Irving Fisher Committee on
Central Bank Statistics

BIS

11th Biennial IFC Conference on “Post-pandemic landscape for central bank statistics”

BIS Basel, 25-26 August 2022

Introduction to and application of SDC rules using self-developed tools[1]

Jannick Blaschke, Matthias Gomolka, Christian Hirsch, Sebastian Seltmann and Harald Stahl,
Deutsche Bundesbank

[1] This presentation was prepared for the conference. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the event.

Introduction to and application of SDC rules using self-developed tools

Jannick Blaschke, Matthias Gomolka, Christian Hirsch, Sebastian Seltmann and Harald Stahl[1]

Abstract

To gain access to confidential microdata via a Research Data Centre (RDC), researchers and the output they produce must adhere to strict confidentiality rules. Typically, RDCs maintain explanatory documents and sophisticated tools to assist researchers in checking the compliance with those rules. However, – as a consequence of diverse and very complex requirements – the amount of available information on SDC rules is constantly increasing. We present concrete examples of three self-developed tools to help on-board researchers at the Deutsche Bundesbank without overloading them with information.

Keywords: Statistical Disclosure Control, output control, tools, INEXDA, data sharing, research data centre, data access, research, microdata

Contents

[1] The authors would like to thank our colleagues at the RDSC for their valuable suggestions and feedback. All views expressed in this report are the authors' personal views and do not necessarily reflect those of the Deutsche Bundesbank or the Eurosystem.

1. A brief introduction to the work of Research Data Centres (RDCs)

Much data collected for public purposes is considered to be a public good and should therefore be as broadly accessible as possible. While some of the data may be released to the general public without any restrictions, most data contain some sort of information that needs to be protected, e.g. personal or market-sensitive information.

One way to make this information accessible is to provide fully aggregated data that can be downloaded, e.g. from the website of the data-collecting institution. Fully aggregated data no longer allow direct or indirect identification of the reporting agents, e.g. households, banks or firms. Although fully aggregated data does ensure a very high level of data protection, much valuable information is lost in the process of aggregating the underlying micro data.

However, depending on the legal basis, anonymised micro data characterised by a lower degree of anonymisation than absolute anonymity[2] may be used for independent scientific research. For this purpose, many data collecting institutions have established special access modes that ensure secure data usage by external researchers. Next to Scientific Use Files[3], Research Data Centres (RDCs) are probably the most common of these access modes.

RDCs provide secure on-site access to confidential micro data for the purpose of scientific research (Ritchie, 2017, 2021). After their data access request has been approved by the RDC, external researchers can come to the premises of the RDC and access the micro data in a secure environment, where they have no access to the internet and are not allowed to use their personal laptops or phones. In addition, researchers must ensure that all results derived from confidential data that leave the secure environment of the RDC following a research visit no longer contain any confidential information. This so-called Statistical Disclosure Control (SDC)[4] or output control is mandatory before results can be released.

The price that data providers and data-using researchers pay for this is a certain degree of complexity in accessing data. Starting with the application, to the access modalities, to arguably the most difficult part, the SDC, working in RDCs is often a new challenge for researchers. RDC staff are well aware that the access processes and rules can be complicated and often require RDC-specific knowledge.

This is particularly evident in the SDC. While this topic is essential for successful work in the RDC and good knowledge determines whether or not results can be published, researchers outside the RDCs usually do not encounter it at all. Therefore, new researchers first have to build up this knowledge, even though it is not part of

[2] Such anonymised micro data is often referred to as formally or factually anonymised data. Formal anonymisation describes the deletion of direct identifiers such as names, addresses, and other identifiers (e.g. LEI) so that no direct identification is possible. Factual anonymisation includes additional perturbation measures so that an identification is only possible with an unreasonable investment of time, cost and manpower.

[3] Scientific Use Files are usually factually anonymised datasets that are sent to approved researchers or can be downloaded from a password-protected area.

[4] An example of an SDC document is the Bundesbank's "Rules for visiting researchers at the RDSC" (Research Data and Service Centre, 2021).

their actual research. After building up the required knowledge, researchers also need to correctly apply this knowledge to their specific use cases. RDCs usually provide user documentation with information aiming to support researchers with both tasks irrespective of their level of knowledge. This strategy works when the number of accessible datasets is small and the complexity of these datasets is low.

In recent years, however, this strategy of a few documents that fit all users and all datasets has come under increased scrutiny. One reason for this is an increase in the number of available datasets as well as an increase in the linking possibilities of datasets in RDCs. In addition, datasets are also becoming increasingly more complex and originate from diverse backgrounds. This development makes it more cumbersome for researchers to comprehend and adhere to SDC requirements. It may also increase the need to amend existing requirements to reflect the new datasets.

These developments also have repercussions for how RDCs present information to their users. In this paper, we argue that ever-expanding user documentation is certainly not the most efficient and helpful way to address the growing complexity and diversity of datasets. Instead, more unstructured information by e.g. extending the SDC document increases the likelihood of researchers not finding the information they need when they need it. This in turn increases the likelihood that they will apply rules incorrectly, resulting in the submission of results that are non-compliant with SDC rules and thus rejected by RDC staff.

The objective of this paper is to present three tools that implement different approaches to address these developments. All three tools have been self-developed by the Research Data and Service Centre (RDSC) of the Deutsche Bundesbank. The first tool, the RDSC Landing Page, pre-sorts all available information according to characteristics deemed helpful to the user. The second tool, "nobsdes5 / nobsreg5" and {sdcLog}, helps users apply information by (semi) automating the SDC checking process. Finally, the Output Submitter applies selected SDC rules automatically to the researcher's output, displays additional information if needed, and streamlines the submission process.

This paper is organised as follows. Section 2 discusses approaches to communicating complex information to a heterogeneous target group in the context of SDC. Section 3 presents the three self-developed tools from the RDSC that are meant to support researchers in learning and performing SDC. Finally, Section 4 concludes with a brief discussion and gives an outlook regarding future areas of improvement.

2. How tools support researchers in understanding and applying SDC rules

Researchers in RDCs generally need information about existing SDC rules and how to apply them to their research project. Let us take as an example a researcher who calculates the mean of a distribution and wants to check whether her result complies with SDC rules. Pointing this researcher to a potentially large pile of information would not be an effective way to support her in this situation. Instead, she would need supplementary information that helps her to isolate specific information on the calculation of the mean of a distribution. After that, she would have to filter this information again depending on her personal skills, e.g. previous experience with SDC rules.

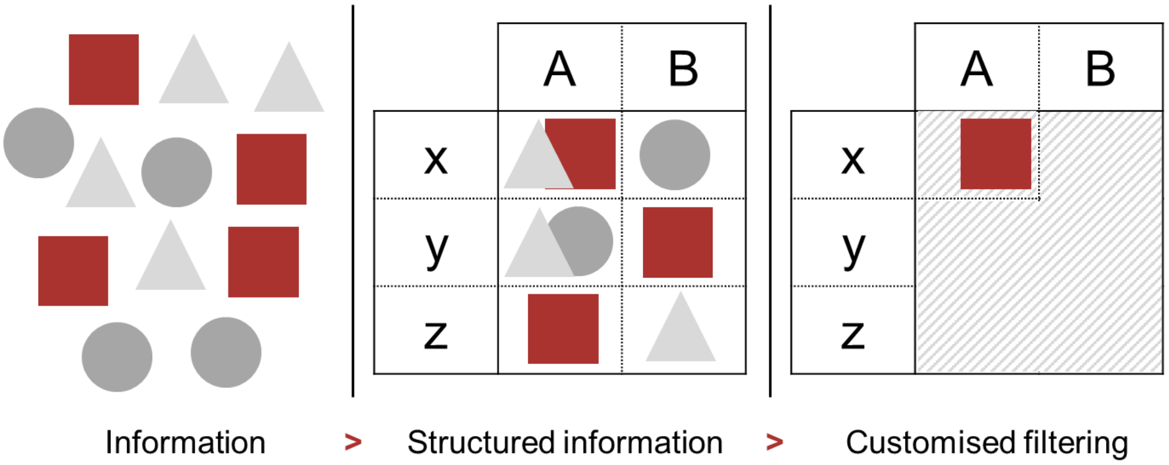
A natural first step to start is to pre-sort all available information into buckets that help researchers find the required documents faster. For example, one could choose categories according to researcher characteristics, the tasks that they perform, and the purpose of the document. Pre-sorting information according to researcher characteristics allows account to be taken of differences in knowledge owing to different levels of experience. At the one end of the spectrum are vastly experienced researchers who require only selective information on how to solve specific problems. First-time users, by contrast, need documents that teach them how to get started. We choose tasks as an additional sorting category because different tasks (potentially) require the application of a different set of rules.

While the first two dimensions are relatively straightforward, the third is often overlooked. Frameworks such as Diátaxis[5] argue that that researchers' success depends both on the theoretical understanding and their ability to apply this understanding to the task. Consequently, documents that cater to these different needs will also look vastly different. Documents aiming to enhance the theoretical understanding of researchers need to be descriptive, state the complete set of rules and explain the rationale behind them (e.g. "explanation" and "reference" documents). In contrast, documents aiming at facilitating the application of rules to a real-world task need to demonstrate how to solve a specific problem and guide researchers through a series of steps (e.g. "tutorials" and "how-to guides").

Pre-sorting heterogeneous information

Structuring different documents* using two dimensions

Figure 1



* The different shapes symbolise the different document types, e.g. legal reference, tutorial, explanation.

Sources: Authors

Panel A and B of Figure 1 depict the transition from unsorted to pre-sorted information, i.e. from unstructured to structured information. For example, the x-axis

[5] "The Diátaxis framework (...) adopts a systematic approach to understanding the needs of documentation users in their cycle of interaction with a product. Diátaxis identifies four modes of documentation - tutorials, how-to guides, technical reference and explanation. It derives its structure from the relationship between them." For more information, see https://diataxis.fr/.

of Panel B could represent different groups of researchers with different experience levels while the y-axis could represent different tasks such as e.g. mean vs. regression. We represent documents with different purposes as differently shaped elements in the grid. Pre-sorting clearly speeds up information discovery compared to the base case (see Panel A), as researchers do not have to browse through all available documents. As such, pre-sorting creates metadata about available information. Pre-sorting is the idea behind the tool "The RDSC Landing Page" that we describe in more detail in the next section.

How useful this pre-sorting is hinges on (i) whether researchers correctly interpret the sorting categories and (ii) the scope and amount of different types of documents in each bucket. Let us begin with the former. This question relates to how researchers assess their own level of experience. For example, does a vastly experienced researcher refer to years working with data or years working with data in the RDC environment? Similarly, the task may not always be clear.

Furthermore, document characteristics may also affect the usefulness of pre-sorting for researchers. For an example, we shall turn to the bucket (A, x) in Panel B which features two distinctly different documents. It may take additional time for a researcher to go through both documents. Along the same lines, very long documents also require researchers to spend additional time finding information appropriate to their task.

To address this challenge, RDCs could either enhance their personal user support, e.g. by assigning staff to assist researchers in finding relevant information, or try to automate this process as far as possible. The advantage of automation is that it is resource-efficient and ensures an equal treatment for all researchers. For example, a tool could present or even apply SDC rules relevant to the current task, e.g. the SDC after the calculation of the mean of a distribution as in the example above.

Panel C of Figure 1 illustrates this point. The hatched area illustrates the tool which relieves the researcher of the decision between the two documents with different purposes that are available for the task and the level of the researcher's experience. Without tools, on the other hand, the researcher would need to go through both documents and then apply the information to the task. Therefore, tools are an avenue to efficient information provision and application. The RDSC Output Submitter is an example of how to implement this idea. This tool automatically applies selected SDC rules to the researcher's output and displays additional information if the output does not comply with applicable SDC rules.

In the next section, we present three examples of how self-developed tools support researchers in performing all tasks related to SDC checking. These examples are taken from the Research Data and Service Centre (RDSC), which is the RDC of the Deutsche Bundesbank[6]. They loosely follow the process of a research project at the RDSC from start to finish. Readers should observe that tasks and therefore information needs of researchers might differ between the various RDCs. This could be due to differences in the underlying legal frameworks, the technical environment in which the data access is granted, or organisational decisions of the RDC.

[6] Readers interested in a more detailed description of the RDSC are cordially referred to "Data Access to Micro Data of the Deutsche Bundesbank" (Schönberg (2019)).

In addition, the information need of a researcher might also vary between different phases of the project. For example, it might be most important for researchers to understand SDC rules at the beginning of a project while the actual application is only relevant when the project is nearing its end. At this final stage, researchers usually submit their results for SDC and need to ensure compliance with the RDC's rules. The tools' focus in the next section will reflect this.

3. Practical examples for self-developed tools supporting researchers at the RDSC of the Deutsche Bundesbank

The project start: Researchers would like to get an overview of all available information – The RDSC Landing Page

The first self-developed tool for automatic filtering of information that we would like to present is the RDSC Landing Page. As the number of data sets available in the RDSC has increased continuously in recent years, the number of rules to apply has also become more and more extensive. In addition, the complexity of the datasets themselves has also increased significantly. Examples of these new challenges include the adequate handling of missing values and duplicates as well as an SDC with multiple variables with entities worth protecting. Finally, the increasing availability of large datasets, and thus the need for advanced programming skills, pose new challenges for the SDC.

As a logical consequence, there are increasingly dataset-specific special rules as well as a variety of auxiliary materials that the RDSC provides to researchers[7]. However, depending on the requested datasets, those challenges might not apply to all research projects and thus not all researchers will need to familiarise themselves with the respective rules. To support researchers visiting the RDSC in finding the right materials at any time and without much effort, the RDSC has built a small application that is available offline in the secure environment. Here, researchers can find a brief summary of relevant information sorted by topic. This allows the RDSC to guide researchers through the information with comparatively little effort and to highlight particularly relevant parts. Detailed support materials, such as researcher guides, are usually only linked.

In our view, the greatest benefit of the RDSC Landing Page is its ability to highlight relations between similar information across different documents and thus point researchers to relevant information for which they were not actually searching. For example, if researchers would like to learn about working with large data, they will most likely go to the respective part in the section on data handling where they will find a helpful guide called "Working with large data at the RDSC" (Gomolka et al, 2021). However, the same report will also be displayed when searching for information on any of the datasets that the RDSC categorises as "large data". Therefore, a researcher who is interested in e.g. a variable overview will likewise find this very helpful report, the existence of which she might otherwise never have known

[7] Examples of such documents are the three technical reports on "Working with large data at the RDSC" (Gomolka et al, 2021), "Statistical Disclosure Control (SDC) for results derived from aggregated confidential microdata" (Blaschke et al, 2022) and "Linking data for MFIs" (Stahl, 2020).

of. This is an example of a relation that is hard to represent otherwise, e.g. with pre-sorting of information.

Technically, the landing page is an NW.js application built from a set of RMarkdown files using the R package {distill} (Allaire, et al, 2018). This makes it easy to improve and extend the landing page, as the RDSC staff need only a working markdown knowledge. Editing or extending the app takes only a few minutes, which allows for quick iteration and development. Note that a mere website would not serve the same purpose, as modern web browsers block any links to local files, which is an essential feature of the landing page because it needs to work within the secure environment where all internet access is blocked.

Figures A.1 and A.2 in the Annex show two examples from the RDSC Landing Page.

During the research: Researchers need context information that applies directly to their analysis – SDC tools

The two software packages "nobsdes5/ nobsreg5" (for STATA users) and {sdcLog} (for R users), both developed by the RDSC, allow researchers to check whether the results they generated in their research project comply with the RDSC's SDC rules. To do this, researchers run a special command directly after generating their results and get immediate feedback from the package.

Thus, researchers receive the appropriate information "just in time". They do not need to go through a full cycle of output submission, checks by RDSC staff and subsequent feedback. Assuming they apply the tools correctly, they know right after the calculation of the result whether or not it complies with the most important SDC rules. Instant feedback allows for quick iteration in case of problems, which is especially helpful for researchers coming a long way to the RDSC premises.

Another strength of both packages is that they provide helpful information for those cases where the results do not comply with SDC rules. Therefore, researchers always know what kind of issue caused the non-compliance and can therefore better react to it. This significantly reduces the workload for both researchers as well as RDSC staff members when performing and checking SDC. However, it is important to understand, that the tools are only semi-automatic and need to be applied appropriately and correctly by researchers, e.g. by modifying the checking commands for their regressions or descriptive tables.

"nobsdes5/ nobsreg5" and "[sdcLog} are both available on the RDSC's website,[8] which means that researchers can also download them before they come to the Deutsche Bundesbank for their on-site work. In addition to the software packages, we also provide practical examples and guidelines to help researchers familiarise themselves with the software. Of course, this documentation is also available in the RDSC's secure environment and can easily be found on the RDSC Landing Page (see Figure A.2).

[8] See https://www.bundesbank.de/en/bundesbank/research/rdsc/your-research-project-at-the-rdsc/output-and-publication-checking-618052. Moreover, the R package "sdcLog" is available on GitHub https://github.com/matthiasgomolka/sdcLog/issues.

Figure A.3 is an excerpt from the {sdcLog} vignette and shows how to use the tool to check whether a mean complies with SDC rules.

Finalising the project: Researchers need to ensure compliance with SDC rules – The Output Submitter

After researchers have concluded their analysis and performed SDC using one of the tools described above to the best of their ability, they need to submit their results to the RDSC for review and approval.

Especially in the case of first-time visiting researchers to the RDSC, many common mistakes and issues occur in this submission: the output is not stored in a permissible format, certain mandatory checks are skipped, or required supplementary information is omitted. The RDSC identifies these issues and discusses them with the researchers, enabling them to improve the quality of their submission to a publishable level. However, this frequently happens after the researchers have already left the RDSC's premises, necessitating another visit, which costs the researchers additional time and effort.

It is therefore beneficial to enable even researchers not entirely familiar with all RDSC rules to identify as many potential issues with their output submission themselves as possible. To achieve this, we have designed and built a self-service application capable of running a battery of automated tests against the intended output submission even before a human RDSC employee looks at the output. Only output which passes these automated checks will then be forwarded to the RDSC staff for final verification.

The properties of a good automated output submission verification process are:

Simplicity: The main beneficiaries of the tool are researchers not yet intimately familiar with the rules and requirements of the RDSC. The usage should therefore be simple enough so that anyone can use it. Time spent learning to use the verification tool could have been spent learning the rules or improving the output and should be minimised. Even experienced researchers can benefit from a simple tool, as it reduces the mental load and allows them to focus on their analysis.

Speed: An automated tool that runs near-instantaneously enables a fast, iterative workflow. Researchers can produce a tentative output, check it for issues themselves, and immediately fix them. Less time and mental energy is spent on non-issues.

Correctness: Errors in output verification (automated or otherwise) fall into two categories: false positives and false negatives. While perfect correctness is desirable and should be strived for, it is not a realistic goal. Therefore, the tool needs to be able to handle such errors gracefully. False positives occur when the checks produce a finding that is not actually a problem. Researchers must be able to mark the false positive as such and explain themselves. The RDSC staff will then choose whether to accept the explanation. False negatives occur when the tool failed to recognise a genuine problem with the output. To mitigate this, it is important to keep both researchers and RDSC staff aware of which kinds of issues can or cannot be automatically detected. This allows human efforts to be focused on the areas too difficult for the automated process to recognise.

Researchers can launch the tool we have created simply by clicking a shortcut in their main project directory. It auto-detects and displays information about the project and the researcher starting the submission. It collects confirmations from the researcher required for regulatory reasons and then proceeds to run the automated checks. These currently include the following:

- Does the project directory structure conform to our expectations?
- Does the analysis code at some point run SDC checks?
- Did the analysis code produce log files?
- Do these log files indicate any issues during the SDC checks?
- Does the code run any commands typically associated with SDC violations because they aggregate away data required for SDC checks? (This would cause only a warning, not outright rejection of the submission.)

The researchers have the opportunity to stop (and later rerun) the tool at any time to go back to their own analysis code and make improvements. They can also make comments regarding any finding of the tool. Once they are content with their output submission, the tool compiles all the information it gathered into a small report for the RDSC staff and allows the researchers to submit it by email.

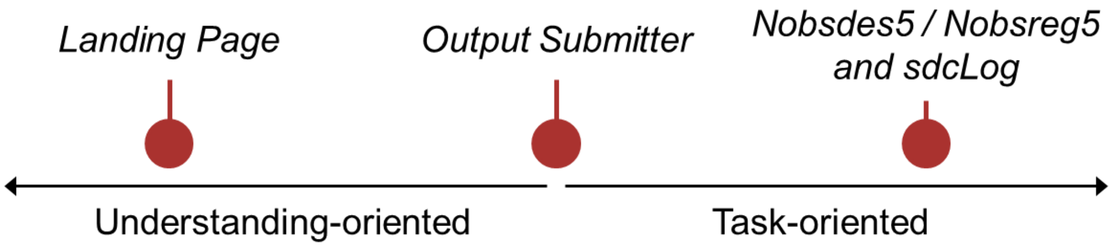
We developed the RDSC Output Submitter ourselves in-house as a native desktop application using web-technologies (javascript, html, css, node). This helps us to modify our tools easily and quickly based on researcher feedback and experience regarding errors in the checks.

Figures A.4 and A.5 in the Annex show two examples from the RDSC's Output Submitter.

Aligning the three tools by information purpose

Information purpose spans from understanding-oriented (e.g. rules in a legal text) to task-oriented (e.g. applying rules to specific results).

Figure 2



Sources: Authors

4. Discussion and future areas of improvement

Due to its importance and comparatively high resource and time costs, SDC is one of the most debated topics in the field of RDCs. This paper presents three self-developed tools intended to support researchers in performing SDC in an RDC secure environment. All tools share the same goals of supporting researchers in understanding and applying SDC rules. However, the tools follow different approaches in pursuit of these objectives. At one end of the spectrum is the RDSC Landing Page that helps researchers faster understand the rules. The "nobsdes5 / nobsreg5" and {sdcLog} tools support researchers in applying SDC rules and accordingly are to be found at the other end of the spectrum, with the Output Submitter located somewhere in the middle. Ultimately, the different approaches reflect the changing focus of researchers during the project's lifecycle.

All tools in this paper present an improvement over a situation in which only unstructured information is available to researchers. One obvious area of improvement would be to move from semi-automated to fully automated tools. While this would clearly benefit researchers as they would receive more help in performing SDC it also complicates the development of these tools, putting some constraints on RDCs to do this.

Ultimately, the different approaches reflect the changing researchers' needs during the project's lifecycle. At the beginning of the project researchers focus most on understanding the rules they need to adhere to. While working on their analysis a researcher's focus shifts from understanding towards applying SDC rules to check compliance of the generated results. After the analysis is complete, the focus shifts back towards understanding as the focus now is on submitting results for output checking.

References

Allaire, J.J., R. Iannone, A. P. Hill and Y. Xie (2018). Distill for R Markdown. Retrieved from https://rstudio.github.io/distill

Blaschke, J., M. Gomolka and C. Hirsch (2022). Statistical Disclosure Control (SDC) for results derived from aggregated confidential microdata, Technical Report 2022-01 – Version 1.0. Deutsche Bundesbank, Research Data and Service Centre.

Gomolka, M., Blaschke, J., and Hirsch, C. (2021). Working with large data at the RDSC, Technical Report 2021-04 – Version 1.1. Deutsche Bundesbank, Research Data and Service Centre.

Procida, D. (2022) "Diátaxis Documentation Framework". Retrieved from https://diataxis.fr/ on 13 July 2022

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.

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

Ritchie, F. (2017). The "Five Safes": A framework for planning, designing and evaluating data access solutions. In Data for Policy 2017: Government by Algorithm? (Data for Policy). Zenodo.

Schönberg, T. (2019). Data Access to Micro Data of the Deutsche Bundesbank. Technical Report 2019-02, Deutsche Bundesbank, Research Data and Service Centre.

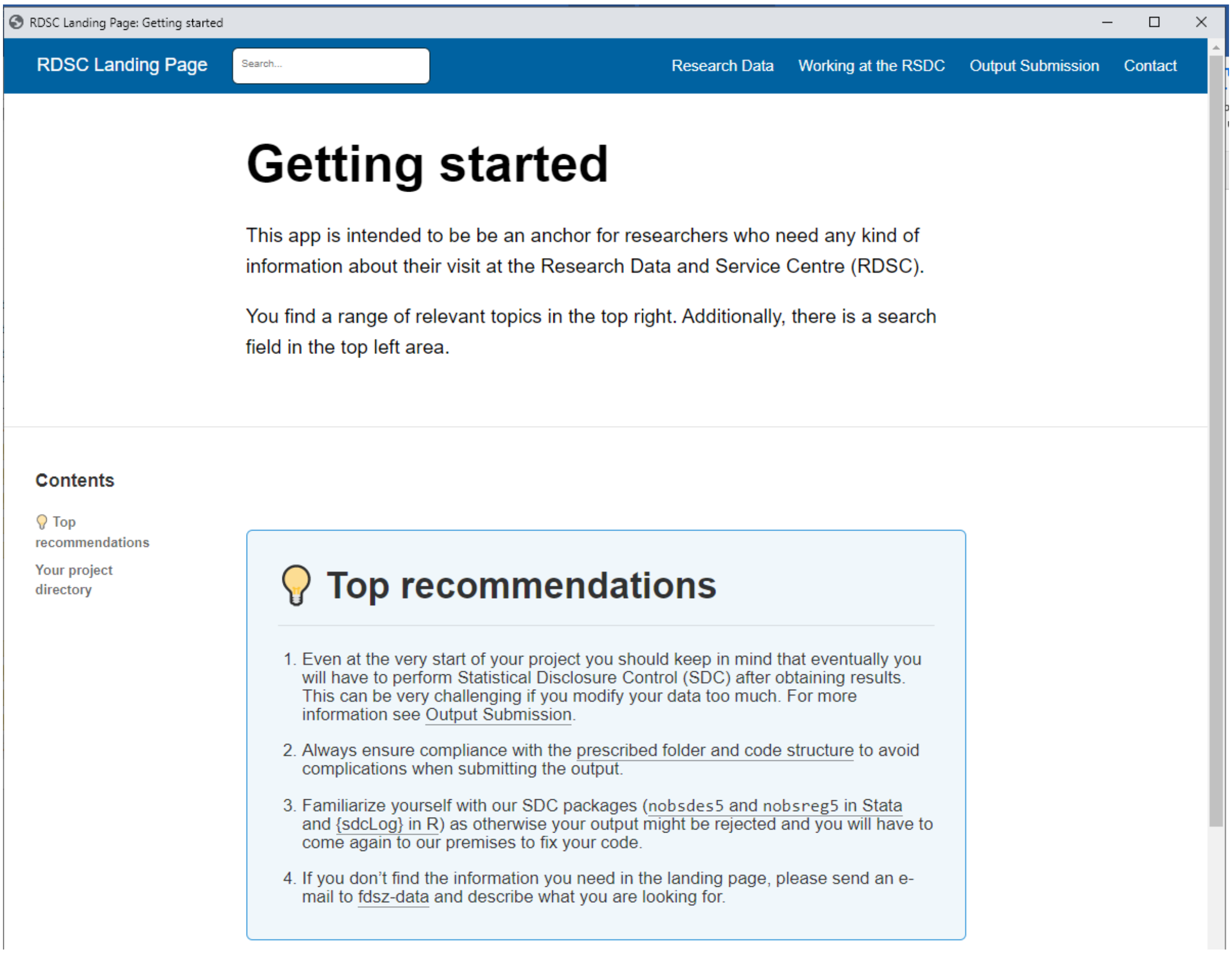
Stahl, H. (2020). Linking MFI data, Technical Report 2020-04 – Version 2. Deutsche Bundesbank, Research Data and Service Centre.

Annex

RDSC Landing Page – Getting started

Figure A.1

This is what researchers see when opening the tool. After reading a brief summary of the most important recommendations, they can easily navigate to other topics.



Getting started

This app is intended to be be an anchor for researchers who need any kind of information about their visit at the Research Data and Service Centre (RDSC).

You find a range of relevant topics in the top right. Additionally, there is a search field in the top left area.

💡 Top recommendations

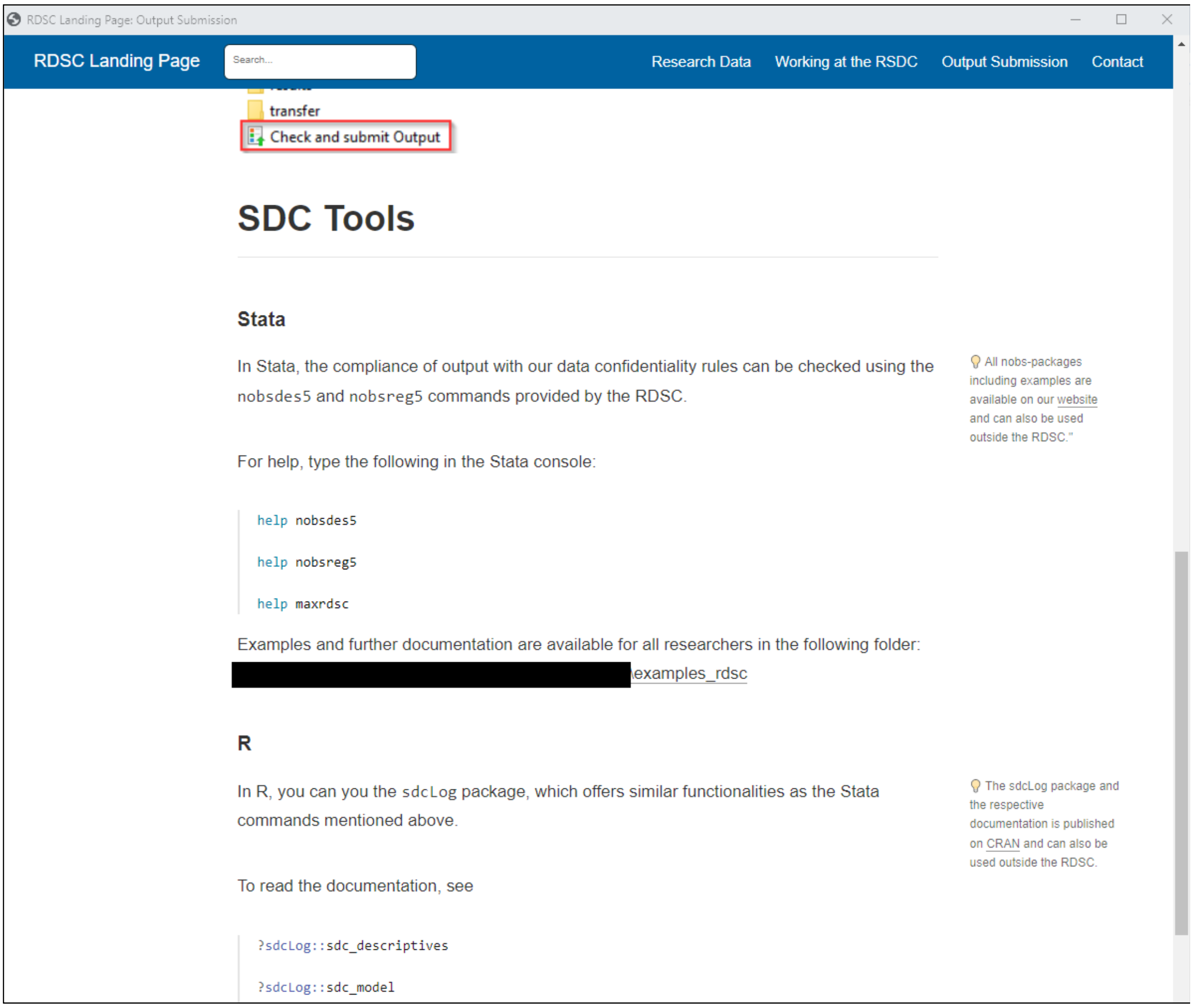
1. Even at the very start of your project you should keep in mind that eventually you will have to perform Statistical Disclosure Control (SDC) after obtaining results. This can be very challenging if you modify your data too much. For more information see Output Submission.
2. Always ensure compliance with the prescribed folder and code structure to avoid complications when submitting the output.
3. Familiarize yourself with our SDC packages (`nobsdes5` and `nobsreg5` in Stata and {sdcLog} in R) as otherwise your output might be rejected and you will have to come again to our premises to fix your code.
4. If you don't find the information you need in the landing page, please send an e-mail to fdsz-data and describe what you are looking for.

Sources: Authors

RDSC Landing Page – Output Submission

Figure A.2

The tool also provides detailed information on SDC rules and available tools for different programming languages as well as recommendations on how to use them.



SDC Tools

Stata

In Stata, the compliance of output with our data confidentiality rules can be checked using the `nobsdes5` and `nobsreg5` commands provided by the RDSC.

All nobs-packages including examples are available on our website and can also be used outside the RDSC."

For help, type the following in the Stata console:

```
help nobsdes5

help nobsreg5

help maxrdsc
```

Examples and further documentation are available for all researchers in the following folder: [redacted]\examples_rdsc

R

In R, you can you the `sdcLog` package, which offers similar functionalities as the Stata commands mentioned above.

The sdcLog package and the respective documentation is published on CRAN and can also be used outside the RDSC.

To read the documentation, see

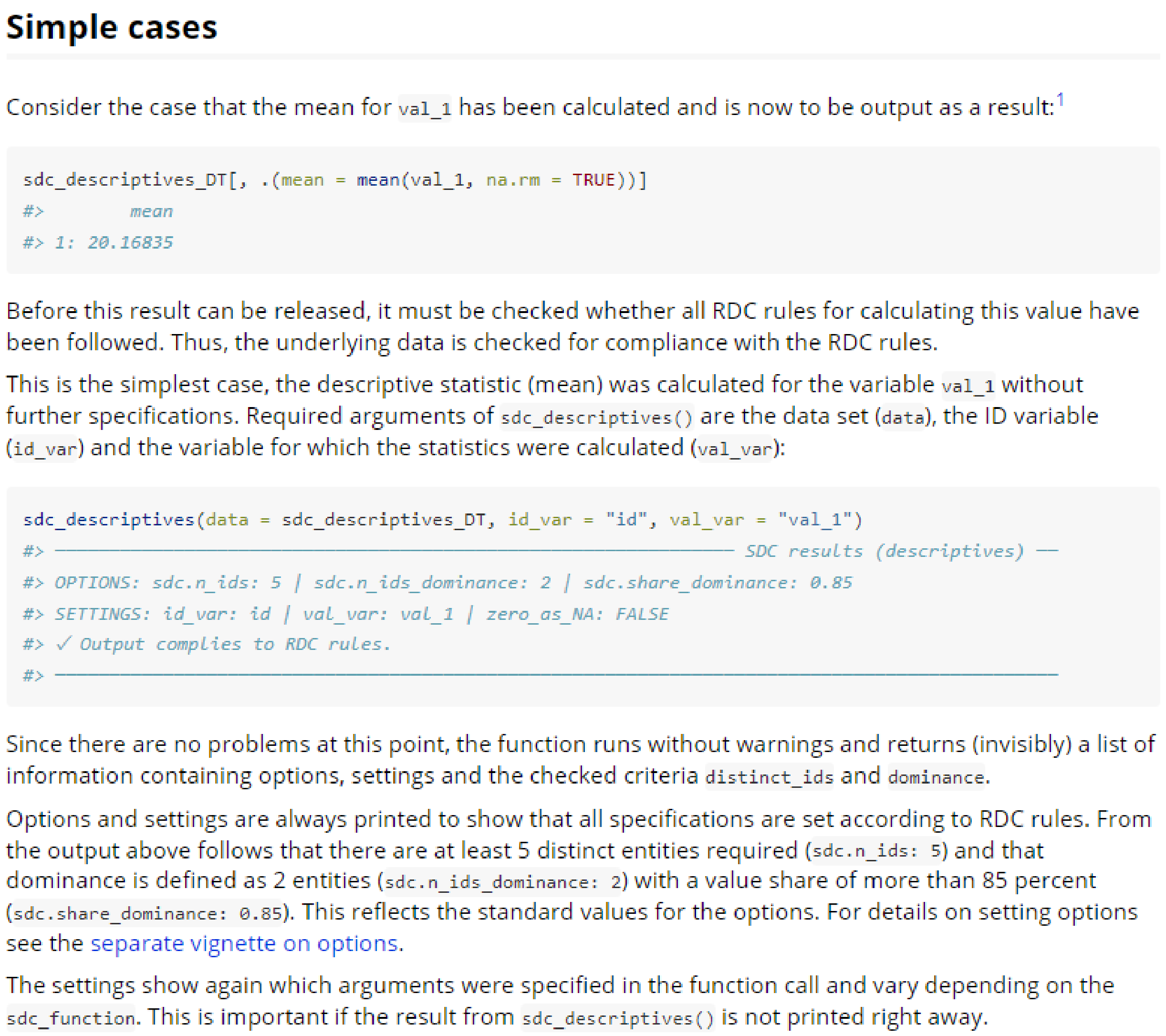
```
?sdcLog::sdc_descriptives

?sdcLog::sdc_model
```

Sources: Authors

{sdcLog} – Excerpt from the documentation

Figure A.3

The following example for {sdcLog} is taken from the vignette and shows a simple example for the application of {sdcLog} to check whether a mean complies with SDC rules.



Simple cases

Consider the case that the mean for `val_1` has been calculated and is now to be output as a result:[1]

```
sdc_descriptives_DT[, .(mean = mean(val_1, na.rm = TRUE))]
#>        mean
#> 1: 20.16835
```

Before this result can be released, it must be checked whether all RDC rules for calculating this value have been followed. Thus, the underlying data is checked for compliance with the RDC rules.

This is the simplest case, the descriptive statistic (mean) was calculated for the variable `val_1` without further specifications. Required arguments of `sdc_descriptives()` are the data set (`data`), the ID variable (`id_var`) and the variable for which the statistics were calculated (`val_var`):

```
sdc_descriptives(data = sdc_descriptives_DT, id_var = "id", val_var = "val_1")
#> ──────────────────────────────────────────────────────── SDC results (descriptives) ──
#> OPTIONS: sdc.n_ids: 5 | sdc.n_ids_dominance: 2 | sdc.share_dominance: 0.85
#> SETTINGS: id_var: id | val_var: val_1 | zero_as_NA: FALSE
#> ✓ Output complies to RDC rules.
#> ─────────────────────────────────────────────────────────────────────────────────────
```

Since there are no problems at this point, the function runs without warnings and returns (invisibly) a list of information containing options, settings and the checked criteria `distinct_ids` and `dominance`.

Options and settings are always printed to show that all specifications are set according to RDC rules. From the output above follows that there are at least 5 distinct entities required (`sdc.n_ids: 5`) and that dominance is defined as 2 entities (`sdc.n_ids_dominance: 2`) with a value share of more than 85 percent (`sdc.share_dominance: 0.85`). This reflects the standard values for the options. For details on setting options see the separate vignette on options.

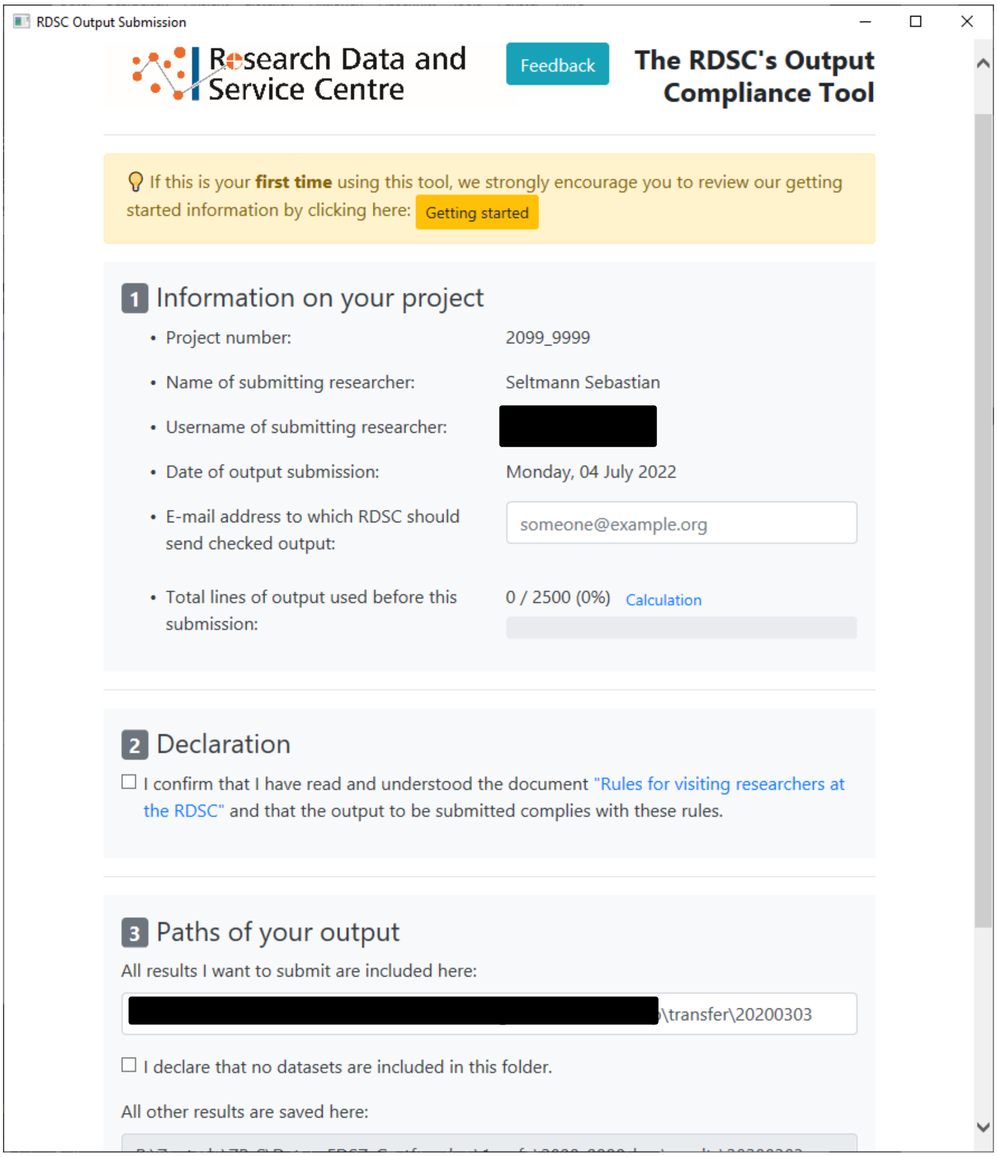
The settings show again which arguments were specified in the function call and vary depending on the `sdc_function`. This is important if the result from `sdc_descriptives()` is not printed right away.

Sources: https://cran.r-project.org/web/packages/sdcLog/vignettes/intro.html

RDSC Output Submitter – Adding check properties

Figure A.4

Screenshot right after launching. Here researchers can read a brief introduction on how to use the tool and indicate which results they wish to submit.



The RDSC's Output Compliance Tool

If this is your **first time** using this tool, we strongly encourage you to review our getting started information by clicking here: Getting started

1 Information on your project

- Project number: 2099_9999
- Name of submitting researcher: Seltmann Sebastian
- Username of submitting researcher: [redacted]
- Date of output submission: Monday, 04 July 2022
- E-mail address to which RDSC should send checked output: someone@example.org
- Total lines of output used before this submission: 0 / 2500 (0%) Calculation

2 Declaration

□ I confirm that I have read and understood the document "Rules for visiting researchers at the RDSC" and that the output to be submitted complies with these rules.

3 Paths of your output

All results I want to submit are included here:

[redacted]\transfer\20200303

□ I declare that no datasets are included in this folder.

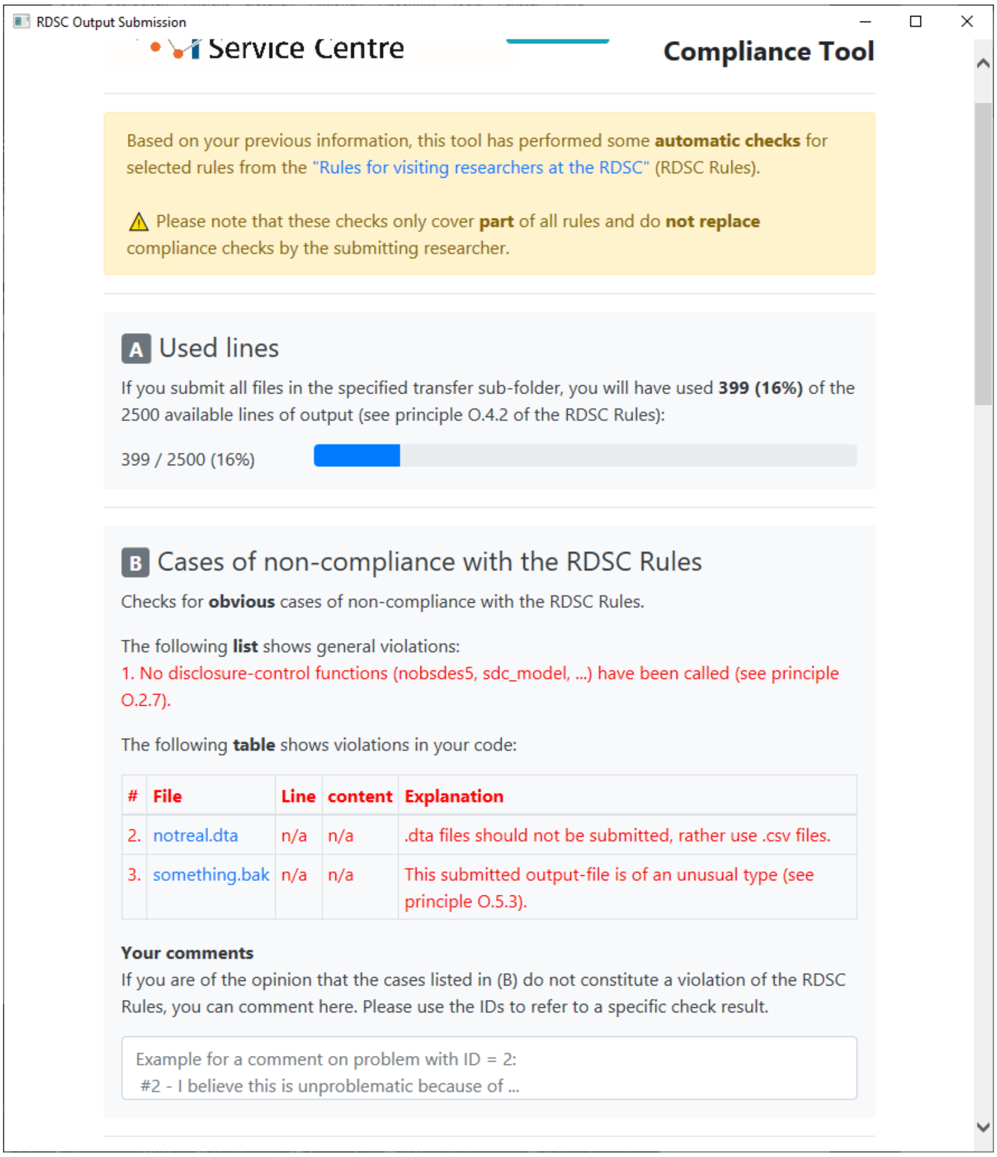
All other results are saved here:

Sources: Authors

RDSC Output Submitter – Automated checks

Figure A.5

Screenshot after running automated checks, before submission to the RDSC staff, giving the opportunity for correction or comment.



Based on your previous information, this tool has performed some **automatic checks** for selected rules from the "Rules for visiting researchers at the RDSC" (RDSC Rules).

⚠ Please note that these checks only cover **part** of all rules and do **not replace** compliance checks by the submitting researcher.

A Used lines

If you submit all files in the specified transfer sub-folder, you will have used **399 (16%)** of the 2500 available lines of output (see principle O.4.2 of the RDSC Rules):

399 / 2500 (16%)

B Cases of non-compliance with the RDSC Rules

Checks for **obvious** cases of non-compliance with the RDSC Rules.

The following **list** shows general violations:

1. No disclosure-control functions (nobsdes5, sdc_model, ...) have been called (see principle O.2.7).

The following **table** shows violations in your code:

#	File	Line	content	Explanation
2.	notreal.dta	n/a	n/a	.dta files should not be submitted, rather use .csv files.
3.	something.bak	n/a	n/a	This submitted output-file is of an unusual type (see principle O.5.3).

Your comments

If you are of the opinion that the cases listed in (B) do not constitute a violation of the RDSC Rules, you can comment here. Please use the IDs to refer to a specific check result.

Example for a comment on problem with ID = 2:
#2 - I believe this is unproblematic because of ...

Sources: Authors

Introduction to and application of SDC rules using self-developed tools

J. Blaschke, M. Gomolka, C. Hirsch, S. Seltmann and H. Stahl (Deutsche Bundesbank)

11th Biennial IFC Conference on "Post-pandemic landscape for central bank statistics"

Session 2 – Microdata disclosure control: a practical perspective

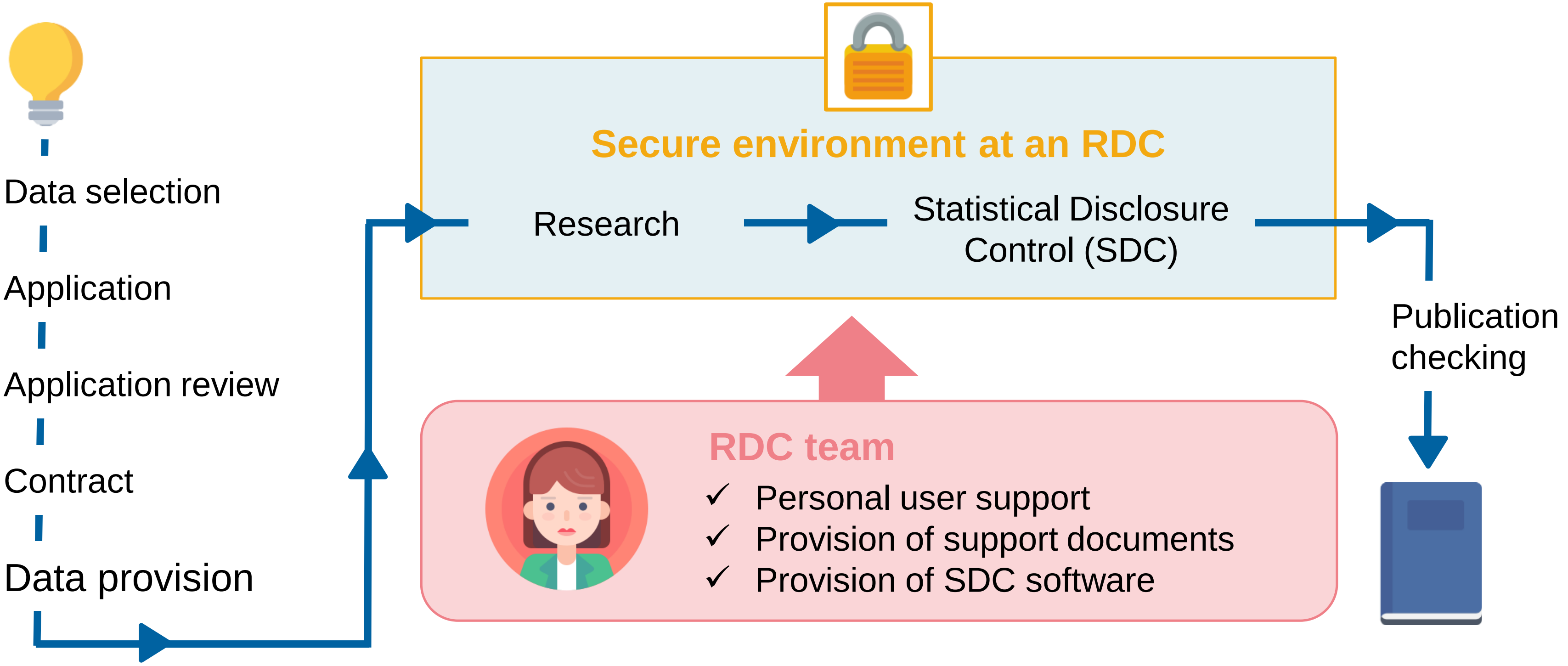
25 August 2022

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

Introduction to and application of SDC rules using self-developed tools

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

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



Data selection

Application

Application review

Contract

Data provision

Secure environment at an RDC

Research → Statistical Disclosure Control (SDC)

Publication checking

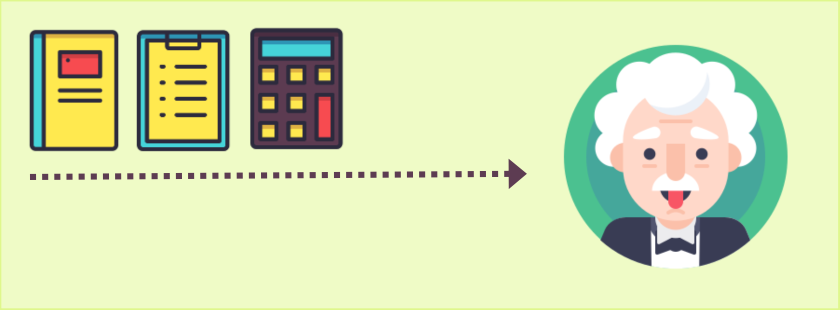
RDC team

- ✓ Personal user support
- ✓ Provision of support documents
- ✓ Provision of SDC software

Introduction to and application of SDC rules using self-developed tools

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

Provision of support materials (e.g. documents, software) can potentially lead to an information overflow



Scenario 1 - Little user support needed

- Small number of datasets
- Easy data structure
- Small dataset size
- Homogeneous legal framework → Similar rules for data access and SDC

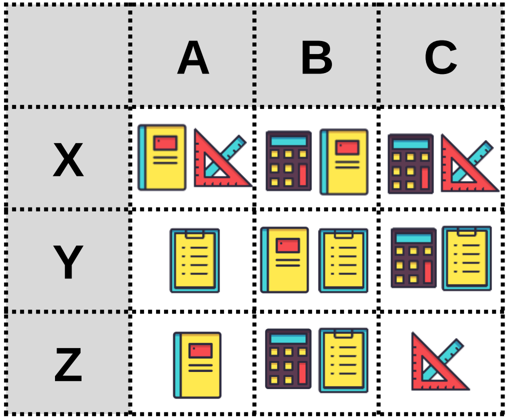
Scenario 2 - Much user support needed

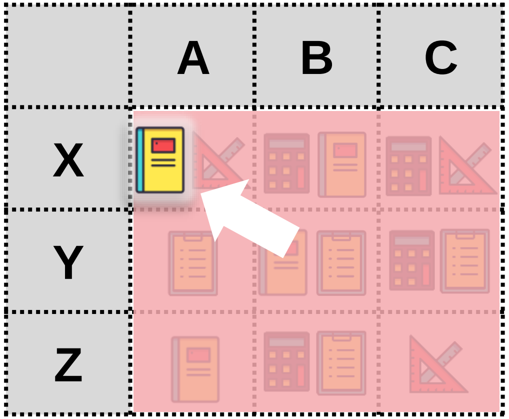
- Large number of datasets
- Complex data structure (e.g. multiple IDs, missing IDs)
- Large dataset size
- Heterogeneous legal framework → Dataset-specific rules for data access and SDC

Introduction to and application of SDC rules using self-developed tools

2. How tools support researchers in understanding and applying SDC rules

***Possible solution for scenario 2**: Pre-sorting information and automating as far as possible*





Case A

- ☐ No structure of material
- ☐ No Customized filtering

Case B

- ✓ Structured information (e.g. by researcher's characteristics and purpose of the document)
- ☐ No Customized filtering

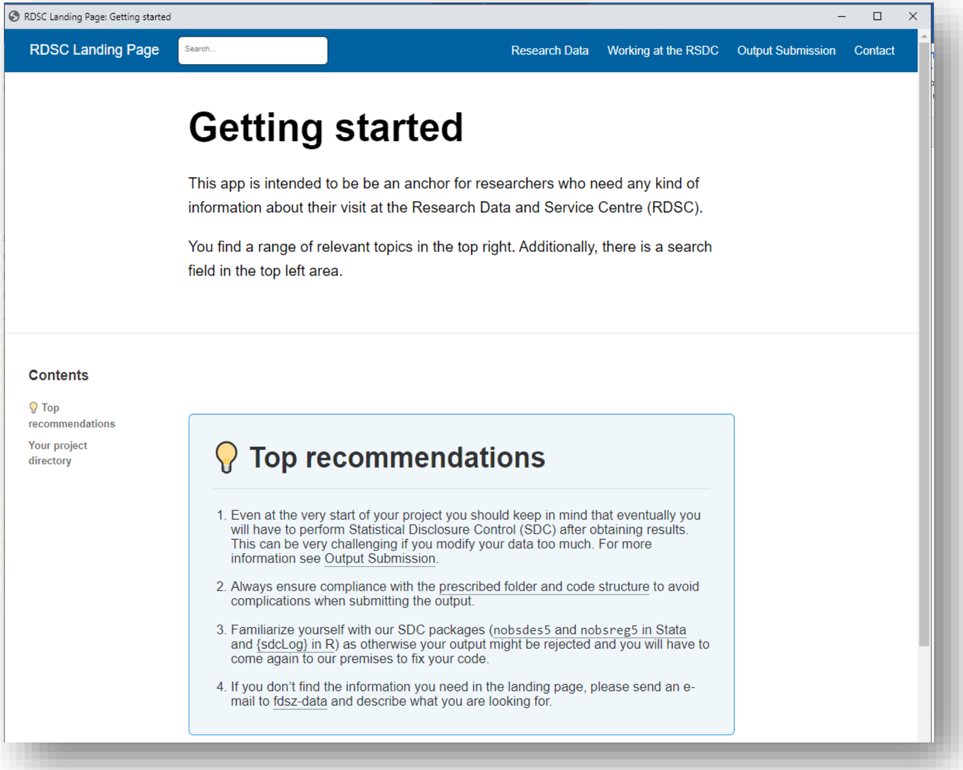
Case C

- ✓ Structured information (e.g. by researcher's characteristics and purpose of the document)
- ✓ Customized filtering

Introduction to and application of SDC rules using self-developed tools

3. Practical examples for self-developed BBk tools - RDSC Landing Page

***The Project start**: Researchers would like to get an overview of all available information*



Getting started

This app is intended to be be an anchor for researchers who need any kind of information about their visit at the Research Data and Service Centre (RDSC).

You find a range of relevant topics in the top right. Additionally, there is a search field in the top left area.

Contents

Top recommendations

Your project directory

Top recommendations

1. Even at the very start of your project you should keep in mind that eventually you will have to perform Statistical Disclosure Control (SDC) after obtaining results. This can be very challenging if you modify your data too much. For more information see Output Submission.
2. Always ensure compliance with the prescribed folder and code structure to avoid complications when submitting the output.
3. Familiarize yourself with our SDC packages (`nobsdes5` and `nobsreg5` in Stata and {sdcLog} in R) as otherwise your output might be rejected and you will have to come again to our premises to fix your code.
4. If you don't find the information you need in the landing page, please send an e-mail to fdsz-data and describe what you are looking for.

Features

- Structured **overview** of all available resources (e.g. documentation, software)
- Show **relations** between similar information across documents → Point researchers to relevant information they were not actually searching

➡ *Structured information*

Technical set-up

- NW.js application, which is built using Rmarkdown and the R package {distill}
- Advantages: Easy to modify, possibility to open links to local files

Introduction to and application of SDC rules using self-developed tools
3. Practical examples for self-developed BBk tools - SDC tools

During the research: *Researchers need information that applies directly to their analysis*

nobsdes5

nobsreg5

{sdcLog}

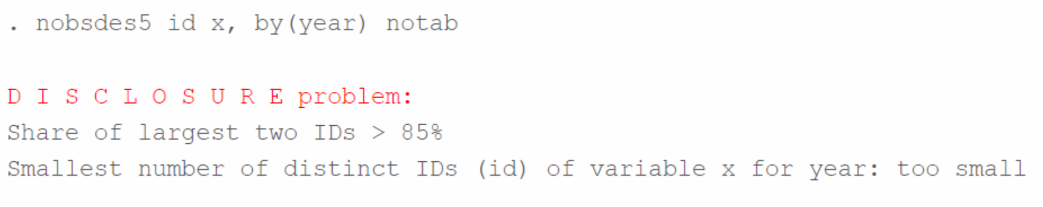
- Both packages are freely available on the RDSC's website.
- {sdcLog} is also available on CRAN https://cran.r-project.org/package=sdcLog

Features

Researchers can use commands after generating a result (e.g. descriptive or regression table) and get immediate feedback, if the table will pass SDC or not

- ➡ No need to program checks themselves
- ➡ Prerequisite: Researchers need to correctly apply the commands

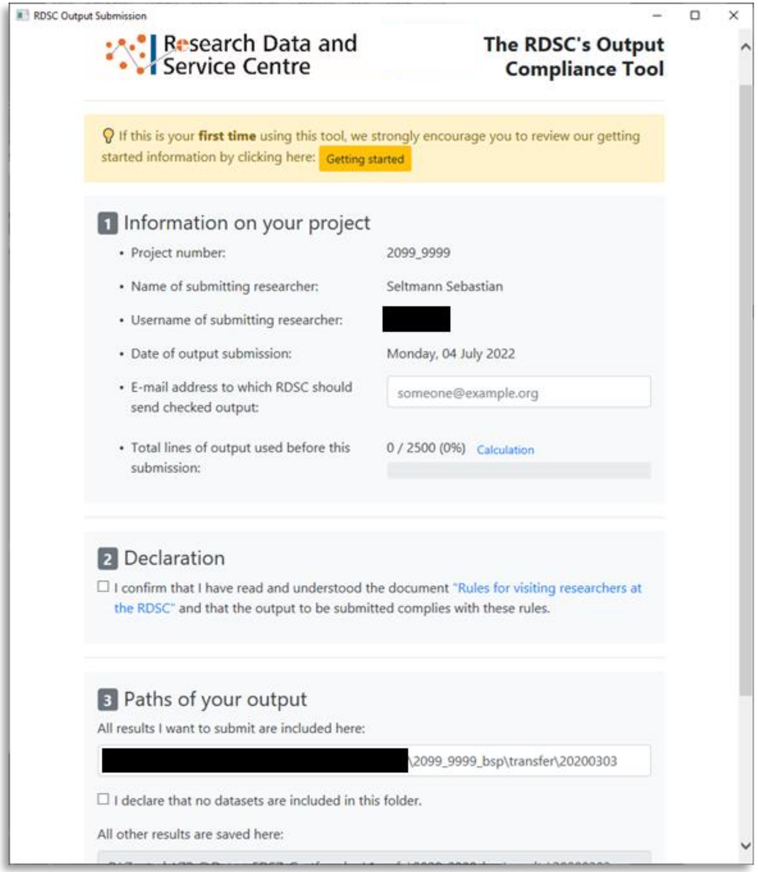
Example from nobsdes5:



```
. nobsdes5 id x, by(year) notab

DISCLOSURE problem:
Share of largest two IDs > 85%
Smallest number of distinct IDs (id) of variable x for year: too small
```

Introduction to and application of SDC rules using self-developed tools

3. Practical examples for self-developed BBk tools - RDSC Output Submitter (1|2)



The RDSC's Output Compliance Tool

If this is your **first time** using this tool, we strongly encourage you to review our getting started information by clicking here: Getting started

1 Information on your project

• Project number:	2099_9999
• Name of submitting researcher:	Seltmann Sebastian
• Username of submitting researcher:	[redacted]
• Date of output submission:	Monday, 04 July 2022
• E-mail address to which RDSC should send checked output:	someone@example.org
• Total lines of output used before this submission:	0 / 2500 (0%) Calculation

2 Declaration

□ I confirm that I have read and understood the document "Rules for visiting researchers at the RDSC" and that the output to be submitted complies with these rules.

3 Paths of your output

All results I want to submit are included here:

[redacted]\2099_9999_bsp\transfer\20200303

□ I declare that no datasets are included in this folder.

All other results are saved here:

***Finalizing the project:** Researchers need to ensure compliance with SDC rules*

Features

- Automated checks for selected rules (e.g. file format, folder structure)
- Warnings for potential rule breaches
- Automated count of remaining lines of output

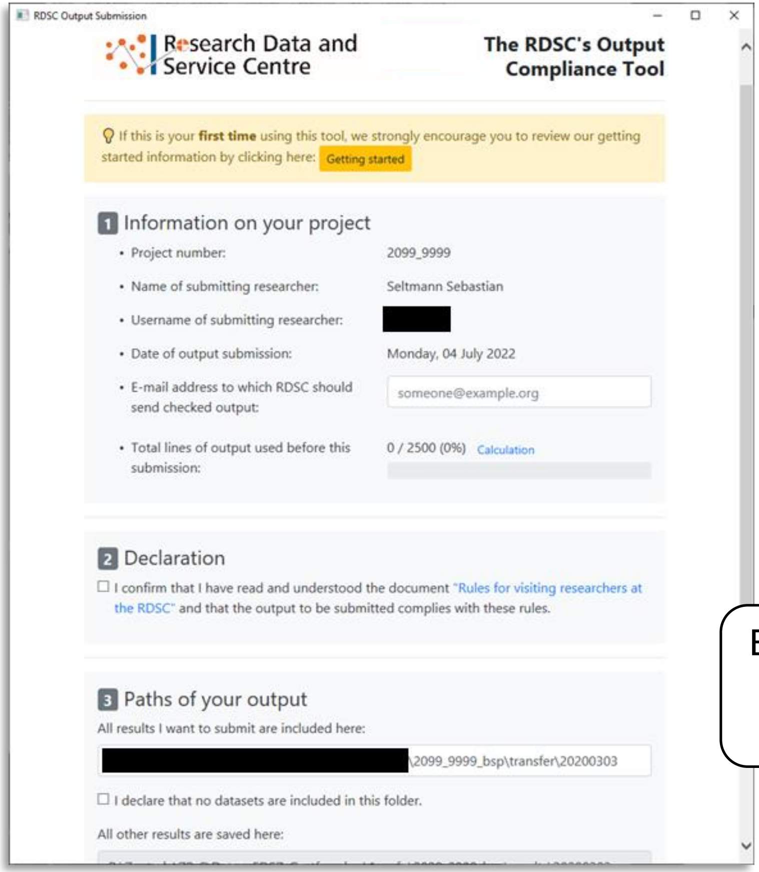
➡ *Customized filtering of information*

Technical set-up

- Web-technologies (javascript, html, css, node)
- Runs locally, allowing access to research project code, logs and output files

Introduction to and application of SDC rules using self-developed tools

3. Practical examples for self-developed BBk tools - RDSC Output Submitter (2|2)



The RDSC's Output Compliance Tool

If this is your **first time** using this tool, we strongly encourage you to review our getting started information by clicking here: Getting started

1 Information on your project

- Project number: 2099_9999
- Name of submitting researcher: Seltmann Sebastian
- Username of submitting researcher: [redacted]
- Date of output submission: Monday, 04 July 2022
- E-mail address to which RDSC should send checked output: someone@example.org
- Total lines of output used before this submission: 0 / 2500 (0%) Calculation

2 Declaration

□ I confirm that I have read and understood the document "Rules for visiting researchers at the RDSC" and that the output to be submitted complies with these rules.

3 Paths of your output

All results I want to submit are included here:

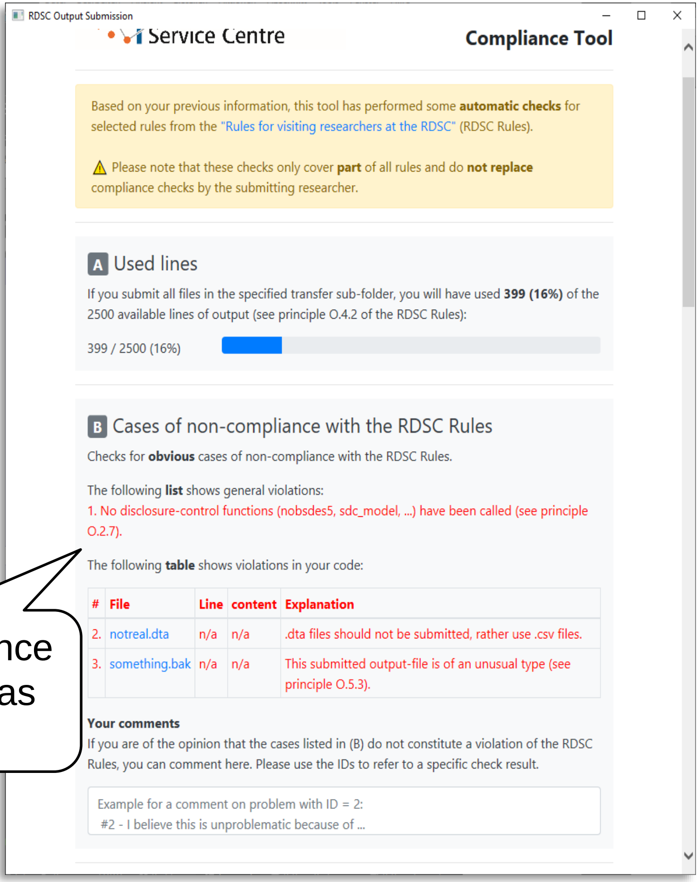
[redacted]\2099_9999_bsp\transfer\20200303

□ I declare that no datasets are included in this folder.

All other results are saved here:

Example

Explicit reference to rule that was breached.



Compliance Tool

Based on your previous information, this tool has performed some **automatic checks** for selected rules from the "Rules for visiting researchers at the RDSC" (RDSC Rules).

⚠ Please note that these checks only cover **part** of all rules and do **not replace** compliance checks by the submitting researcher.

A Used lines

If you submit all files in the specified transfer sub-folder, you will have used **399 (16%)** of the 2500 available lines of output (see principle O.4.2 of the RDSC Rules):

399 / 2500 (16%)

B Cases of non-compliance with the RDSC Rules

Checks for **obvious** cases of non-compliance with the RDSC Rules.

The following **list** shows general violations:

1. No disclosure-control functions (nobsdes5, sdc_model, ...) have been called (see principle O.2.7).

The following **table** shows violations in your code:

#	File	Line	content	Explanation
2.	notreal.dta	n/a	n/a	.dta files should not be submitted, rather use .csv files.
3.	something.bak	n/a	n/a	This submitted output-file is of an unusual type (see principle O.5.3).

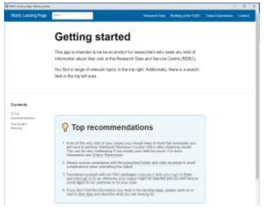
Your comments

If you are of the opinion that the cases listed in (B) do not constitute a violation of the RDSC Rules, you can comment here. Please use the IDs to refer to a specific check result.

Example for a comment on problem with ID = 2:
#2 - I believe this is unproblematic because of ...

Introduction to and application of SDC rules using self-developed tools

4. Conclusion and discussion



RDSC Landing Page

Research Data · Working at the RDSC · Output Submission · Contact

Getting started

This app is intended to be be an anchor for researchers who need any kind of information about their visit at the Research Data and Service Centre (RDSC).

You find a range of relevant topics in the top right. Additionally, there is a search field in the top left area.

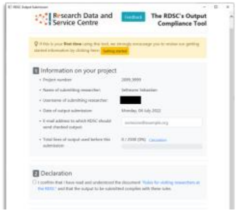
Contents

Top recommendations

Your project directory

Top recommendations

1. Even at the very start of your project you should keep in mind that eventually you will have to perform Statistical Disclosure Control (SDC) after obtaining results. This can be very challenging if you modify your data too much. For more information see Output Submission.
2. Always ensure compliance with the prescribed folder and code structure to avoid complications when submitting the output.
3. Familiarize yourself with our SDC packages (nobsdes5 and nobsreg5 in Stata and sdcLog in R) as otherwise your output might be rejected and you will have to come again to our premises to fix your code.
4. If you don't find the information you need in the landing page, please send an e-mail to fdsz-data and describe what you are looking for.



RDSC Output Submission

Research Data and Service Centre

Feedback

The RDSC's Output Compliance Tool

If this is your **first time** using this tool, we strongly encourage you to review our getting started information by clicking here: Getting started

1 Information on your project

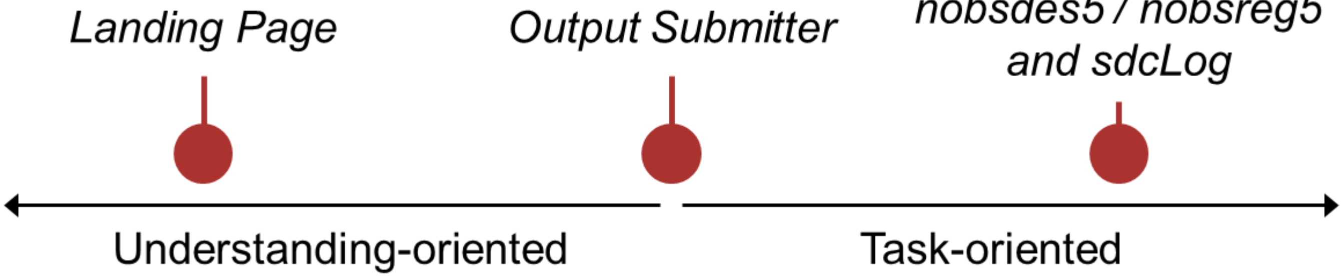
- Project number: 3999_9999
- Name of submitting researcher: [illegible]
- Username of submitting researcher: [redacted]
- Date of output submission: Monday, 04 July 2022
- E-mail address to which RDSC should send checked output: someone@example.org
- Total lines of output used before this submission: 0 / 2500 (0%) Calculation

2 Declaration

☐ I confirm that I have read and understood the document "Rules for visiting researchers at the RDSC" and that the output to be submitted complies with these rules.

STATA

R



Landing Page

Output Submitter

nobsdes5 / nobsreg5 and sdcLog

Understanding-oriented

Task-oriented

Future areas of improvement

Move from semi-automated to **fully automated tools.** While this would clearly benefit researchers as they get more help in performing SDC it also complicates the development of these tools putting some constraints on RDCs to do this.

Thank you

Jannick.blaschke@bundesbank.de

Website: www.bundesbank.de/rdsc

Contact: fdsz@bundesbank.de