INEXDA
The Granular Data Network

22.04.2024

Main outcomes of the INEXDA working group on Statistical Disclosure Control (SDC)

SUMMARY: This document provides an overview of the primary results of the INEXDA working group on Statistical Disclosure Control (SDC), which was initiated in June 2023 and concluded in April 2024. During this period, the group focused on the two primary categories of SDC methods, namely anonymization and output control. The main objective of the working group was to describe the current SDC requirements and procedures employed by INEXDA members and to encourage knowledge sharing and standardization within INEXDA in the field of SDC.

Contributors to this report

Members of the INEXDA Working Group on Statistical Disclosure Control (SDC)

Institution	Name
Banca d'Italia	Daniele Piras
	Elena San Martini
Banco Central de Chile	Patricia Medrano
Banco de España	Ricardo Arcos
	Cristina Barceló
	Laura Crespo
	Ana Esteban (chairperson)
	Sandra García Uribe
	Eugenia Koblents
	Emma Pérez
Banco de México	Claudia Katheri Velázquez Villegas
Banco de Portugal	Joana Pimentel
	Gustavo Iglesias
	Rita Sousa
Banque de France	Vincent Georgelin
Central Bank of Turkey	Rifat Aras
Deutsche Bundesbank	Christian Hirsch
	Hariolf Merkle
Eurostat	Aleksandra Bujnowska
	Frank Espelage
	Fabio Ricciato
	Marco Stocchi

Contents

1 Executive summary

This document summarizes the main outcomes of the INEXDA working group (WG) on Statistical Disclosure Control (SDC), which was launched in June 2023 and completed in March 2024. The WG focused on the two main families of SDC methods, namely anonymization and output control. The main goal of this WG was to review current SDC needs and procedures used among INEXDA members and to promote knowledge sharing and harmonization within INEXDA in the area of SDC.

In the context of this WG, an extensive survey was conducted among INEXDA members to identify current SDC needs, procedures, and software tools. The results of this survey were analyzed and presented via an interactive dashboard. This WG was run in a hybrid format, involving three virtual meetings as well as one in-person meeting, which took place in Madrid in November 2023. During this in-person meeting, the most relevant SDC use cases developed within INEXDA were presented and discussed. The goal of this document is to present a comprehensive summary of all the valuable information shared by INEXDA members on their SDC procedures during this WG, including an overview of the main features of the Research Data Centers (RDC) participating in the project, the survey results, the use case descriptions, and the main lessons learned, which can be useful for new teams addressing similar challenges.

The survey has revealed that all participating RDCs apply primary anonymization techniques, and only a few need to implement secondary anonymization techniques for certain datasets. In this context, data utility and confidentiality are the most relevant aspects for RDCs that apply secondary anonymization. Additionally, all RDCs perform output control. While some of them acknowledge progress in semi-automation, overall, this task is predominantly carried out manually. This circumstance leads participants to consider it challenging or very challenging in terms of human effort and researcher training.

2 Introduction

This introduction includes the mandate of this WG and summarizes the content of this document. Statistical disclosure control (SDC) is a major challenge faced by data laboratories of national central banks and by national statistical institutes when making sensitive microdata available to external researchers. The goal of SDC methods is to prevent the disclosure of sensitive information of individual statistical units (respondents), which can be triggered by their direct or indirect re-identification from the original micro dataset or from any aggregated data derived from it. This motivation led INEXDA to propose the creation of this WG on SDC to be run by Banco de España.

The SDC process consists of two main tasks: (1) microdata anonymization, and (2) output control. Microdata anonymization aims to protect the original microdata before making it available to researchers. Access modes that may require anonymization include on-site and remote access within a secure data lab environment, as well as public distribution of sensitive data. Output control procedures ensure that individual respondents cannot be re-identified in the results to be published outside the data laboratory as a consequence of the research performed.

This WG provided an overview of both SDC tasks in the context of INEXDA. In particular, the main goals of this WG were the following:

1. Identify the current SDC needs, procedures, and tools used among INEXDA members.
2. Foster harmonization in the area of SDC within INEXDA, leveraging the experience of senior members. This may result in the production of best-practice recommendations and the sharing of software tools.
3. Identify open challenges in the area of SDC and define a plan to address them.

These goals have been successfully achieved by the WG, with close collaboration among all INEXDA members and especially between the WG members. The main tasks addressed during the WG duration were the following:

1. Conduct a survey among INEXDA members to identify current SDC needs, procedures, and software tools (both open source and proprietary) used for different access modes and levels of confidentiality. Collect, summarize, and present the survey results, identifying lessons learned, open challenges, and future research and development areas.
2. Provide an overview of past and present use cases.
3. Conduct a technical session to review specific use cases and discuss implementation details.
4. Define a plan to address open challenges.

This document presents the main outcomes of the INEXDA SDC WG and is organized as follows: Section 2 provides an introductory summary of this document and presents the WG goals and main tasks. Section 3 summarizes the main outcomes of the WG, including an overview of the main characteristics of the RDSc participating in the project, a summary of the main survey results, and a description of the main SDC use cases reported by INEXDA members. Section 4 summarizes the main lessons learned during the WG, and finally, Section 5 presents the conclusions and next steps.

3 Main outcomes of the working group

This section provides an overview of the main outcomes of the SDC WG, including a summary of the main characteristics of the participating INEXDA RDC in Section 3.1, a summary of the main survey results in Section 3.2, and a description of the presented SDC use cases in Section 3.3.

3.1 General overview of the Research Data Centers

The following institutions participated in the WG: Banca d'Italia, Banco Central de Chile, Banco de España, Banco de México, Banco de Portugal, Banque de France, Central Bank of Turkey, the Deutsche Bundesbank and Eurostat.

This section presents basic information for each RDC to provide a comprehensive and comparable overview of their situation. Knowing the characteristics of each research center in advance helps to have a better understanding on the generalities and particularities related to SDC in each RDC discussed in the following sections.

Banca d'Italia

Overview: The RDC of Banca d'Italia was established in 2019, based on the 2016 Strategic Plan for the following three years, in order to expand and improve the information resources available to the public. The RDC, as part of the Economics and Statistics Department, is the single entity centralizing the collection and the dissemination of the microdata within the Banca d'Italia, for research purpose.

Access Modes	Public use files, remote execution
Type of researchers	Internal, External
Datasets	>60
Primary anonymization	Some datasets
Secondary anonymization	Yes
Perform Output Control	Yes, to some datasets
Output volume limit	Volume in remote execution
Output code reproduction	All
Active Projects	36

Banco Central de Chile

Overview	
Access Modes	
Type of researchers	Internal, external (co-authoring)
Datasets	
Primary anonymization	Some datasets
Secondary anonymization	No
Perform Output Control	Yes, to some datasets
Output volume limit	No
Output code reproduction	Data analysis code
Active Projects	

Banco de España

Overview: The Banco de España Data Laboratory (BELab) was established in 2019 as a pilot project, and three years later, it evolved into a fully consolidated service within a dedicated unit. Its primary objective is to enhance access for the research community to high-quality confidential microdata from Banco de España. This access is provided in a controlled environment, ensuring data confidentiality. The BELab operates under the umbrella of the Statistics Department.

Access Modes	On site, Remote, Remote execution, Public use files
Type of researchers	External
Datasets	14
Primary anonymization	Some datasets
Secondary anonymization	Yes

Banco de España

Perform Output Control	Yes, to all datasets offered
Output volume limit	No
Output code reproduction	Data analysis code
Active Projects	40

Banco de México

Overview: The EconLab is the data research laboratory of Banco de México at the General Directorate of Economic Research. The main objectives are further our understanding of the Mexican economy by supporting evidence-based research, foster ties with the academic community to advance the quantity and quality of economic research carried out at Banco de México, provide, through co-authorships, a framework for joint research projects that require the use of confidential microdata. The EconLab only provide access for academic research, it is not possible to provide access for policy-oriented research (e.g. policy briefs, Boxes in institutional reports, etc.) or other purposes.

Access Modes	On site, Remote
Type of researchers	Internal, external
Datasets	>15
Primary anonymization	All datasets
Secondary anonymization	No
Perform Output Control	Yes, to all datasets offered
Output volume limit	No
Output code reproduction	Cooming soon
Active Projects	>50

Banco de Portugal

Overview: The Banco de Portugal Microdata Research Laboratory (BPLIM) started its activity in 2016 as an autonomous unit within the Economics and Research Department. Its primary objective is to facilitate the development of research projects and studies focusing on the Portuguese economy. Through BPLIM, both internal and external researchers gain, in a controlled environment, access to well-documented and anonymized microdata sets that can be tailored to suit their specific requirements. Moreover, researchers have the opportunity to utilize the computational resources offered by the laboratory. By offering remote access to its data, BPLIM aims to attract the attention of both national and international researchers.

Access Modes	On site, Remote Access, Remote execution
Type of researchers	Internal, external
Datasets	>10
Primary anonymization	All datasets
Secondary anonymization	Yes
Perform Output Control	Yes, to all datasets offered
Output volume limit	No

Banco de Portugal

Output code reproduction	All
Active Projects	136

Banque de France

Overview: In 2016, the Banque de France opened its data to external researchers in a room with three stations giving access to some 400 million lines of anonymized statistics, initially in Paris, then in New York in 2018. In 2020, the second version of the "Open Data Room" offers a remote secure access to data for around twenty projects, but for the entire team. Since 2022, access to anonymized data from Banque de France for external research has been outsourced to the CASD (Centre d'Accès Sécurisé aux Données - *Secure Data Access Center*) which allows cross-referencing with other French institution data. Banque de France, however, still ensures the data anonymization and their transfer to the CASD.

Access Modes	Remote, Remote execution
Type of researchers	Internal, external
Datasets	Internal (>250), external (> 60)
Primary anonymization	All datasets
Secondary anonymization	Yes for external
Perform Output Control	Yes, to all datasets offered to external
Output volume limit	No
Output code reproduction	No
Active Projects	40

Central Bank of Republic of Turkey

Overview	
Access Modes	On site access
Type of researchers	Internal, external
Datasets	
Primary anonymization	All datasets
Secondary anonymization	No
Perform Output Control	Yes, to all datasets offered
Output volume limit	No
Output code reproduction	No
Active Projects	

Deutsche Bundesbank

Overview: To meet the increased internal and external demand for microdata and to data confidentiality requirements, in 2013 the Bundesbank established the Integrated Microdata-based Information and Analysis System (IMIDIAS) and established the Research Data and Service Centre (RDSC) (for more detail, refer to [7], [5]). Today, the RDSC provides standardised access to selected microdata collected by the Bundesbank in accordance with its statutory mandate to

Deutsche Bundesbank

be used in independent scientific research projects. Since its foundation, projects started at the RDSC resulted in over 150 scientific papers and over 130 (graduate) degrees [6].	
Access Modes	On site, Remote execution, Scientific use files
Type of researchers	Internal, external
Datasets	>25
Primary anonymization	All datasets
Secondary anonymization	Yes
Perform Output Control	Yes, to some datasets
Output volume limit	Volume of output
Output code reproduction	Only if deemed it necessary
Active Projects	480

Eurostat

Overview: Eurostat provides access to European microdata for scientific purposes for more than 20 years and constantly improves its microdata access services. Eurostat offers a single-entry point to harmonised microdata sets of EU countries, EFTA and candidate countries. The access procedures are the same for all microdata sets available for research at Eurostat.	
Access Modes	On site, Remote (both to Secure use files) Public use files, Scientific use files
Type of researchers	Internal, external
Datasets	14
Primary anonymization	All datasets
Secondary anonymization	Yes (Public use files, Scientific use files)
Perform Output Control	Yes, to some datasets (secure use files)
Output volume limit	Volume of output, code and logs
Output code reproduction	No
Active Projects	>1000

3.2 Survey on SDC. Main results

As one of the initial tasks of the WG, a survey was conducted among participating INEXDA members to identify current SDC needs, procedures, and software tools used for different access modes and levels of data confidentiality. The survey was created using Microsoft Forms and contained a total of 40 questions, grouped into four main sections, namely primary anonymization, secondary anonymization, output control, and other general questions.

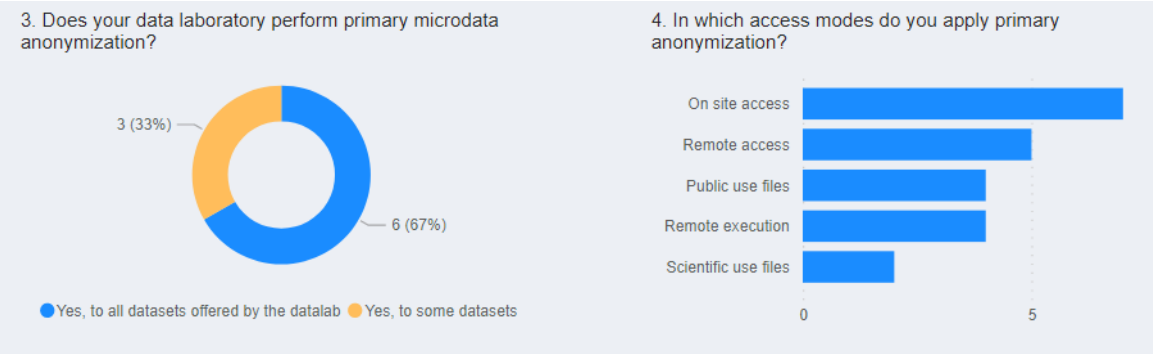
All WG participants responded to the survey. Answers have been compiled into an interactive dashboard[1] that also allows to select and compare different institutions. It is available in this link. The main findings are summarized below.

3.2.1 Primary anonymization

All institutions apply primary anonymization (PA), either to all datasets or only to some, across all access modes but mostly in on-site and remote access (Figure 1). Overall, PA is identical (same procedures and parameters) across research projects, and it is performed mostly with in-house tools and code or dedicated packages in different programming languages. The merging of datasets is allowed, either performed by the data lab or by the researcher.

In general, external researchers are the main users of the data, along with internal researchers in most institutions.

Figure 1. RDC performing primary anonymization and access modes



3.2.2 Secondary anonymization

Only 5 out of 9 participating institutions apply secondary anonymization (SA), namely, Banco de España, Bundesbank, Banco de Portugal, Eurostat, and Banca d'Italia. Only their answers are considered in this section.

All responding institutions apply SA to only a subset of the datasets. The application of SA depends on the confidentiality level of each dataset and the type of released files (Figure 2). The techniques applied are varied, with each institution using at least 2 different ones, Bundesbank and Eurostat using up to 6. All of them use internal tools alone or in combination with sdcmicro or mu-argus.

Regarding challenges, data utility and confidentiality are the most relevant aspects, as all respondents consider them to be very challenging or challenging. Human effort is the next most challenging aspect. Figure 3 shows these results.

[1] Open-ended questions are not included in the dashboard.

Figure 2. RDC performing secondary anonymization and access modes

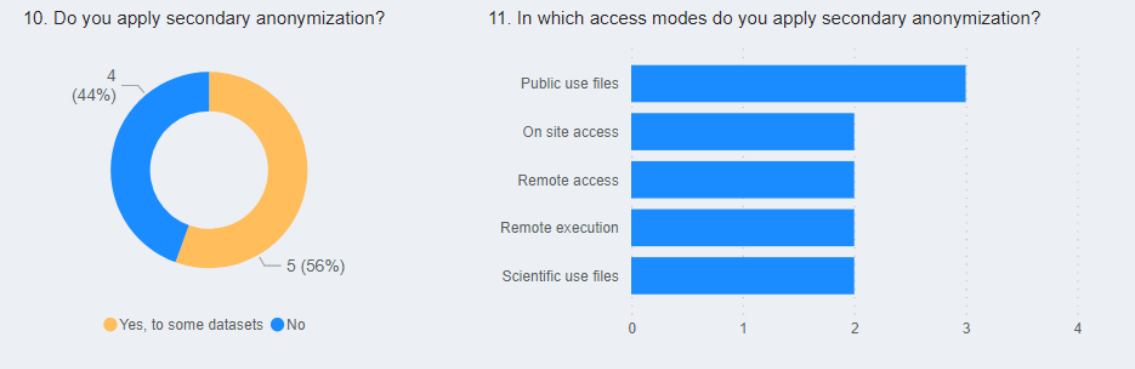
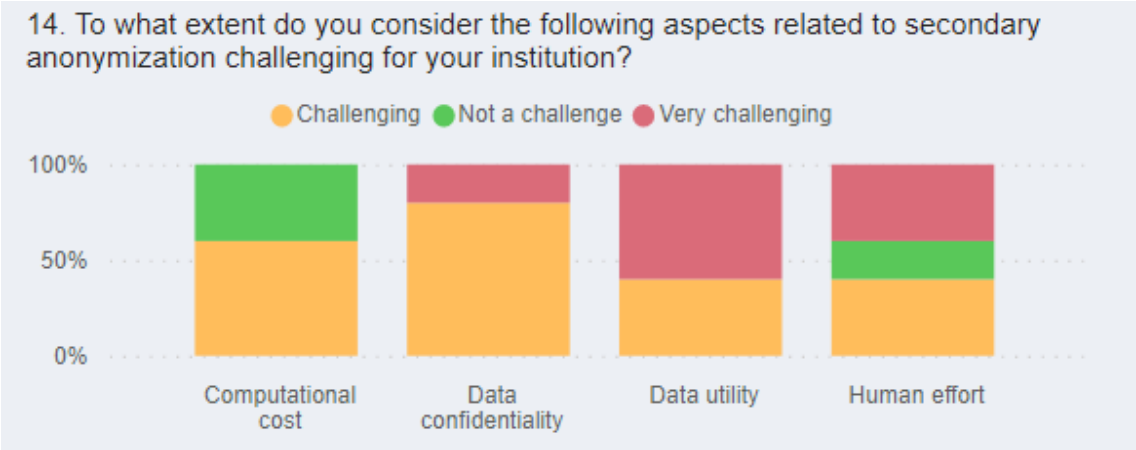


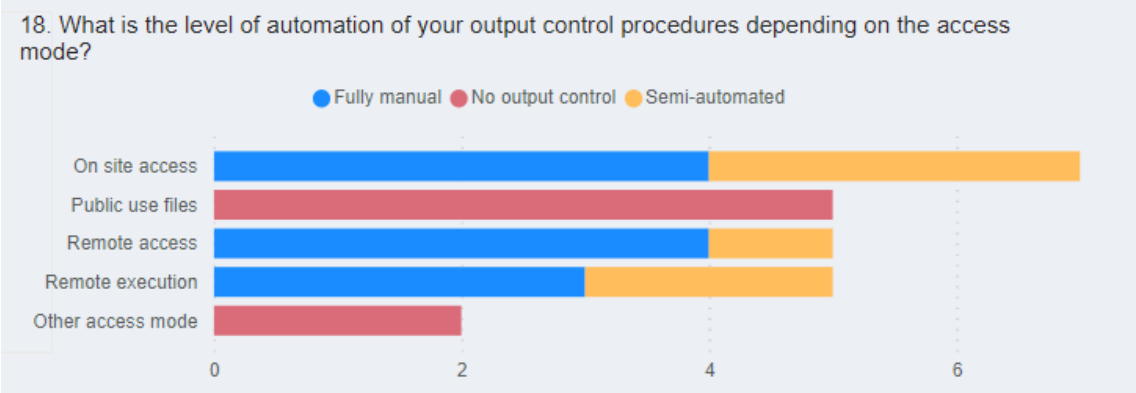
Figure 3. Secondary anonymization main challenges



3.2.3 Output control

All institutions apply output control, either to all datasets or only to some, using their own criteria or in combination with the DwB guidelines promoted by Eurostat[2]. The output control process for remote execution/access and onsite access is being semi-automated in some institutions, but in general it is still mostly manual, as presented in Figure 4. Automation of output checking process.

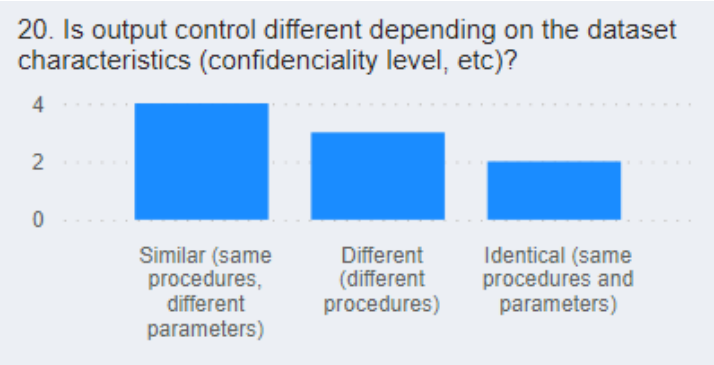
Figure 4. Automation of output checking process



[2] These guidelines were developed in the context of the Data without Boundaries European research project.

The output checking process is identical or similar across datasets in most institutions (Figure 5), in the latter case making use of different parameters depending on the characteristics of the dataset. Three institutions apply different procedures depending on the data.

Figure 5. Types of output control across datasets



Most institutions do not limit the volume of output/code/logs to be checked per project. A few institutions do apply restrictions to the volume of output and/or logs. Just three institutions enforce a fixed structure to organize research projects. Half of the institutions do not reproduce the code (or do it only when deemed necessary), and others reproduce either the whole code or just some parts. In terms of programming languages, researchers most commonly use **STATA**, followed by **R** and **Python**. The least commonly used languages are **MATLAB** and **SAS.**

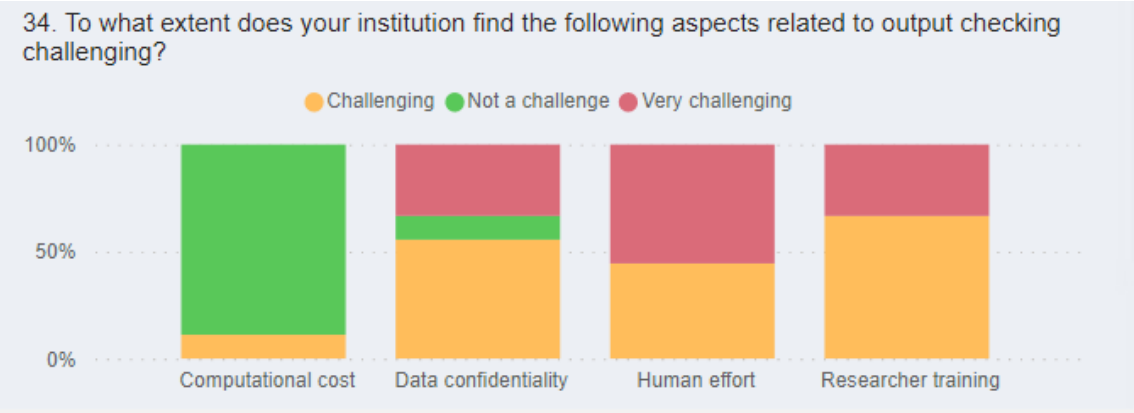
In terms of output formats, CSV and Excel formats are widely used to release results, followed by images, PDF, and LaTeX. All institutions allow the release of means and percentiles as descriptive statistics. Some also allow graphs and modes, while only three institutions allow extracting maximum and minimum values. 50% of the institutions require a minimum of 3 observations for extraction, while the rest have higher thresholds.

There is no consensus on the dominance rule. Three institutions apply the 85% rule for the largest contributor, while others apply it to the two largest contributors. Some institutions have no fixed threshold, or one depending on the dataset. Around 50% of institutions check confidentiality across multiple tables.

Specific guidelines and ad-hoc meetings are widely used to assist researchers on adhering to control rules. Additionally, example code is also used. However, six institutions encounter problems regarding compliance with the rules sometimes or often. Most institutions report a response time for output control requests of two weeks or less, with some even having one week or less.

Regarding the main challenges, human effort and researchers' training are considered challenging or very challenging by all institutions, followed by data confidentiality. Computational cost is not a concern in this case: most consider it not to be a challenge (as displayed in Figure 6. Output control main challengesprobably due to the lack of automation.

Figure 6. Output control main challenges



3.2.4 Other general questions

Most institutions offer a number of different access modes (at least two, except the Central Bank of the Republic of Turkey, which only provides onsite access). Seven institutions provide onsite access, mostly in combination with other access modes like remote access and remote execution. There are two institutions that do not provide onsite access. Public use files (PUFs) and Scientific use files (SUFs) are used by some institutions too.

There are several professional profiles involved in anonymization and output control duties: data scientists and economists are present in most institutions, followed by statisticians and methodologists. To a lesser extent, we also find data engineers, assistants and software developers.

Most institutions have no experience with alternative privacy enhancing techniques (PETs), with the exception of EUROSTAT, Bundesbank and Banco de España.

3.3 Use cases

This section provides an overview of the eleven SDC use cases reported by INEXDA members participating in the WG, classified into the following topics: anonymization, output control, data sharing, and PETs.

3.3.1 Anonymization

3.3.1.1 Disclosure avoidance in the Spanish Survey of Household Finances

Cristina Barceló, Banco de España

We explain the practices in the Spanish Survey of Household Finances (EFF) for avoiding the identification of the households that participate in this wealth survey conducted by the Banco de España. This survey collects detailed information on household assets, debts, income and consumption, as well as labour income and demographic and labour characteristics of all household members. The survey data do not contain personal identity information; the household identification code is completely anonymized.

The goal of this survey is to allow different studies of household portfolio composition and inequality, the EFF has an oversampling of the rich as also implemented in other wealth surveys such as the Survey of Consumer Finances (SCF) in the US. For this purpose, it is key to preserve the true distribution of variables and relationships among them. Thus, methods such as rounding off monetary values, making a perturbation of data values, interchange values or gather observations into grouped values are not appropriate.
To avoid the disclosure of sensible information for identifying households, we remove confidential information such as the month of birth, geographical information and the place of birth, among others. We also provide information in a more aggregated way (1-digit occupation instead of 2 digits), we censor household member age at 85 and also the number of years living in the house at 85 years at most. Finally, we keep different anonymized household identification codes across different datasets, and we store separately public multiply imputed survey data from the files that contain confidential information (National Statistics Institute (INE), Tax Office, interviewers' metadata and fieldwork data).

3.3.1.2 Anonymization Algorithm for trade transactions

Claudia Velázquez, Banco de México

During this presentation, the anonymization algorithm used to obfuscate the trade database was introduced. This algorithm allows for the publication of information at a more granular level than official statistical data. It achieves factual anonymization of the data, preventing the identification of amounts from any participating company while presenting information at a product, month, and country level. The rules utilized by the algorithm were explained in detail to illustrate how it covers all possible combinations of data situations.

3.3.1.3 Joint analysis of categorical and continuous key variables for secondary anonymization of sensitive microdata

Eugenia Koblents, Banco de España

An SDC approach for secondary anonymisation of sensitive time series microdata has been developed at the Banco de España's BELab data laboratory in order to protect confidentiality of the recently published CIR dataset. This dataset contains yearly microdata on loans to legal persons, including multiple variables describing loans and debtors. The main challenge faced in this work is the fact that the set of key variables, i.e. those that may allow debtor re-identification, includes both categorical and numerical loan and debtor variables. Additionally, debtors may have multiple loans and loans may have multiple debtors, which makes the direct use of existing SDC software tools for microdata protection (mu-argus, sdcMicro) unfeasible. For these reasons, a novel SDC procedure has been designed and implemented in order to protect the debtors appearing in the CIR dataset against re-identification, while jointly analysing categorical and numerical variables and addressing time series data protection. The proposed procedure has been designed in close collaboration with the CIR dataset provider and consists of the following steps:

1. Identification of continuous and numerical debtor and loan key variables.
2. Global recoding of selected key variables significantly reducing the disclosure risk.
3. Creation of full debtors' profiles that incorporate information on all their loans throughout the full time series.
4. Local suppressions on debtor profiles with sdcMicro to guarantee k-anonymity.

5. Reinsertion of supressed zeros back into debtor profiles.
6. Local suppressions based on nearest neighbours to protect remaining risky debtors.
7. Transfer of local suppression patterns of debtors to the original loans dataset.

Table 1 shows a summary of the number and percentage of local suppressions performed on each of the identified key variables. The overall number of suppressions was below 1%.

Table 1. Summary of the number of local suppressions per variable in the full time series.

Debtor and loan key variables	Number of suppressions	Percentage of suppressions
Residence	70,720	0.27
Institutional sector	81,436	0.31
Legal form	528,404	1.98
Economic activity	3,006,248	11.29
Enterprise size	447,127	1.68
Currency	44,428	0.17
Guarantee	15,623	0.06
Drawn amount	2,993	0.01
Undrawn amount	2,993	0.01
Investment region	100,244	0.38
TOTAL	**4,300,076**	**0.95**

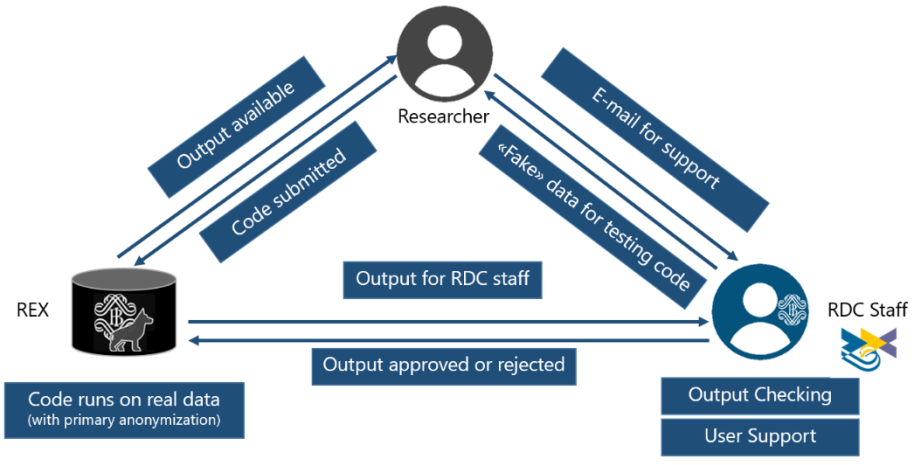
This work has been published in the Eleventh IFC Conference on "Post-pandemic landscape for central bank statistics" [1] and in the Privacy in Statistical Databases Conference in 2022 [2].

3.3.1.4 Bank of Italy remote execution system

Daniele Piras, Banca d'Italia

One of the methods used by Banca d'Italia to disseminate data to external researchers is remote execution. As of January 2023 a new tool was released: REX, a Remote Execution platform that facilitates script submissions by researchers and output control by RDC staff (Figure 7).

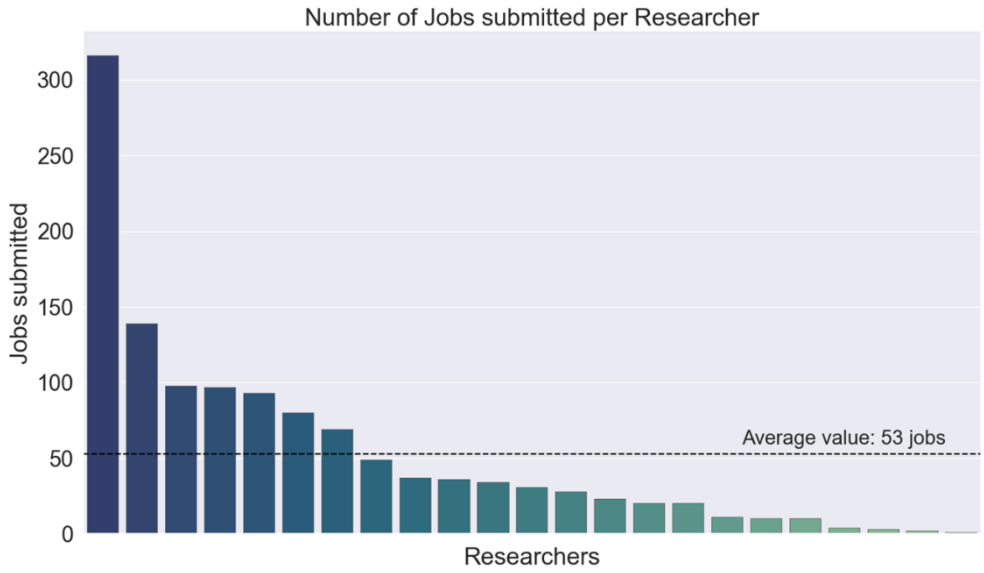
Figure 7. Interactions among the remote execution platform REX, researchers and RDC Staff in Banca d'Italia.



Through REX, 7 datasets are available to external researchers, but these can be enriched with additional data from researchers, allowing them to have their own customized dataset. Starting from January 2023, the number of active researchers increased throughout the year, peaking in September with 13 active researchers.

While this may not seem like a high number, it should be noted that researchers experiment some difficulties due to the fact that they do not have direct access to data, as a consequence, on average, a researcher needs slightly more than 50 jobs to get the desired output (Figure 8). However, the RDC provides to researchers fake datasets to test the scripts.

Figure 8. Number of jobs submitted by researchers.



Researchers who need more jobs may compromise the proper functioning of the system, as output checking is performed manually. However, RDC staff manages to maintain an excellent level of efficiency: the average response time to an output checking request is around 20 hours.

The REX system guarantees confidentiality at three levels:

1. Users: must submit an application to have access to remote execution system, providing: personal info (valid ID required), details on the research project and the signature of a formal agreement including privacy law and deontological code.
2. Data: are not completely anonymized, but identification variables are expunged, stratification variables are collapsed and extreme values may be censored.
3. System: programs submitted are automatically processed, a parser system identifies and rejects jobs containing some blacklisted instructions, but only the ones that can compromise the integrity of the system. However, RDC manually reviews all outputs before sending results back.

The system is constantly evolving and further improvements are planned for 2024, both in terms of technology (limiting submissions to 3 pending jobs per user and automatic execution of code on fake data to avoid syntax errors) and content (enriching the data offer).

3.3.1.5 BPLIM's approach to protecting sensitive data

Joana Pimentel, Banco de Portugal

The Banco de Portugal Microdata Laboratory (BPLIM) primary goal is to promote external research on the Portuguese economy by making available datasets collected and maintained by Banco de Portugal (BdP). Given that some of these datasets contain highly sensitive information, BPLIM had to implement data access solutions that preserve the confidentiality of the data. The solutions adopted are based on the following principles: (1) access is free of charge and only for scientific purposes; (2) all data must be analyzed on the servers of the bank; (3) external researchers are granted remote access to the server; (4) confidential datasets placed on the server have to always be "modified"; (5) researchers can always ask BPLIM staff to run their scripts on the original data.
BPLIM classifies datasets into 3 levels of confidentiality: low, medium, and high. If the level of confidentiality is low, external researchers access anonymized data.
If the level of confidentiality is medium or high then, on top of the anonymization, the data must be "modified". BPLIM uses 3 strategies to create "modified" datasets: perturbation, shuffling, and randomization (dummy data). For each one of these strategies, BPLIM has developed different tools.

For anonymization, we developed the 'anonimizanif' and 'validarnif' tools. To automate the process of data perturbation we developed the 'perturbdata' tool. The 'bpmask' tool allows us to automate the process of shuffling data while trying to preserve some features of the original data. Finally, we developed a tool – 'dummyfi' - to create random/dummy data with some underlying structure, captured by the metadata of the original data.

Therefore, for medium/high confidential data sets, what is placed in the account of the researcher are modified versions of the data. The researcher implements all scripts based on the data he has available and produces the outputs required for the project. Since these outputs are obtained from the modified data, they do not contain valid results that can be used by the researcher.

However, once researchers complete this task, they can ask BPLIM staff to rerun their scripts, this time using the original confidential data. These outputs are then subject to standard output control checks for confidential data and delivered to the researcher.

Replication App

When BPLIM reruns the code developed by the researchers on the original data it is fact doing an exercise of replication. To improve our workflow, we decided to raise awareness of our researchers for the need to implement good practices in reproducible research. First, we need to be assured that the computing environment used by the researcher on BPLIM's external server was identical to that used by BPLIM staff when reproducing the code. Thus, for the case of researchers who work with open-source software, we have been incentivizing them to work with Singularity containers (for more information on these, see our GitHub repository: https://github.com/BPLIM/Containers).

This facilitates our work because we are sure that our reproducibility check is implemented in the same self-contained environment that was used by the researcher.

More recently, we have worked on shifting the burden of the reproducibility check to the researcher itself. We have developed an application targeted mainly at researchers who use BPLIM's (modified) confidential data sets, but we hope to eventually convince other users to take advantage of it. In our GitHub repository, we make available the source code (see https://github.com/BPLIM/ReplicationApp).

3.3.2 Output Control

3.3.2.1 Tools and resources for output checking at the RDSC of Bundesbank

Hariolf Merkle and Christian Hirsch, Bundesbank

The RDSC of Bundesbank provides access to most of its microdata for scientific research via workstations in a secure environment. This means that all output leaving the secure environment has to be checked for compliance with the RDSC's rules for visiting researchers.
How time-consuming this output checking depends, firstly, on the complexity of the microdata structure. Complexity hinges on, among other aspects, (i) the number of identifier to be protected, (ii) how often these identifiers occur in a dataset, and (iii) the number of lines in the dataset. The RDSC offers both microdata with simple (e.g. surveys and balance sheet information) and complex data structures (e.g. loan or transaction data). In addition, complexity could increase further because research projects regularly combine microdata in projects.

The second determinant is the experience of the researcher working with the microdata. Generally, experienced users need less time to perform output checking compared to users doing this for the first time. The RDSC provides a combination of technical reports and tools to assist researchers with checking adherence regarding statistical disclosure control (SDC) rules. The provided materials need to take into account heterogeneous user groups and data as well as different statistical applications. To achieve this, all technical reports are written with a specific user need in mind using the Diataxis framework (https://diataxis.fr/) as guidance. Examples are the technical report "Rules for visiting researchers at the RDSC" [8], which is an information-oriented reference or the technical report "Statistical Disclosure Control (SDC) for results derived from aggregated confidential microdata" [9] and the technical report "Statistical Disclosure Control (SDC) for results derived from combined confidential microdata" [9] which are learning-oriented tutorials.

The RDSC also provides two different types of tools to assist researchers with checking adherence regarding SDC rules. The first set is intended to inform researchers directly regarding the admissibility of results of their analysis. An exemplary case is providing a warning in case a single result does not exhibit the minimum number of underlying identities. These tools are currently implemented in R (sdcLog) and Stata (.ado-files). The second type of provided tools is the RDSC's output submitter to check the results being verifiable and inform the data user in case the requested output will not be able to be released. Both types of output checking tools provide guidance to the researcher to only submit output to the RDSC fulfilling a certain level of SDC-requirements right before the actual checking by the RDSC takes place. As a result, RDSC employees process output checking more quickly due to recognizable procedures and proven functionality.

Tool Overview:

Name	Description	Author	Availability
nobsdes5, nobsreg5	.ado-files for Stata	Harald Stahl	Freely available: https://www.bundesbank.de/en/bundesbank/research/rdsc/your-research-project-at-the-rdsc/output-and-publication-checking-618052
sdcLog	R-Package	Matthias Gomolka [aut, cre], Tim Becker [aut], Pantelis Karapanagiotis [ctb]	Freely available: https://cran.r-project.org/web/packages/sdcLog/index.html
Output Submitter	Stand-alone software	Sebastian Seltmann	Not freely available (tailor-made for RDSC)

3.3.2.2 Automated checking of research output (ACRO)

Marco Stocchi, Eurostat

Eurostat introduced two tools (namely "ACRO" and "S-ACRO") suitable to perform automatic checking of research output, developed by a team of experts of the University of West England. Both tools are implemented using high-level programming languages, popular in the research community. They offer user-friendly interfaces and minimum training required; they can be installed in research computation environments with low administration burden. The purpose of the tools is to reduce (as much as possible) the error-proneness and the overall workload of the output checking activities by automating several rule-based output checking computations, otherwise traditionally performed manually by the designated officials.

The first use case of the ACRO tool is its deployment in a *KIOSK* information system for remote access to European microdata, developed and maintained jointly by Eurostat and the Directorate General for Informatics of the European Commission. Researchers who already used ACRO have provided Eurostat with valuable feedback, encouraging the development of extra features and the need for extending the compatibility of the tool with different programming languages.

According to the feedback received, Eurostat steered the project to further enhance the output checking abilities of the tool (such as risk analysis on plots) and to enable researchers to use several other technologies (such as "R" and "Python").

Both ACRO and S-ACRO source codes are open and made available in public Git repositories, to the benefit of the whole community of practitioners in official statistics.

3.3.2.3 Measures to ease code reproducibility for output control

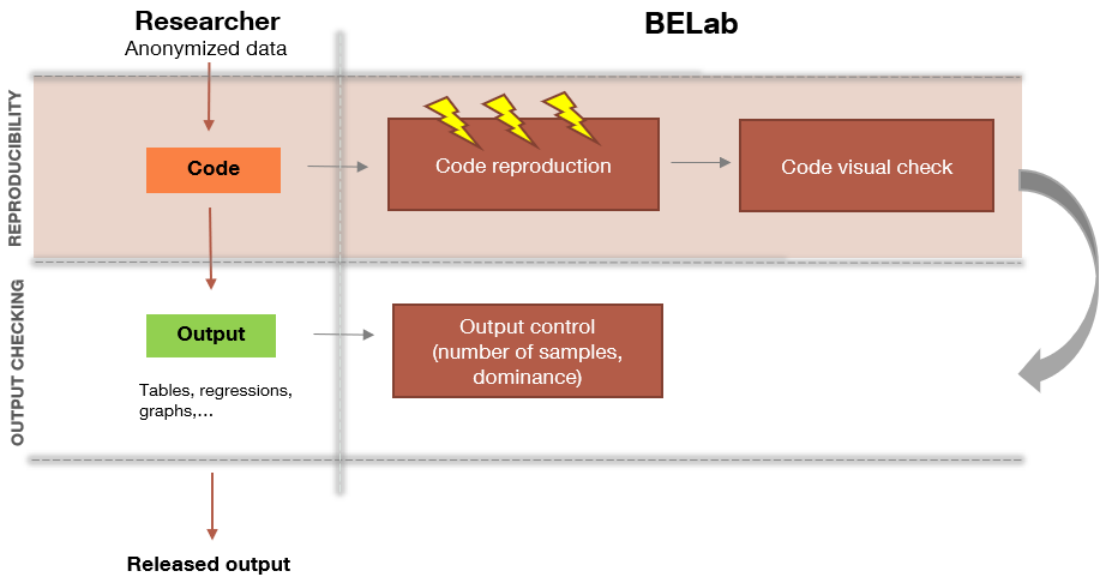
Ricardo Arcos and Emma Pérez, Banco de España

This use case first presents the operating context and circumstances of the data laboratory of the Banco de España (BELab) when undertaking the tasks of reviewing and reproducing the information extractions requested by external researchers.

Reproducibility of results is an essential requirement to ensure the accuracy and compliance of anonymization and aggregation rules of the extracted results, and should be easily and quickly performed by the staff.

The working environment that BELab used to provide to researchers included access to the complete dataset (all years and variables) and freedom to manage the folder structure. In this setting, they carried out their research producing both codes and outputs that are reviewed by BELab's staff to ensure their reproducibility and safety for extraction according to the output control rules (see Figure 9).

Figure 9. BELab output control process.



This setup combined with those of the researchers themselves (variability of programming skills and poor performance in complying with rules and recommendations) added to impractical user guides, made the output review process and, in particular, the reproducibility of the results, a very tedious and time-consuming process.

The implemented solution revolves around three elements:

- Fixed folder structure: aligned with the nature of the elements used in research projects (data, code, results…).
- New user guides: more explanatory and with an eminently practical nature.
- Example code: templates designed in the three available programming languages (R, Stata and Python) to provide practical solutions to researchers' organizational problems.

3.3.2.4 Output control practices in the Spanish Survey of Household Finances

Cristina Barceló, Banco de España

We describe the practices in the Spanish Survey of Household Finances (EFF) for allowing researchers to use confidential information in their studies without revealing relevant information to identify households. To avoid the disclosure of household identity, we remove from the survey key information for identification (month of birth, geographical information and place of birth, that is to say, whether the household member is a native or an immigrant). This unrevealed information may be very relevant for economic research and policy evaluation, although geographical information in the survey is not representative of the Spanish population.

Therefore, we allow researchers to request for the use of confidential information by executing their research programs in remote by the staff at the Banco de España. Researchers can make use of the usual software for handling microdata in their research (Stata, R, Python and Matlab). We give researchers detailed instructions on how to prepare their programs and data files, and we encourage researchers to make as many calculations as possible and needed in their project to reduce the number of iterations with the Banco de España. For this purpose, we provide relevant information for the analysis of confidential variables (variable names, their storage format, their possible values and a brief description of the variable), instructions on how to allocate their own data, how to link their data with the confidential information, and the structure of the folders in their programs to store data and their output, among other guidelines.

To facilitate the use of confidential geographical information, variables are coded with the same classifications as those of the National Statistics Institute (INE), and we provide a code of -9999 for missing information in confidential data. Users can export their results in any format of general use (ASCII, Excel, pdf. Latex, images, etc.), but they cannot bring log files with them.

In order not to reveal household identity, users are not allowed to calculate some descriptive statistics based on measures of position, such as percentiles and maximum and minimum values. We allow the use of means, correlations and variances conditional on a group of geographical regions if they are representative. The results must be computed using a sufficiently high number of observations to be considered representative or non-biased (especially in the case of non-representative information in the EFF such as geographical information) in order not to bring sensible conclusions. This is the most time-consuming task for our staff in the output control, i.e. to check a proper use of confidential information.

3.3.3 Data Sharing and Privacy Enhancing Technologies (PET)

3.3.3.1 Utility and confidentiality assessment of synthetic financial data - Pilot in collaboration with the European Commission

Eugenia Koblents, Banco de España

This use case describes the participation of Banco de España in the synthetic data pilot launched by the European Commission in 2022 to support the creation of the future EU

Digital Finance Platform's Data hub. The goal of this pilot was to evaluate the accuracy and privacy of the generated synthetic data in order to determine the utility of this technology for several purposes, including research, generation of statistics, and test exercises in IT developments and auditing, among others.

Two datasets currently available to researchers in BELab data laboratory have been used in the pilot. On one hand, the CBI microdata set containing economic and financial variables built from the annual accounts reported by Spanish non-financial firms, has been used for data utility assessment. On the other hand, the CIR microdata set containing very sensitive information on loans extended to legal entities resident and non-resident in Spain, has been mainly used for confidentiality assessment. The main outcomes of the pilot are the following:

- The synthetic data generation software allows to accurately reproduce univariate and bivariate distributions as well as the correlations present in the real data.
- However, the provided fit-for-purpose utility metrics do not guarantee high utility for any analysis, which needs to be evaluated for each dataset and use case. In particular:
  - Outlier suppression heavily affects summary statistics.
  - Linear and non-linear relationships among variables are often not preserved.
- Merging synthetic datasets with different, correlated variables is not possible with the technology used.
- No significant privacy issues have been identified. Only a very low number of low-dimensional unique matches have been identified, which should be supressed to minimize reputational risk.

Figure 10 shows a comparison between the percentual imbalance of five linear balance equations, computed from real (left histograms) and synthetic (right histograms) data. In the real dataset most deviations are below 0.2% while in the synthetic dataset many samples present deviations above a 20%. Only 6% of samples present an imbalance below 10% for these five linear relationships (while many more exist).

Figure 10 Comparison between the percentual imbalance of five linear balance equations, computed from real (left histograms) and synthetic (right histograms) data.

This work has been presented at the 5th WCARS Conference [3] and at the IFC-Bank of Canada Satellite Seminar on Granular Data in 2023 [4].

3.3.3.2 *Towards a shared European Statistical System infrastructure for collaborative confidential computing*

Fabio Ricciato, Eurostat

Eurostat is developing the concept of a shared infrastructure to enable members and partners of the European Statistical System (ESS) to perform collaborative confidential computing tasks on demand. Such infrastructure is currently being referred by the name Multi-Party Secure Private Computing-as-a-Service (MPSPCaaS). It would be based on Input Privacy technologies (also known as "Secure Private Computing" technologies) at both software and hardware levels, possibly involving a combination of Trusted Execution Environment (TEE) and Secure Multi-Party Computation (SMPC) based on secret sharing. In the proposed model, the MPSPCaaS system would allow two or more "input parties" to launch computation tasks on their confidential data and deliver the final result to the intended "output parties" with no need for the data holders to transmit their data in intelligible form neither to each other nor to other third party. The proposed model would not replace but rather complement existing data sharing agreements, offering a viable alternative in scenarios where the traditional exchange of data in intelligible form would not be accepted (e.g, due to the very high risk that transmission of plain data would involve). The envisioned MPSPCaaS system will incorporate technological and organizational components combined to enforce "by design" the GDPR principles data minimization, purpose limitation, storage limitation, integrity and confidentiality, and lawfulness. From a legal perspective, the MPSPCaaS would offer a ready-to-use consistent set of supplementary technical and organizational measures, as required by GDPR Art. 89.

In 2023 Eurostat has launched an open call for tender for the "*specification, feasibility analysis and prototype demonstration*" of a MPSPCaaS that is currently under evaluation. The resulting project is expected to start in Q1/2024 for a total duration of 24 months. In the final phase of the project, Eurostat plans to invite interested partner organisations to carry out usability test and pilots, on a voluntary basis, based on the system prototype developed within the project. Future updates about the project will be made available on the PET4OS webpage.

4 Lessons learnt on SDC applied to a Research Data Center. Questions and answers

Ensuring the confidentiality of data is one of the most relevant aspects in the creation and operation of RDCs in any institution.

The "Five Safes" framework [12] is a set of principles that allows data centers to provide secure access to data for research. The framework originated in the UK ONS (Office for National Statistics) and has become a good practice in data protection.

1. Safe data
2. Safe projects
3. Safe people
4. Safe settings
5. Safe outputs

The WG on SDC has been working to learn about the practices carried out in the participating institutions regarding points 1. Safe data and 5. Safe outputs. The WG wants to go one step further and establish recommendations that help other data services comply with these two points. Below are the questions that a RDC starting its activity can ask itself in relation to these two points, and example answers, based on the experience of the participants in this WG.

4.1 Safe data

What techniques can RDCs use to maintain data confidentiality? Mainly techniques of suppression or obfuscation of direct identifiers such as ID numbers, names, addresses, and other identifiable information (primary anonymization). In some cases, secondary anonymization techniques (like global recoding, local suppression, PRAM, top-bottom coding, micro-aggregation, and perturbation) are also applied to ensure that individual respondents cannot be re-identified from the primarily anonymised data.

Can different anonymization techniques be applied depending on the mode of access to the data?
Yes, the following main modes of access can be distinguished: on-site access, remote access, remote execution, sharing of micro data files and publication of public use files. In some institutions, different anonymization techniques are applied, depending on the mode of access to the data.

Is it possible to match different datasets if they are anonymized?
Yes, as long as the same procedure starting with the same seed is used to anonymize the common identifier of the different datasets.

Is it necessary to guarantee that data cannot be indirectly re-identified?
Yes, for some types of microdata sets. It also depends on the legislation behind each dataset. The legal departments of the institutions have to analyze and determine the level of anonymization required.

What techniques do RDCs use to guarantee that data cannot be indirectly re-identified (secondary anonymization)?
The techniques mainly used are: global recoding, local suppression, PRAM, top-bottom coding, micro-aggregation, and perturbation.

Are there any software tools that implement secondary anonymization techniques?
Yes, applications such as sdcmicro or mu-argus can be used, but they may be insufficient for anonymizing time series or for combining categorical and numerical variables. If this is the case, the final solution must be complemented with proprietary software.

4.2 Safe outputs

Is it necessary to check every result?
Every request to release an output should lead to that output being checked by the RDC staff to prevent confidential data from being released. However, researchers tend to

produce more results than those requested for release and published. That latter output has not to be checked. Some institutions perform output checking on only a sample of results. There are also institutions not doing it at all, relying on self-output checking by researchers.

Is it convenient to create a fixed folder structure for researchers to organize their project?
Yes, it is very convenient because it facilitates the subsequent reproducibility of the output by the RDC staff and helps researchers organize their work in line with appropriate criteria.

Is it convenient to provide guidelines to researchers on how to organize their code?
Yes, it is very convenient to provide guidelines or trainings on best practices for organizing their code so that it is easily reproducible and understandable by a third party.

What requirements does a good output have to meet to facilitate the work of the output checker?
A good output must be easy to understand (clear labels on graphs etc.) and well explained (methodology and interpretation of results). By contrast, bad outputs contain little or no explanation about how the results were achieved and what they represent, and they are often little more than log files created by analysts as part of their daily work.

Is it convenient to establish limits on the volume of output to be reviewed per project?
Yes, in this way the researcher is obliged to send only the output necessary for the publication of the results of their project and not to add unnecessary output whose review consumes resources of the RDCs. However, each RDC should decide whether to apply this restriction or not.

Can output checking be fully automated?
Some RDCs have developed applications that allow partial automation of output review. However, the intervention of RDC staff will always be necessary for a validation.

What statistics are usually allowed to be released in the final output?
Descriptive statistics presented in different ways (e.g. tables, histograms, charts) are generally accepted. They can include a measure for the center of a distribution (mean, median or modus), a measure of the shape of the distribution (variance, standard deviation, skewness, kurtosis), and a measure for how often a certain combination of attributes is observed (frequencies, relative frequencies and counts); percentiles and correlation coefficients are also accepted. Maximum and Minimum values are usually not accepted because they refer to single observations [11].

What rules are required to ensure that the released output is safe?
There are different rules depending on the statistical indicator. For example, in the case of tables, a minimum number of observations per cell is required and the dominance rule must be met to prevent an observation from a cell from having a very high weight. The limits for these rules are defined by each RDC based on the specific characteristics of the dataset.

Is it convenient to review the reproducibility of the code?
It depends on the mode of access. In the modes of access in which real data is accessed (on-site access or remote access), it is convenient to verify the reproducibility of the code

to guarantee the code creates the output the researcher wants to release, there is no malicious code (false number of records, etc.), and the provided output meets aggregation rules and is safe to be released (num_samples, dominance).

What are the programming languages commonly utilized by researchers to produce their output?
STATA is the most used language by researchers, followed by R and Python. The least used are MATLAB and SAS.

What output formats are allowed?
CSV and Excel are widely allowed to release results, followed by images, PDF and latex.

5 Conclusions and next steps

This document summarizes the main outcomes of the INEXDA WG on SDC. The group's goal was to review and discuss current SDC needs and procedures used among INEXDA members and to promote knowledge sharing and harmonization within INEXDA in the area of SDC. We consider these goals have been successfully achieved thanks to a productive collaboration among all participating institutions. In particular, the in-person meeting allowed to better understand the procedures and challenges faced by other institutions and has been a very positive experience. The survey conducted by the WG revealed that while there are similarities in the type of controls implemented in each RDC, differences persist in the quantitative rules and the statistics that researchers are allowed to release. This WG represents an initial step towards approaching harmonization in the field of SDC within INEXDA.

However, there is still room to reach a higher harmonization in the area of SDC within INEXDA. This was one of the objectives identified in the mandate of this working group. Furthermore, the RDCs works and experiences are constantly evolving in relation to all aspects related to SDC techniques. Addressing both, harmonization and the exchange of experiences, we propose organizing a virtual follow-up meeting annually. Additionally, every three years, a workshop could be held with the aim of sharing the most outstanding innovations in this field among all participants.

Annex 1: Glossary

Five Safes Framework: decision-making framework used to manage and protect confidential or sensitive data. It is commonly applied to research access to statistical data held by government agencies and data archives. The framework considers five dimensions: projects, people, settings, data, and outputs, aiming to ensure safe and responsible data use. These dimensions help address questions related to appropriateness, trustworthiness, access control, disclosure risk, and statistical results.

INEXDA (International Network for Exchanging Experience on Statistical Handling of Granular Data): international cooperative project involving central banks, Eurostat, and other organizations. Its goal is to exchange experiences related to the statistical handling of granular data for research purposes. The network facilitates collaboration and knowledge sharing among institutions dealing with detailed data.

Output checking refers to the process of assessing research results based on microdata files available in Research Data Centers (RDCs). It ensures that the disclosure risk of these outputs is minimized while maintaining data utility. Techniques such as compliance, consent, control, and auditing are used to verify that the released data adheres to privacy and confidentiality standards.

Primary Anonymization (PA): involves transforming raw data to remove personally identifiable information (PII) or making it less identifiable. Techniques include pseudonymization (replacing identifiers with placeholders) and generalization (grouping individuals to reduce re-identification risk). PA aims to prevent direct identification of individuals.

Privacy Enhancing Technologies (PETs): technologies that uphold data protection principles by minimizing personal data use, enhancing data security, and empowering individuals. They allow users to protect their personally identifiable information (PII) while interacting with online services. PETs include techniques like pseudonymization, informed consent, and transparency.

Real Data: refers to actual data collected from real-world sources, such as surveys, administrative records, or transaction logs. It contrasts with synthetic or simulated data generated for research or testing purposes. Real data often contain sensitive information and require careful handling to protect privacy and confidentiality.

Research Data Center (RDC): secure facility where researchers can access and analyze sensitive data, such as government surveys or administrative records. RDCs provide controlled environments to ensure data privacy while enabling valuable research. Researchers must follow strict protocols within RDCs to prevent unauthorized disclosure.

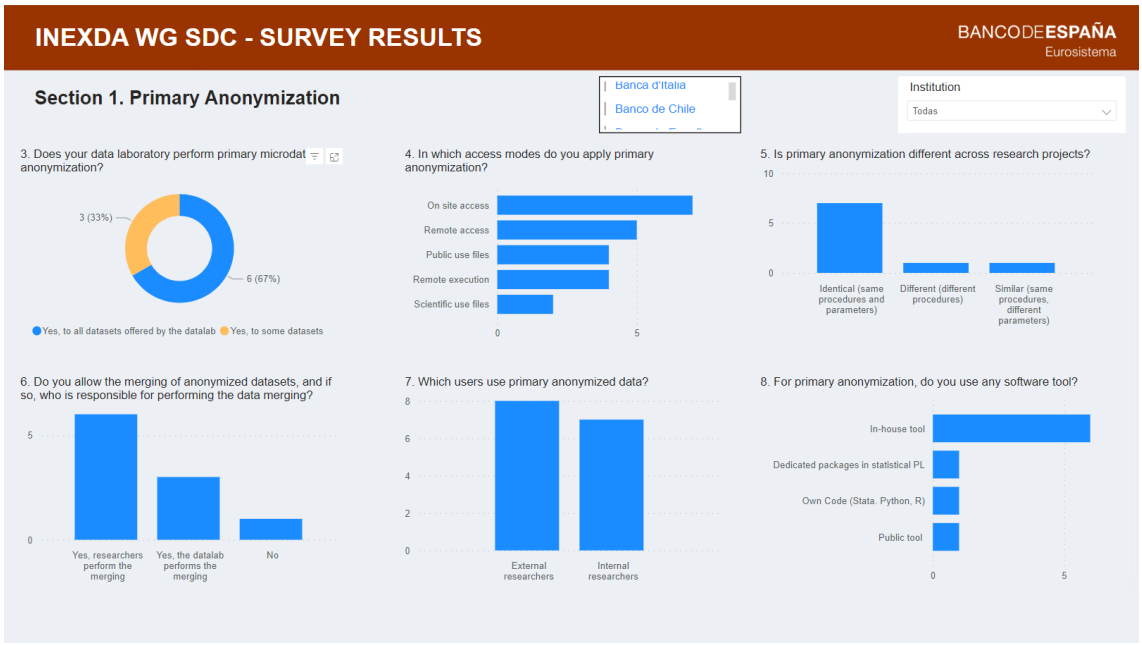
Secondary Anonymization (SA): involves further enhancing the anonymity of data that has already undergone primary anonymization. It aims to reduce the risk of re-identification by applying additional techniques, such as perturbation or aggregation. SA ensures that even if someone gains access to the data, they cannot easily link it back to specific individuals.

Statistical Disclosure Control (SDC): refers to a set of methods and techniques used to protect sensitive data from unauthorized disclosure while maintaining data utility for statistical analysis. SDC includes strategies like microdata anonymization, and output control. Its goal is to balance privacy protection and data usability.

Annex 2: Dashboard SDC Survey

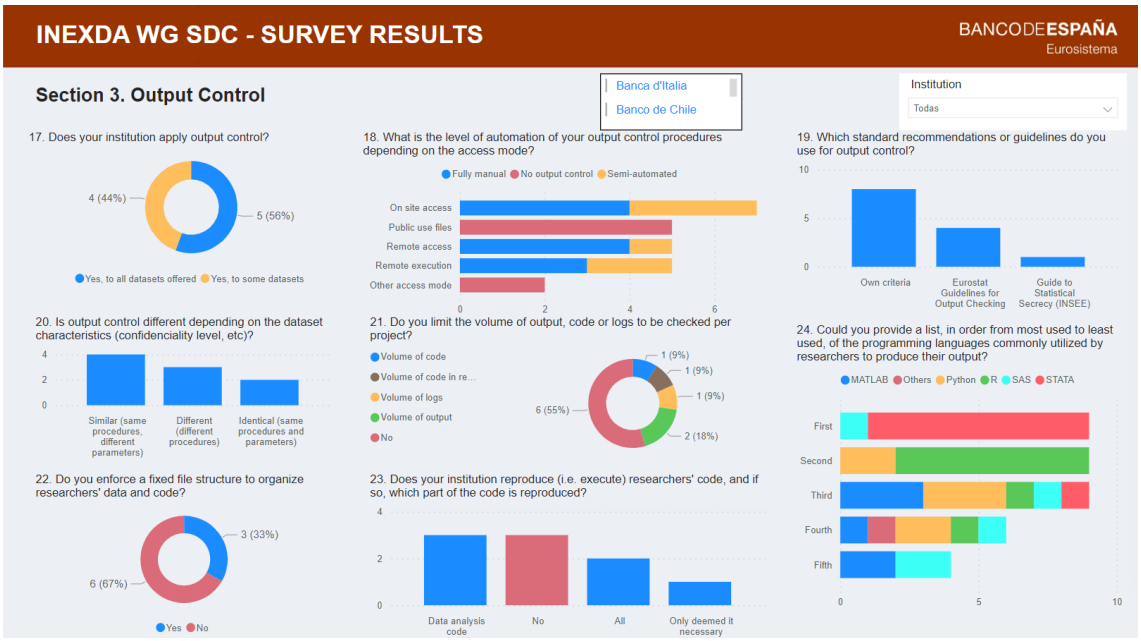
The interactive dashboard is available in this link. Below are the static images of the dashboard for the different sections of the survey

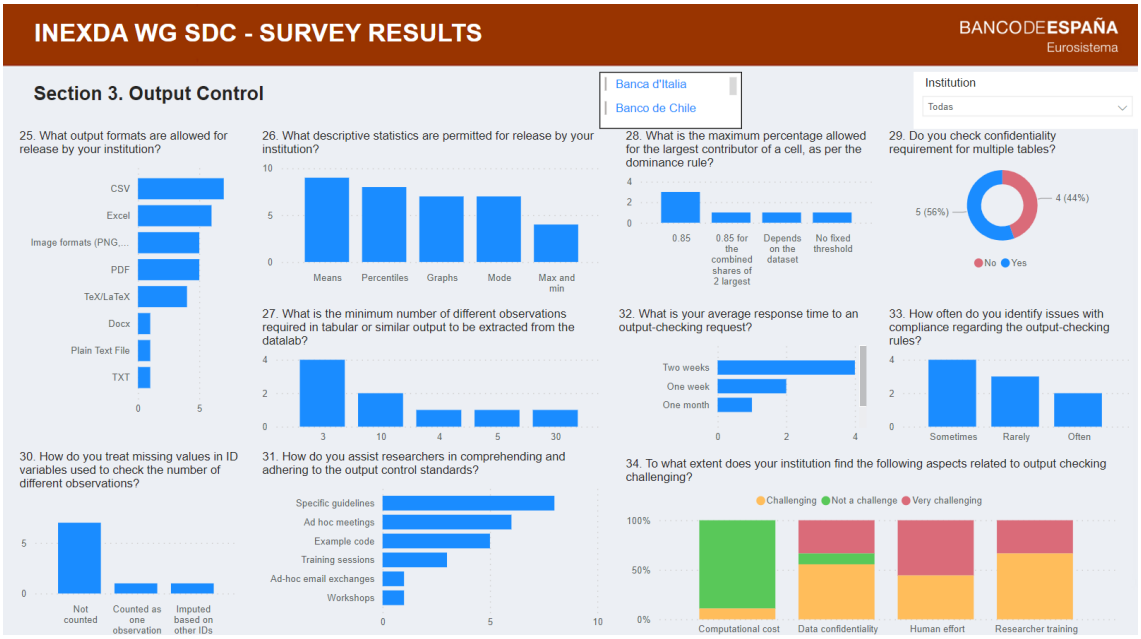
Section 1 Primary Anonymization



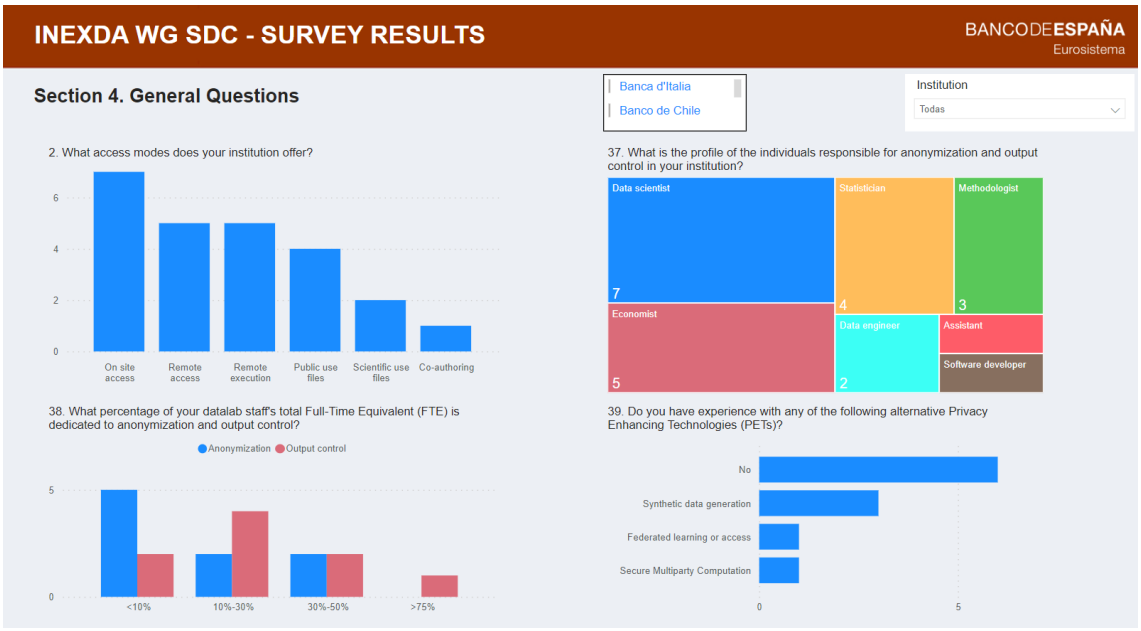
Section 2 Secondary Anonymization

Section 3 Output Control





Section 4 General Questions



References

[1] Koblents, E., & Megia, A. L. (2023). Joint secondary anonymisation of categorical and numerical variables in sensitive time series microdata-novel approach for Statistical Disclosure Control of a sensitive microdata set published in BELab data laboratory. IFC Bulletins chapters, 58.

https://www.bis.org/ifc/publ/ifcb58_09.pdf

[2] Koblents, E., & Megia, A. L. (2023). Joint secondary anonymization of categorical and numerical variables in sensitive time-series microdata. Privacy in Statistical Databaes Conference, 2022.

https://crises-deim.urv.cat/psd2022/index.php?m=program

[3] Koblents, E. Utility and confidentiality assessment of synthetic financial data. 56th World Continuous Auditing and Reporting Symposium, 2023.

https://www.bde.es/f/webeventos/Eventos/56_WCARS_Programa.pdf

[4] Koblents, Eugenia. Utility and confidentiality assessment of synthetic company data, IFC-Bank of Canada Satellite Seminar on Granular Data, 2023.

https://www.bis.org/ifc/events/ifc_230715_agenda_final.pdf

[5] Bender, S. and P. Staab, P. (2015). The Bundesbank's Research Data and Service Center (RDSC), Gateway to treasures of microdata on the German financial system. IFC Bulletin No 41, Irving-Fisher Committee on Central Bank Statistics

[6] Blaschke, J., and Hirsch, C. (2023). On the value of data sharing: Empirical evidence from the Research Data and Service Centre, Technical Report 2023-08 – Version 1.0. Deutsche Bundesbank, Research Data and Service Centre.

[7] Kalckreuth, U. von (2014). A Research Data and Service Centre (RDSC) at the Deutsche Bundesbank – a draft concept. IFC Bulletin No 37, Irving-Fisher Committee on Central Bank Statistics.

[8] Research Data and Service Centre (2021). Rules for visiting researchers at the RDSC, Technical Report 2021-02 - Version 1-0, Deutsche Bundesbank, Research Data and Service Centre.

[9] Blaschke, J., Gomolka, M., Hirsch, C. (2022). Statistical Disclosure Control (SDC) for results derived from aggregated confidential microdata. Technical Report 2022-01, Deutsche Bundesbank, Research Data and Service Centre.

[10] Blaschke, J., Hirsch, C., Kollmann, R. (2023). Statistical Disclosure Control (SDC) for results derived from combined confidential microdata. Technical Report 2023-02, Deutsche Bundesbank, Research Data and Service Centre.

[11] Steve Bond (ONS), Maurice Brandt (Destatis), Peter-Paul de Wolf (CBS) (2010). Guidelines for the checking of output based on microdata research. Data without Boundaries - DwB.

[12] Desai, T., Ritchie, F., & Welpton, R. (2016). Five safes: designing data access for research. Economics Working Paper Series, 1601, 28.