Irving Fisher Committee on
Central Bank Statistics

BIS

12th biennial IFC Conference: "Statistics and beyond: new data for decision making in central banks"

22-23 August 2024

Main outcomes of the INEXDA working group on Statistical Disclosure Control (SDC)[1]

Ricardo Arcos, Ana Esteban, Eugenia Koblents and Emma Pérez,
Bank of Spain

[1] This contribution was prepared for the conference. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

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

Ricardo Arcos, Ana Esteban, Eugenia Koblents and Emma Pérez[1]

Abstract

INEXDA provides a platform for research data centres to exchange practical experiences among their members. In this context, a working group on Statistical Disclosure Control (SDC) was created. The WG, active from June 2023, is run by Bank of Spain. The group focused on two main SDC methods: anonymization and output control. Its primary goal was to outline current SDC practices among INEXDA members and promote knowledge sharing and standardization within the network. This document summarizes the outcomes of this WG.

Keywords: Statistical Disclosure Control, output control, anonymization, INEXDA, research data centre

Content

[1] The authors would like to thank our colleagues of the INEXDA Working Group on Statistical Disclosure Control for their contributions and valuable suggestions and feedback.

1 Introduction

Statistical disclosure control (SDC) is a major challenge faced by Research Data Centres (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. This motivation led INEXDA to propose the creation of a WG on SDC.

The SDC process consists of two main tasks: (1) microdata anonymization, and (2) output control. This WG provided an overview of both tasks. In particular, the main goals were:

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

To achieve these objectives, the working group has carried out the following actions:

1. A survey to identify current SDC needs, procedures, and software tools.
2. A technical session to review specific use cases and discuss implementation details.
3. A plan to address open challenges.

This document presents the main outcomes of the INEXDA SDC WG and is organized as follows: Section 2 summarizes the main outcomes of the WG: an overview of the RDCs participating in the project, the main survey results, and a description of the SDC use cases reported by INEXDA members. Section 3 summarizes the lessons learned, and finally, Section 4 the conclusions and next steps.

2 Main outcomes of the working group

2.1 General overview of the Research Data Centres

Banca d'Italia, Banco Central de Chile, Banco de España, Banco de México, Banco de Portugal, Banque de France, Central Bank of Turkey, Deutsche Bundesbank and Eurostat participated in the WG. This section presents information for each RDC to provide a comprehensive overview of their situation.

Banca d'Italia

Overview: The RDC was established in 2019, as part of the Economics and Statistics Department. It is the single entity centralizing collection and dissemination of 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

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

Banco de España

Overview: Banco de España Data Laboratory (BELab) was established in 2019 as a pilot project under the umbrella of the Statistics Department. 3 years later, it evolved into a consolidated service within a dedicated unit. Its primary objective is to enhance access for external researchers. This access is provided in a controlled environment, ensuring data confidentiality.

Access Modes	On site, Remote, Remote execution, Public use files
Type of researchers	External
Datasets	14
Primary anonymization	Some datasets
Secondary anonymization	Yes
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, foster ties with the academic community to advance the economic research carried out at Banco de México, provide a framework for joint research projects. The EconLab only provide access for academic research.

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

Banco de México

Active Projects	>50

Banco de Portugal

Overview: The Banco de Portugal Microdata Research Laboratory (BPLIM) started its activity in 2016 within the Economics and Research Department. Its primary objective is to facilitate the development of research projects focusing on the Portuguese economy. Through BPLIM, both internal and external researchers gain, in a controlled environment, access to anonymized microdata sets that can be tailored and have the opportunity to utilize the computational resources offered.

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
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, initially in Paris, then in New York in 2018. In 2020, offers a remote secure access to data. 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) which allows cross-referencing with other French institution data.

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

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

Deutsche Bundesbank

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

Eurostat

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

2.2 Survey on SDC. Main results

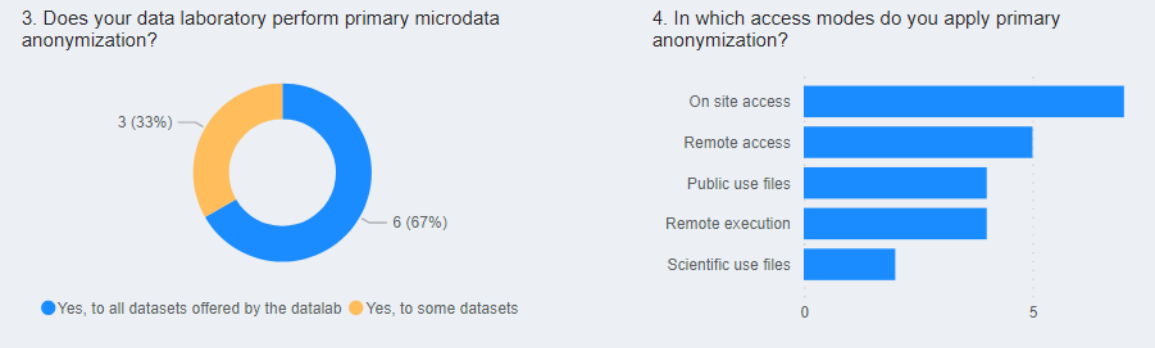
A survey was conducted to identify current SDC needs, procedures, and software tools used for different access modes and levels of data confidentiality. The survey contained a total of 40 questions, grouped into four main sections: primary anonymization, secondary anonymization, output control, and other general questions.

Answers have been compiled into an interactive dashboard that also allows to select and compare different institutions. The main findings are summarized below.

2.2.1 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). PA is usually identical across research projects, performed either with in-house tools and code or either with dedicated packages.

RDC performing primary anonymization and access modes — Figure 1

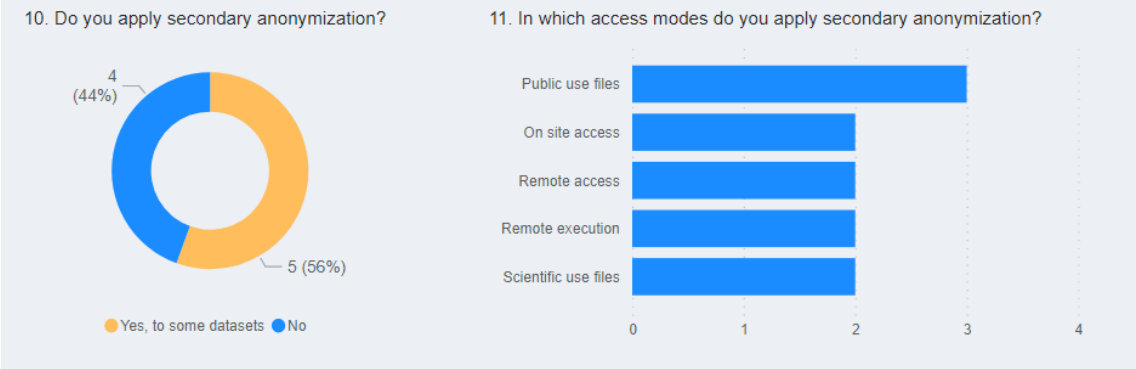


2.2.2 Secondary anonymization

Five out of nine participating institutions apply secondary anonymization (SA) (Figure 2). Techniques vary, each institution using at least two methods. Tools include internal tools, sdcmicro, or mu-argus. Challenges in SA include data utility and confidentiality (Figure 3).

RDC performing secondary anonymization and access modes

Figure 2



Secondary anonymization main challenges

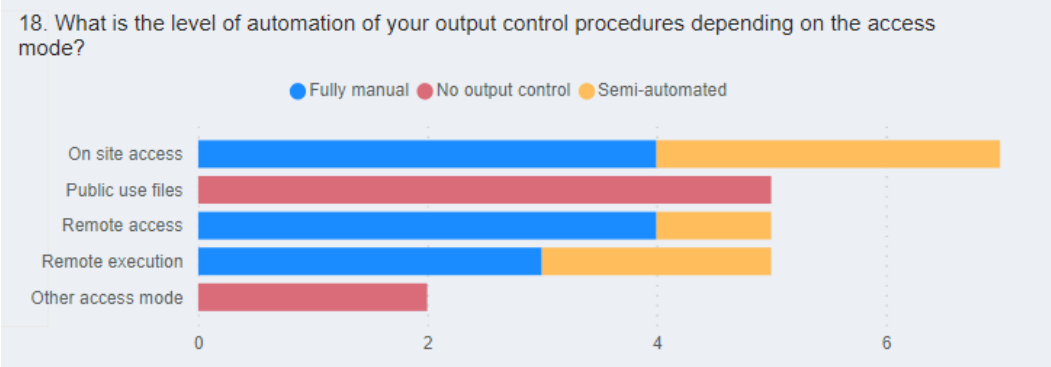
Figure 3

2.2.3 Output control

All institutions apply output control, either to all datasets or only to some, using their own criteria or in combination with DwB guidelines promoted by Eurostat[2]. The process is semi-automated in some institutions but mainly manual (Figure 4).

Automation of output checking process

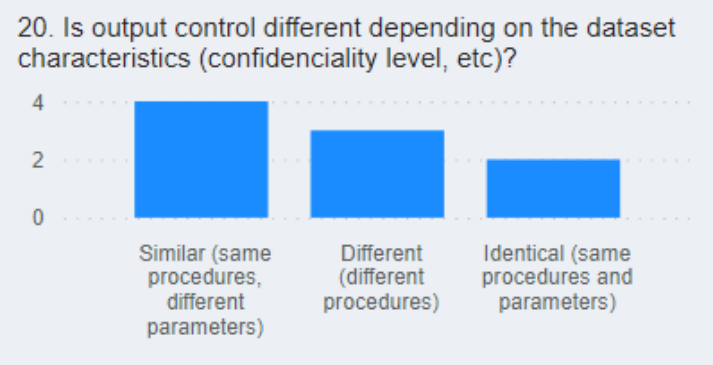
Figure 4



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

The output checking process is identical (same parameters) or similar (different parameters) across datasets (Figure 5). Three institutions vary procedures.

Types of output control across datasets — Figure 5



Most institutions do not limit output/code/log volumes. Three enforce fixed project structures. Half reproduce code selectively; others reproduce whole or part code. Researchers mainly use STATA, followed by R and Python.

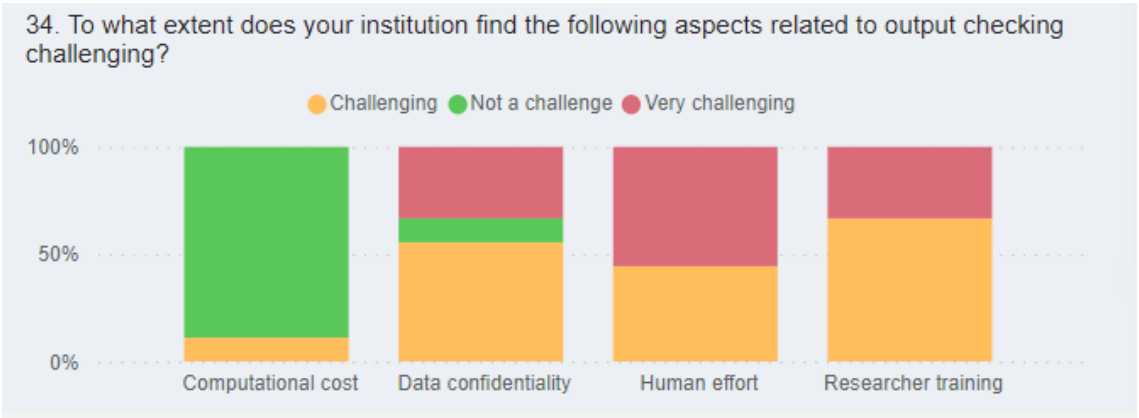
Common output formats are CSV, Excel, followed by images, PDF, and LaTeX. All allow means and percentiles as descriptive statistics; some allow graphs and modes.

Half require a minimum of three observations for extraction; others have higher thresholds. Dominance rules vary: three use 85% for the largest contributor, others for the top two, or no fixed threshold.

Guidelines, ad-hoc meetings, and example code help researchers to know the rules. Response time for output control requests is generally two weeks or less.

Finally, the main challenges are human effort, researcher training, and data confidentiality. Computational cost is not a concern (Figure 6).

Output control main challenges — Figure 6



2.2.4 Other general questions

Most institutions offer multiple access modes. Most provide onsite access with other modes like remote access and remote execution. Some use Public use files (PUFs) and Scientific use files (SUFs).

Professional profiles include data scientists, economists, statisticians, methodologists, data engineers, assistants, and software developers.

Most institutions lack experience with alternative privacy enhancing techniques (PETs), except EUROSTAT, Bundesbank, and Banco de España.

2.3 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.

2.3.1 Anonymization

2.3.1.1 Disclosure avoidance in the Spanish Survey of Household Finances. *Cristina Barceló, Banco de España*

It explains the practices in the Spanish Survey of Household Finances (EFF) for avoiding the identification of the households that participate in this 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 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. For this purpose, it is key to preserve the true distribution of variables and relationships among them. To avoid the disclosure of sensible information for identifying households, confidential information is removed and information in a more aggregated way is provided. Different anonymized household identification codes across different datasets are kept, and store separately public multiply imputed survey data from the files that contain confidential information.

2.3.1.2 Anonymization Algorithm for trade transactions. *Claudia Velázquez, Banco de México*

This use case introduces the anonymization algorithm used to obfuscate the trade database. This algorithm allows the publication of information at a more granular level. 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 covers all possible combinations of data situations.

2.3.1.3 Joint analysis of categorical and continuous key variables for secondary anonymization of sensitive microdata. *Eugenia Koblents, Banco de España*

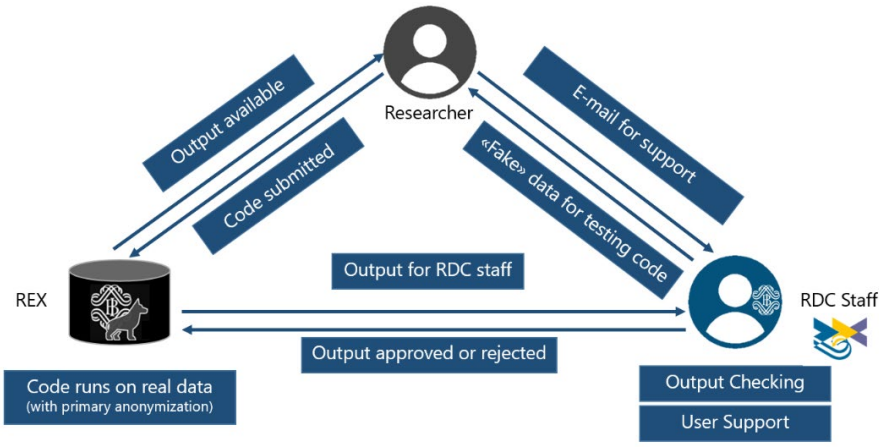
An SDC approach for secondary anonymisation of sensitive time series microdata has been developed at the Banco de España in order to protect confidentiality of Central Credit Register (CCR) dataset. This dataset contains yearly microdata on loans to legal persons, including multiple variables describing 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 CCR dataset against re-identification, while jointly analysing categorical and numerical variables and addressing time series data protection. 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].

2.3.1.4 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).

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

Figure 7



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.

The REX system guarantees confidentiality at three levels:

1. Users: must submit an application to have access to remote execution system.

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 and content.

2.3.1.5 BPLIM's approach to protecting sensitive data. *Joana Pimentel, Banco de Portugal*

BPLIM implemented data access solutions that preserve the confidentiality of the datasets that contain highly sensitive information.

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 it 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.

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

To improve our workflow, we decided to raise awareness of our researchers for the need to implement good practices in reproducible research. For researchers who work with open-source software, we have been incentivizing them to work with Singularity containers (ttps://github.com/BPLIM/Containers).

We have developed an application targeted mainly at researchers who use BPLIM's (modified) confidential data sets and other users, (see https://github.com/BPLIM/ReplicationApp).

2.3.2 Output Control

2.3.2.1 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 secure workstations. Requires output leaving the secure environment to comply with RDSC rules for visiting researchers.

The Complexity of the Microdata structure determines the time required for output checking. Another factors affecting complexity are: (i) Number of identifiers to be protected. (ii) Frequency of identifier occurrence in the dataset. (iii) Number of lines in the dataset. Combining different microdata may increase the complexity.

The second determinant is the experience of the researcher working with the microdata. Generally, experienced users need less time to perform output checking. RDSC provides technical reports and tools for adherence to statistical disclosure control (SDC) rules. Technical reports are tailored to user needs using the Diataxis framework (https://diataxis.fr/). Examples include information-oriented references and 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 (i.e. number of underlying observations). 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.

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)

2.3.2.2 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. They can be installed in research computation environments with low administration burden. The purpose of the tools is to reduce the error-proneness and the overall workload of the output checking activities.

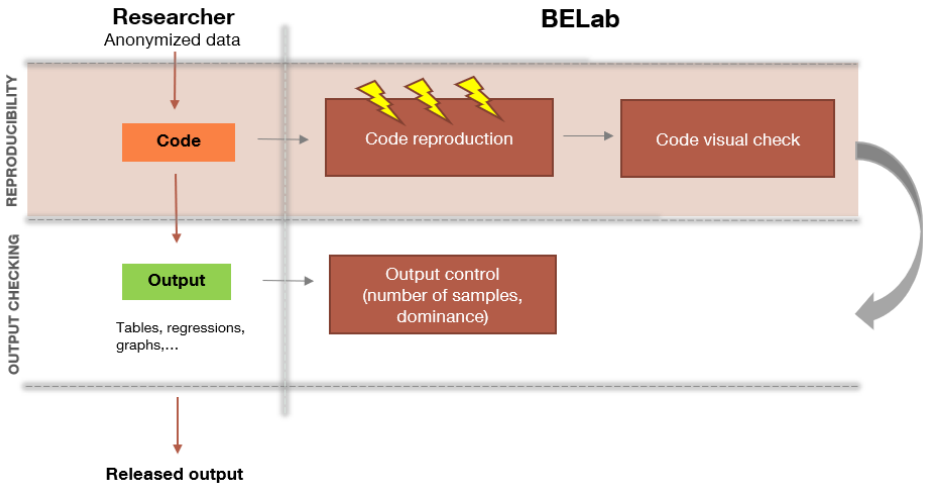
According to the feedback received, Eurostat steered the project to further enhance the output checking abilities of the tool (such as risk analysis on plots) and to enable researchers to use several other technologies (such as "R" and "Python").

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.

2.3.2.3 Measures to ease code reproducibility for output control. *Ricardo Arcos and Emma Pérez, Banco de España*

Reproducibility of results is an essential requirement to ensure the accuracy and compliance of output rules. Researchers carry out their research producing codes and outputs. Both are reviewed by BELab's staff to ensure their reproducibility and safety for extraction according to the output control rules (see Figure 9).

BELab output control process — Figure 9



This setup, combined with different programming skills of researchers made BELab implemented solutions around three elements to facilitate the task of reproducibility:

- Fixed folder structure.
- Explanatory and practical nature user guides.
- Example code: templates designed in R, Stata and Python.

2.3.2.4 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 without revealing relevant information. To reach that, we remove from the survey key information for identification (mostly geographical).

Researchers can request access to confidential information by executing their research programs remotely. Usual software is available for handling microdata. Researchers receive detailed instructions

on preparing programs and data files to minimize interactions with Banco de España. To facilitate the use of confidential geographical information, variables are coded with the same classifications as those of the National Statistics Institute (INE). Users can export their results in any format of general use, but they cannot bring log files with them.

To protect household identity, users cannot calculate certain descriptive statistics based on measures of position. Allowed descriptive statistics include means, correlations, and variances conditional on geographical regions if they are representative. The results must be computed using a sufficiently high number of observations in order not to bring sensible conclusions. This is the most time-consuming task for our staff in the output control.

2.3.3 Data Sharing and Privacy Enhancing Technologies (PET)

2.3.4 Data Sharing and Privacy Enhancing Technologies (PET)

2.3.4.1 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.

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.

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].

2.3.4.2 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 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.

The project is expected to start in Q1/2024 for a total duration of 24 months. Future updates about the project will be made available on the PET4OS webpage

3 Lessons learnt. 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 is a set of principles that allows data centres 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.

The WG on SDC learnt about the practices carried out in the participating institutions regarding points 1. Safe data and 5. Safe outputs. The WG wanted to go one step further and established recommendations to help other data services. 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.

3.1 Safe data

What techniques can RDCs use to maintain data confidentiality? Mainly techniques of suppression or obfuscation of direct identifiers (primary anonymization). In some cases, secondary anonymization techniques are also applied.

Can different anonymization techniques be applied depending on the mode of access to the data? Yes; 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.

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.

What techniques do RDCs use to guarantee that data cannot be indirectly re-identified (secondary anonymization)? 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. If this is the case, the final solution must be complemented with proprietary software.

3.2 Safe outputs

Is it necessary to check every result? 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 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, 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? Easy to understand and well explained. By contrast, bad outputs contain little or no explanation about how the results were achieved and what they represent.

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. 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. 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? Measure for the centre of a distribution (mean, median or modus), shape of the distribution (variance, standard deviation, skewness, kurtosis), a certain combination of attributes (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? 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.

Is it convenient to review the reproducibility of the code? Yes, to guarantee the code creates the output the researcher wants to release, there is no malicious code (false number of records, etc.) and output meets aggregation rules.

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

4 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 within INEXDA in the area of SDC. 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.

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.

BANCO DE ESPAÑA
Eurosistema

USO INTERNO

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

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.

International Network of Exchanging Experiences on Statistical Handling of Granular Data

BANCO CENTRAL DE CHILE
DEUTSCHE BUNDESBANK EUROSYSTEM
BANQUE DE FRANCE EUROSYSTÈME
BANCO DE PORTUGAL EUROSISTEMA
BANCO DE ESPAÑA Eurosistema
BANCA D'ITALIA
eurostat
EUROPEAN CENTRAL BANK EUROSYSTEM
BANCO DE MÉXICO
TÜRKİYE CUMHURİYET MERKEZ BANKASI
BANK OF ENGLAND
Office for National Statistics

INEXDA WG ON STATISTICAL DISCLOSURE CONTROL

Why create this WG in INEXDA?

BANCO DE ESPAÑA
Eurosistema

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, goals and tasks

BANCO DE ESPAÑA
Eurosistema

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.

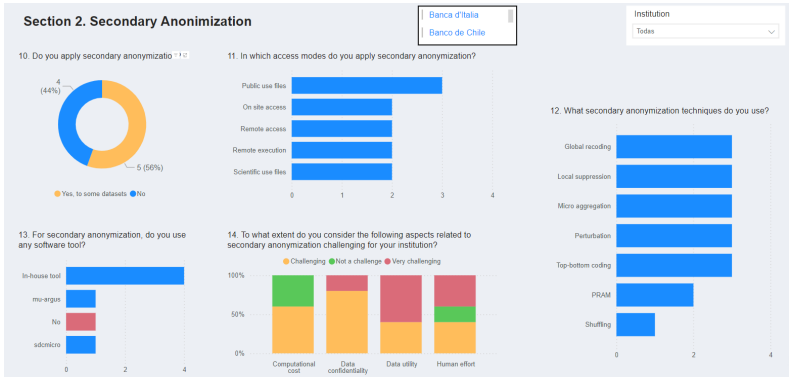
Tasks

- 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.

.

Primary Anonymization

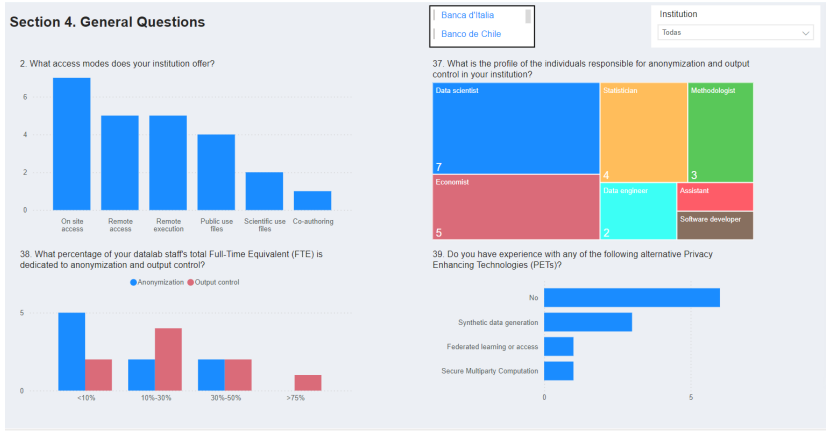
Secondary Anonymization



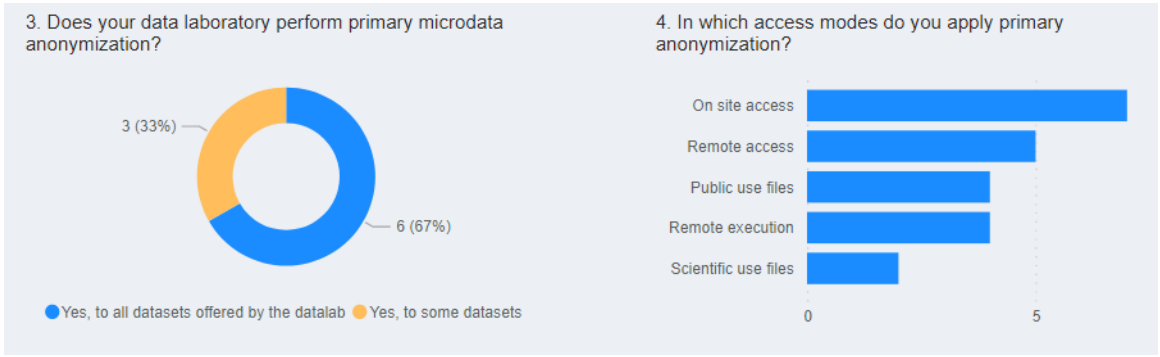
SURVEY CONTENT

Output Control

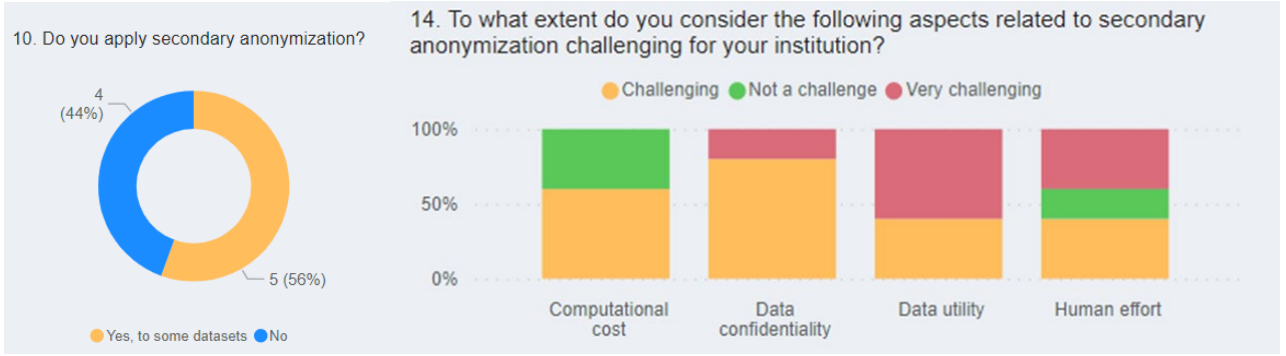
General Questions



Primary Anonymization

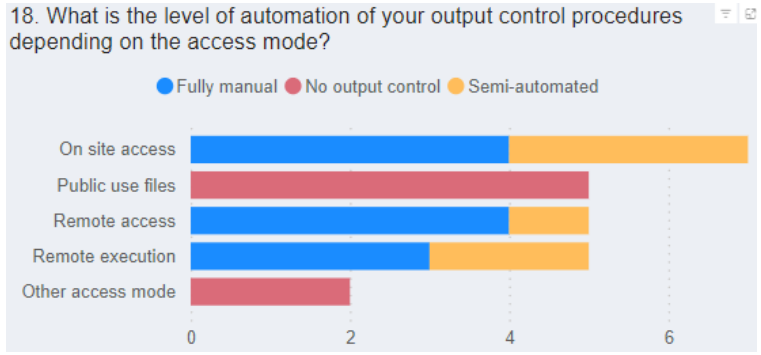


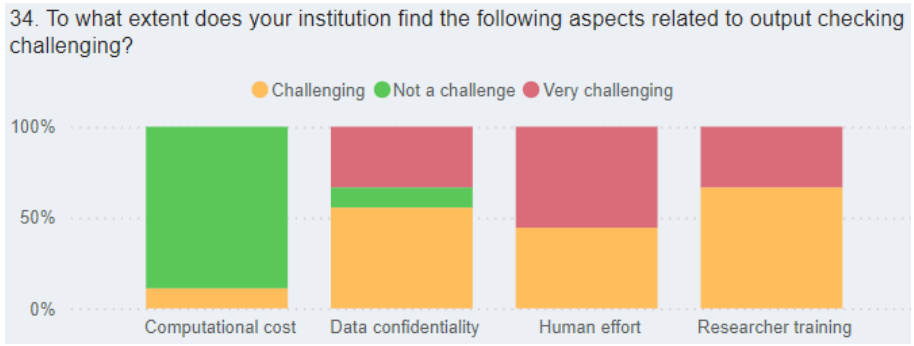
Secondary Anonymization



SURVEY: MAIN RESULTS

Output Control





INEXDA WG ON STATISTICAL DISCLOSURE CONTROL

Use cases

BANCO DE ESPAÑA
Eurosistema

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)

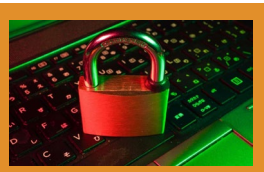
- **Utility and confidentiality assessment of synthetic financial data- Pilot in collaboration with the European Commission** *(Eugenia Koblents, Banco de España)*
- **Towards a shared European Statistical System infrastructure for collaborative confidential computing (***Fabio Ricciato, Eurostat)*

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?**
- **…**

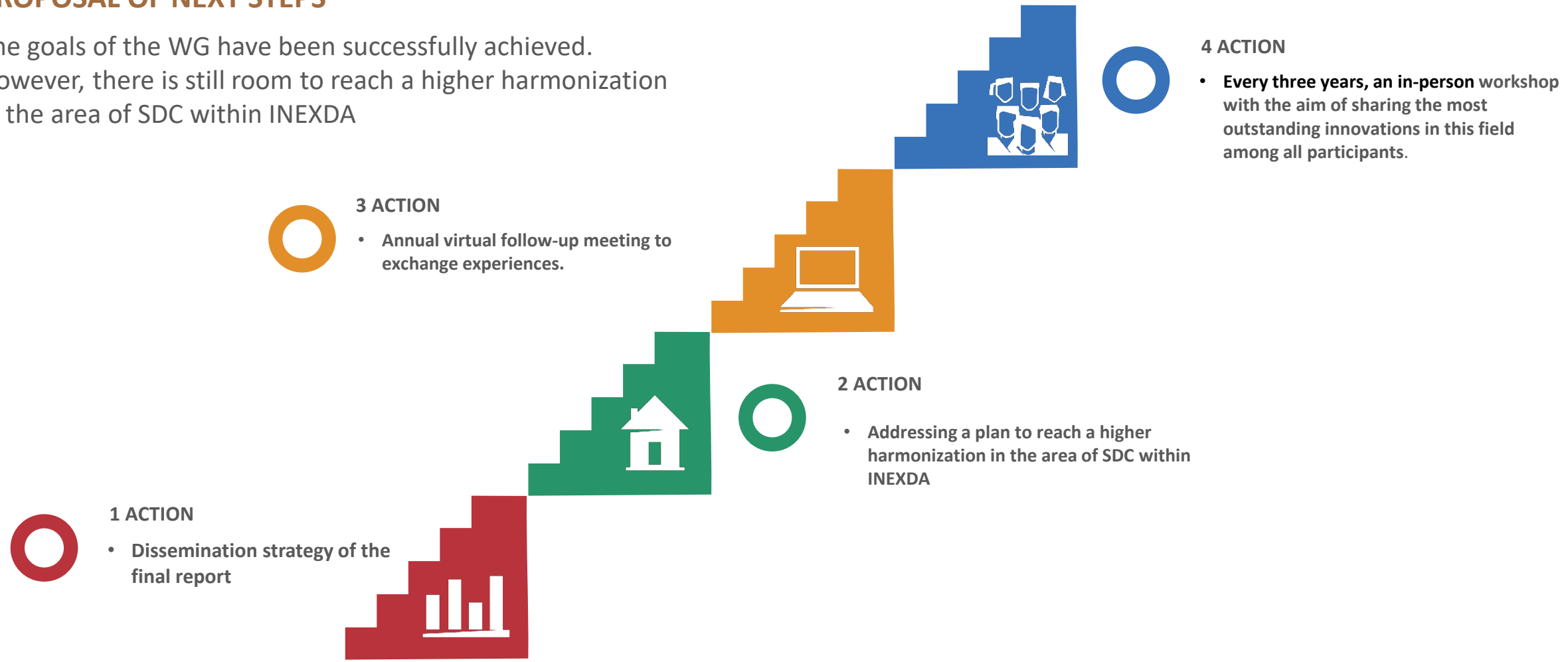
SAFE OUTPUT



- **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.**

BANCO DE ESPAÑA
Eurosistema

THANK YOU FOR YOUR ATTENTION