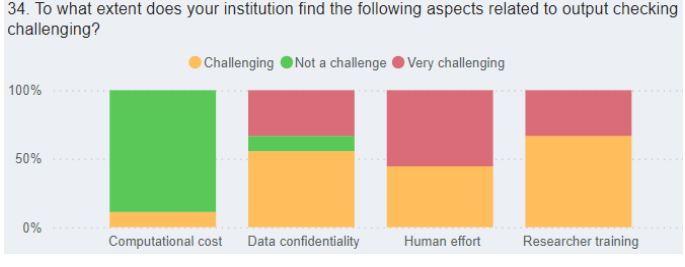
Secondary Anonymization



SURVEY: MAIN RESULTS



Output Control



Eleven SDC use cases were presented by the members of the WG in Madrid, classified into the following topics:

ANONYMIZATION

- **Disclosure avoidance in the Spanish Survey of Household Finances** (*Cristina Barceló, Banco de España*)
- **Anonymization Algorithm for trade transactions** (*Claudia Velázquez, Banco de México)*
- **Joint analysis of categorical and continuous key variables for secondary anonymization of sensitive microdata** (*Eugenia Koblents, Banco de España*)
- **Bank of Italy remote execution system** *(Daniele Piras, Banca d'Italia)*
- **BPLIM's approach to protecting sensitive data** *(Joana Pimentel, Banco de Portugal)*

OUTPUT CONTROL

- **Tools and resources for output checking at the RDSC of Bundesbank** *(Hariolf Merkle and Christian Hirsch, Bundesbank)*
- **Automated checking of research output (ACRO)** (*Marco Stocchi, Eurostat)*
- **Measures to ease code reproducibility for output control (***Ricardo Arcos and Emma Pérez, Banco de España)*
- **Output control practices in the Spanish Survey of Household Finances** *(Cristina Barceló, Banco de España)*

DATA SHARING AND PRIVACY ENHANCING TECHNOLOGIES (PET)

- **Utility and confidentiality assessment of synthetic financial data- Pilot in collaboration with the European Commission** *(Eugenia Koblents, Banco de España)*
- **Towards a shared European Statistical System infrastructure for collaborative confidential computing (***Fabio Ricciato, Eurostat)*

Questions that a new RDC may inquire ask in relation to SDC and example **answers**, based on the experience of the participants in this WG.

SAFE DATA

- **What techniques can RDCs use to maintain data confidentiality?**
- **Can different anonymization techniques be applied depending on the mode of access to the data?**
- **Is it necessary to guarantee that data cannot be indirectly re-identified?**
- **…**

SAFE OUTPUT

- **What requirements does a good output have to meet to facilitate the work of the output checker?**
- **Can output checking be fully automated?**
- **What rules are required to ensure that the released output is safe?**
- **…**

PROPOSAL OF NEXT STEPS

The goals of the WG have been successfully achieved. However, there is still room to reach a higher harmonization in the area of SDC within INEXDA

1 ACTION

- **Dissemination strategy of the final report**

2 ACTION

- **Addressing a plan to reach a higher harmonization in the area of SDC within INEXDA**

3 ACTION

- **Annual virtual follow-up meeting to exchange experiences.**

4 ACTION

- **Every three years, an in-person workshop with the aim of sharing the most outstanding innovations in this field among all participants.**