October/30/2015

Guidelines for Access to and the use of Data at CIRRAU servers

Malene Thygesen[12], Esben Agerbo[12], Hanne Birgitte Hede Jørgensen[12], Carsten Bøcker Pedersen[12]

[1] Centre for Integrated Register-based Research, CIRRAU, Aarhus University, Aarhus, Denmark

[2] National Centre for Register-Based Research, Aarhus University, Business and Social Sciences, Aarhus, Fuglesangs Alle 4, 8210 Aarhus V, Denmark

CONTENT

Guidelines for research utilizing CIRRAU's infrastructure

This document describes how to access and use CIRRAU data and provides guidelines to best practice for responsible use of our common computer resources. Most importantly, researchers must be aware of and limit their use of disk space. Please read and comply with these guidelines for the benefit of all users.

When conducting research on personally sensitive data ("personfølsomme oplysninger" in Danish), you must ensure that personally sensitive information is not transferred to any unauthorized persons. Researchers may access personally sensitive information for the approved research, only, and must never reveal any personally sensitive information to anyone outside the project. This is by far the most important criterion for any register-based study of individuals.

The solution offered by CIRRAU consists of VPN access to a server located at Statistics Denmark which contains the personally sensitive data. The server is accessed through the researcher's own pc. Downloading personally sensitive data to the researcher's own pc is neither possible, nor permitted. However, through e-mail it is possible to access small files with aggregated data from the server, provided that these files contain no personally sensitive data. The advantage to this solution is that researchers are protected from accidentally violating the Danish Act on Processing of Personal Data. However, before transferring output from the Statistics Denmark environment, make sure that you understand and comply with the rules regarding transfer of files (For more details – see section 3).

A challenge when applying such a set-up is that all researchers share computer resources, including CPU, virtual memory and most importantly disk storage. Given the large size of many population-based registers, a single researcher with an (possibly inadvertently) unfortunate practice can impact negatively on all other researchers' possibility to perform research.

To ensure the best performance of our common computer resources users should:

1) keep only one working dataset in addition to the rawdata and not make copies for each minor change in the dataset. Instead, the user should keep program code that updates the one working dataset with changes. This will also facilitate documentation and overview in the long run (For more details- see section 1).

2) keep track of your own disk use. A user is not allowed to generate more than 20GB per project per year. Disk use can easily be checked by placing the mouse on your personal folder in either "WORKDATA" or "KODE", make a right click, scroll down and click on "properties" and look at "size on disk". If the size of the data exceeds 20GB and has been generated in less than a year, then you are using too much disk space and you need to delete some of your data.

3) limit the use of virtual memory and CPU. For example, users are encouraged NOT to run several processes at the same time and if possible "heavy jobs" should be run at night or during the weekends.

4) remember to "log off" properly through "START", otherwise all temporary files will not be deleted, and this may cause disk space problems.

5) if you delete permanent files also remember to empty the “Recycle Bin” which is found on the desktop in the upper left corner. Simply right click on the recycle bin and choose empty recycle bin.

6) do not save files at disk C – files should only be saved in the WORKDATA area, in your personal folder “KODE” or at your home directory disk Z. The latter has an upper limit of 50MB.

To ensure the performance of our common computer resources, researchers who repeatedly fail to comply with the guidelines described in this document will be locked out and their data will be removed from the server to ensure good working conditions for everyone else ☺

1. Best practice in data management

1: Never keep more than one new dataset

Keeping track of the research progress is the most important key to avoid unnecessary permanent datasets, and thereby avoiding use of unnecessary disk space. Typically, register-based research utilizes large datasets, and several copies of these datasets will take up disk space, impacting negatively on all users.

Many new users make a new permanent copy of a dataset for each minor change. The worst scenario is that you get access to data1, and based on this you make a new variable called var2 and save this dataset as data2. Then you decide to make another variable called var3 based on the modified dataset data2, and save the new data as data3. If you end up making 10 new variables, you will have 10 nearly identical permanent copies of the same dataset. After a few years you may request an update to data1; now you will have to repeat all these 10 steps to implement the update, while also saving additional 10 new permanent copies of the same datasets. This is the worst case scenario for data handling regarding permanent copies of datasets. Please feel free to make as many temporary datasets as needed as long as they are stored only temporarily on the server. ‘Temporarily’ means that it is only stored during the current logon session.

2: Keep track of your research progress

Never make changes to the rawdata. Please note that if data are delivered through Statistics Denmark and placed in the RAWDATA folder, it is not possible to make changes to your rawdata.

Use version control. When you prepare a dataset to perform analyses, such as the relation between income and survival:

1) save a copy of the SAS/STATA/R program/syntax performing these tasks, and name this program inc_surv01.sas (or .r). All file references in this program should be directly back to rawdata, thereby documenting all changes needed to prepare data for analyses. The perfect program/syntax includes a description of the reason for each of the tasks performed by the program/syntax.

2) save a copy of the dataset when it is completely ready to be analyzed, and name this copy inc_surv01.sas7bdat. Using the same first name you will indirectly keep track of which program/syntax generated your dataset for analyses.

3) When changes are needed to make new analyses, generate a new program/syntax for this task based on the former version 01, and name the new version 02. As before, in the new version of the

program/syntax (inc_surv02.sas) all file references should be directly back to rawdata, and never to the former version of these data (e.g. inc_surv01.sas7bdat).

4) If you follow these guidelines, you will have a clear track of the research progress, and it will be possible at any stage to easily implement a new update to your data. In addition, it is possible to delete all intermediate datasets, and also regenerate data again if needed.

5) Refrain from using version control such as final, very final, second final, definite final, final2 or the like. There will never be a final version (except when the paper has been published and you no longer work on the project).

In addition, you are advised to create a text file called 'readme.txt' in each directory where you shortly describe the contents of files in this directory as well as planned future steps, planned changes, along with dates. This will also help you keep track of your research process.

Unfortunately, in our experience, many researchers fail to keep track of the research process, making it difficult to later implement changes and update data in the project. Typically, when implementing new changes these are very obvious and there is no need to document them. However, when the paper comes back from review 9-12 months later many researchers have forgotten the underlying reason for the implemented changes.

2. Introduction to the CIRRAU servers

CIRRAU has 4 hosted servers at Statistics Denmark called Srvfsencrr3, Srvfsencrr4, Srvfsencrr5 and Srvfsencrr6. When researchers have their project request accepted at Statistics Denmark they are also told on which server the project will be located. The configuration varies slightly between the 4 servers.
On the desktop of each server you find a document called "Intro_srvfsencrrX.docx" with a description of the system, applications installed, network locations, settings etc. We recommend that you read this document before you start working on the server.

Signing out of server 4 and 5:

The configuration of server 4 and 5 is slightly different from server 3. When signing out of these servers you need to point at the right upper corner and you will see an icon "START". Click "START" and then click on your name and a "sign out" will appear. Click on sign-out.

Signing out of server 6:

Click "start", click "your username", click "sign out". Ensure that you lock out properly as soon as your programs have finished. Otherwise your programs will use resources, impeding all other users' response times.

Sending files FROM the servers

It is possible to send files containing results from analyses from the servers to a researcher's own email address. However, it is of great importance that the researcher has made sure that the files do not contain personally identifiable information (for more details – see section 3). Statistics Denmark saves all files that

have been sent for 6 months and conduct random inspection of files to make sure that users comply with the rules. **If rules are violated, the penalty ranges from a personal warning to a permanent lock-out of all users on all the institution's projects at Statistics Denmark. Therefore, make sure that no microdata are transferred.**

To send files to your email click on the icon "Send fil" on the desktop. Write the email address, browse and attach the file. Click "Hjemtag" and the file has been sent.

It is only possible to send files smaller than 3.000 KB and you are not allowed to send compressed files.

Also, it is only possible to send these specific types of files:

Type of file	**Filekstension**
Excel	xls xlsx xlsm xlm xml
Grafik	eps png wmf tif jpg gif emf tif jpeg svg bmp ppt pptx pptm tiff
HTML	htm html mht
Programkode	sas r sps do doh ado
Logs	log
SAS	lst sas7bdat sas7bcat
SPSS	dat
R	rdata rda spv
GAUSS	raw sav spo
Stata	dta gph smcl ster
Tekst	csv cvs tab txt
Latex	tex
PDF	pdf xps
Word	doc docx rtf

For more details, please consult the guidelines (in Danish) at Statistics Denmark's homepage http://www.dst.dk/da/TilSalg/Forskningsservice/Vejledninger

Sending files TO the servers

To place information on the server that is not personally identifiable, e.g. syntax or documentation, please send this to Susanne Bang Vind sbv@dst.dk, with information about project no. and where to place the file. This is typically in WORKDATA.

When data from other sources outside CIRRAU and Statistics Denmark are needed in the project and the data contain personally identifiable information there are two ways to send the data files to Statistics Denmark:
1) Files can be sent by recommended mail to Statistics Denmark on DVD, CD-ROM or USB-stick. They should be sent to "Service desk" and ATTENTION Susanne Bang Vind also with information on project no.

2) Files can be send by email with encryption provided that the institution has obtained Statistics Denmark's certificate. The certificate is found at www.dst.dk/certifikater - choose forskerpost@dst.dk – the VCR-format is recommended for Outlook. The encrypted data files are sent to forskerpost@dst.dk with a

specification of project number and contact person at Statistics Denmark (Susanne Bang Vind) in the subject field.

To speed up the process of reading in the files to the server it is highly recommended that you send the files as SAS files and attach a description of the data files, including an indication of variables that need to be de-identified.

3. Working with micro data and data security

Data in CIRRAU projects are to be treated as confidential information according to the Danish Act on Processing of Personal Data. Violation of the data security is considered a very serious breach of the agreement between the researcher and Statistics Denmark. Non-compliance with the terms may exclude a researcher from access to data at Statistics Denmark for a period of time or permanently. In worst cases, the entire research environment may also be excluded from the research services for at least one month. For that reason, **make sure that you understand and comply with the Statistics Denmark "Terms of Agreement"** (Appendix 1).

TRANSFER OF MICRO DATA IS NOT ALLOWED:

All data in the CIRRAU project database is micro data, which is data concerning individuals, single firms or single institutions. All micro data must be treated as confidential information and must remain on the secure servers at Statistics Denmark. Even though all identifiers such as CPR numbers, addresses and CVR numbers have been replaced by scrambled identifiers, data must be treated as micro data and may not be transferred from the secure servers at Statistics Denmark. Only aggregated data or otherwise completely anonymized data may be copied from the secure servers at Statistics Denmark. Even if you delete identifying variables such as the de-identified CPR number it is still micro data and must not be transferred. The de-identification does not prevent an individual or a company from being identified since many **other variables can indirectly identify individuals** (e.g. date of birth, level of education) or companies (e.g. revenue, value added). As long as the file that you want to transfer contains individual observations NO MATTER what the variables contain, it is NOT allowed.

Most CIRRAU users experienced restricted access to Statistics Denmark for 1 month in December 2014/January 2015 because of an unintended violation of the rules. We would like to avoid a similar situation, so we have imposed the rule that users working under the authorization of NCRR (Center for Registerforskning) are ONLY ALLOWED to transfer TABLES and FIGURES that are intended for publications and do not violate the security rules set up by Statistics Denmark.

This means that users are NOT ALLOWED to transfer log-files, syntax and other types of output. The reason for this decision is that the majority of breaches of the security rules were unintended and caused by transferring e.g. log files with direct or in-direct personally identifiable data.

CIRRAU also impose the rule that all new users initially cannot transfer any files from the servers - not even TABLES and FIGURES that are intended for publications and do not violate the security rules set up by Statistics Denmark. Extraction of files should be made by and in agreement with the project leader. After a grace period of not less than 6 months and after at least 10 successful guided extractions of non-personally identifiable information from the servers, the project leader may grant the user access to extract files from

the servers. These rules are mandatory for all new users under NCRR's approvals at Statistics Denmark, and they are recommended for all new users working under other approvals.

What CAN be transferred from Statistics Denmark?

You are ONLY ALLOWED to transfer aggregated tables or figures in which it is impossible to identify e.g. individuals, households, families, firms or other units with an identifiable number. This applies also even if the identifiable number has been removed. The content of the tables or figures should have a form that would enable you to present it in a publication without breaking the security rules.

Tables should have at least 4 observations in a cell. In business statistics you also need to take into consideration if the two largest companies in a cell account for more than 85% of the total e.g. revenue of the companies in the cell. If this is the case you need to have more observations in this particular cell.

If you transfer several tables make sure that it is not possible to identify individuals, firms etc. by combining some of the tables that you are transferring.

Also be aware that exact medians, minimums, maximums or percentiles must be truncated to values that represent e.g. more than three individuals and that outliers in figures can also be regarded as identifiable information.

What CANNOT be transferred from Statistics Denmark?

- LOG FILES - since especially STATA log files may contain error messages or listings with identifiable information
- TABLES with less than 4 observations in a cell
- TABLES that, in certain combinations, make it possible to identify individuals etc.
- FIGURES with outliers or single data points
- Listings with de-identified variables, such as the personal identification number (PNR), the Business Registration number (CVR), serial workplace numbers (LBNR), address numbers (KOM/BOPIKOM) etc. Even if the numbers are de-identified and the file does not contain anything apart from the de-identified number itself, it is still not allowed.
- Files with information about variables, such as date of birth, level of education, date of death, income, number of children etc. which can be linked to individuals or firms. Even if the de-identified variables as CPR or CVR number have been removed.
- SYNTAX, since may contain identifying information that has been directly coded into the syntax. Syntax transfers are still allowed between projects on a server.

Finally, all output must be manually checked before it is transferred. Transfer of uncontrolled output is not allowed and is also considered a violation of the security rules as users must know exactly what they are transferring.

If you discover that the rules set up by Statistics Denmark have been broken unintendedly please contact Susanne Bang Vind SBV@dst.dk immediately (with Carsten Bøcker Pedersen in copy) since it will be regarded as mitigating circumstances if Statistics Denmark is informed in advance about mistakes.

Remember that the same rules also apply if you have sent your own micro data to Statistics Denmark.

If you are in doubt about whether you are allowed to transfer specific information from the server, then you should aggregate the output further. Or if it is very important contact Jørn Korsbø Petersen, JKP@dst.dk , from Statistics Denmark before transferring, since an unintended violation of the rules can have serious consequences for you and the entire research environment.

Please consult Statistics Denmark's paper for more information on Data Security
http://dst.dk/da/TilSalg/Forskningsservice/Dataadgang

and Statistics Denmark's guidelines regarding transfer of files from the servers
http://www.dst.dk/da/TilSalg/Forskningsservice/Vejledninger

4. Use of macros and formats on the servers

Statistics Denmark formats
Statistics Denmark provides a number of formats which can be used in SAS, STATA and SPSS. The formats allow different categorizations of many variables, such as different levels of education, geography, marital status, the groupings used in Statbank Denmark etc.

As an example, to access the formats in SAS you need to write the following syntax in your SAS editor

```
libname fmt '\\srvfsenas3\formater\SAS formater I Danmarks Statistik\FORMATKATALOG' access=read only;

Options fmtsearch=(fmt.times_personstatistik
                   fmt.times_bbr
                   fmt.times_erhvervsstatistik
                   fmt.brancher
                   fmt.uddannelser
                   fmt.geokoder
                   fmt.disco fmt.statistikbank);
```

Statistics Denmark has set up guidelines on how to use the formats. The guidelines are found through the shortcut "DST-Formater" (looks like a star) on the desktop.

NCRR default formats for SAS on server 3 and 6
NCRR also provides formats in SAS which allow various categorizations of data, such as country of birth, degree of urbanization at place of birth in Denmark, geographical region of birth, etc.

For a description on how to setup in SAS please see the documentation on the desktop intro_srvfsencrrX.docx, where X refer to the server number.

5. Documentation of registers and variables

In CIRRAU, we have compiled an overview of variables from Statistics Denmark and variables from the Family Relations Database in the CIRRAU project database, and you may also find documentation in English for selected variables from Statistics Denmark at:
http://cirrau.au.dk/data-resources/data-documentation/

Documentation of selected registers and variables are also located **at E:\data\documentation** on server 6. You can also access the folder from server 3, 4 and 5.

Descriptions of all variables can be found at Statistics Denmark's homepage (in Danish only).
The 600 most used variables have detailed variable descriptions, including information on data breaches. This documentation is called High Quality Documentation and can be found here:
http://www.dst.dk/da/TilSalg/Forskningsservice/Dokumentation/hoejkvalitetsvariable.aspx

Documentation of other variables produced by Statistics Denmark can be found on the following pages:
Current variables:
http://www.dst.dk/da/Statistik/dokumentation/times.aspx

Historical variables:
http://www.dst.dk/extranet/staticsites/TIMES3/html/Start000-0000-0000-0000-000000000000.htm

6. Tasks supported by CIRRAU

CIRRAU provides access to data through Statistics Denmark for research aimed at interdisciplinary research at Aarhus University, including administrative help in obtaining the relevant permissions for gaining access to data.

When a project has been approved by the relevant authorities, CIRRAU provides free online access to the data sources included in the Family Relations Database as well as the project datasets at Statistics Denmark. CIRRAU provides detailed documentation of the registers included in the Family Relations Database as well as selected variables included in the project dataset (see webpage). CIRRAU associates are encouraged to report documentation on variables in the project database to CIRRAU so that it can be added to the overall documentation. CIRRAU provides online access to copies of the relevant datasets in either SAS or STATA format covering the relevant population. We do not provide technical help regarding data management, usage of formats, analytic methods, or the like.

7. Payment

The CIRRAU project database was established as a long-term data resource from which researchers may access interdisciplinary data capabilities. All data that are already included in the CIRRAU project database can be provided free of charge to associated researchers, whereas additional data must be paid for by the researchers' own funds.

When data are needed from other sources, Statistics Denmark charges a fee for adding the data to the project on the CIRRAU server. This fee is also paid by the researcher.

When researchers wish to add registers and variables that are included in the CIRRAU project database to an existing project it is also free of charge for researchers as long as it is only occasionally, since CIRRAU pays a fee to Statistics Denmark every time new registers and variables are added to an existing project.

The only "payment" to CIRRAU for the data and services provided is affiliation on research papers: on papers based on CIRRAU data, we request that at least one author states CIRRAU affiliation. The authors decide among themselves, whom and how many authors should state affiliation. For details on format, please see the wording used on page1.

APPENDIX 1.

STATISTICS DENMARK TERMS OF AGREEMENT

1. The data sets to which access is given shall be treated as confidential information in accordance with Section 27, subsection 3 of the Danish Administration Act and Section 152 of the Danish Penal Code.

2. Processing of the basic data may only be conducted from the research environment for which the authorization has been granted, or access can also via the authorized research/analysis environment be switched to linked-up home computers in accordance with the guidelines determined by Statistics Denmark.

3. A computer linked up to Statistics Denmark may not be placed at the disposal of other persons, and the connection shall be completely turned off or disconnected, when the computer is not used, i.e. protected against unauthorized use.

4. Passwords, which are supplied by Statistics Denmark for the project are strictly personal and shall not be passed on to any third party.

5. Basic data as well as derived data sets shall not, neither directly nor in-directly, be downloaded.

6. All transfers of output (tables, analytical results), etc. for printing or for further statistical processing shall only take place in accordance with the guidelines and methods determined by Statistics Denmark. A logging of these transfers is conducted by Statistics Denmark.

7. Confidential data shall not be printed, including data at the level of individuals or firms, and all output shall be aggregated in such a manner that it is impossible to identify individual persons or individual firms directly or indirectly. Attempts at identifying individual persons or firms are not permissible.

8. Access to the data is given for the period: Two years with the possibility of extension.

9. No information from the project in which it is possible to identify an individual person or individual firm may be published.

10. Published information from the project shall be submitted to Statistics Denmark for scrutiny.

11. If a “token”, has been provided for the project, it shall be returned to Statistics Denmark when the agreement expires.

On attached researchers in particular.

12. The responsible person signing the agreement of authorization for the authorized Danish institution, shall approve and assume the responsibility that all existing rules governing access to micro data are observed by the associated researcher.

13. It shall be the responsibility of the authorized Danish institution to inform the attached researcher of the rules governing the use of micro data, including the rules of confidentiality in force as well as the rules governing downloading of data

14. The associated researcher's access to micro data shall pass through the authorized Danish institution and can also be switched to linked-up home computers in accordance with the rules governing work from home

15. The authorized Danish institution appoints a contact person undertaking the responsibility for all contact with the attached researcher and Statistics Denmark.

16. All invoices concerning the attached researcher are forwarded to and paid by the authorized Danish institution in question in accordance with the terms of invoicing applicable to the institution.

A breach of the provisions of this agreement will imply that access to the data is immediately denied. Furthermore, the person who has signed this agreement will in future be excluded from using any of Statistics Denmark's research schemes. In the case of minor breaches, the person will be excluded from Statistics Denmark's research schemes temporarily for a period of not less than three years.

This agreement may be terminated by either party at 3 months' notice. If the authorization of the research/analysis environment expires or is changed, this agreement is simultaneously cancelled.