Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

Remote Data Access and On-Site Use at the FDZ of the BA at the IAB

Source: http://doku.iab.de/fdz/access/Vorgaben_DAFE_EN.pdf (Version 8; November 13th, 2017)

Contents

First Steps

Dear user,

the following instruction is supposed to facilitate the work with our data products and to clarify modes of data access. We give advice on how to prepare for working via remote data access, i. e. remote data execution/processing, or on-site use.

For this purpose, we provide test data, frequency tables, template programs and data reports for each data product on our website. We now also offer a compilation of template do-files covering the processes of data preparation, description, analysis and presentation of results.

In the following, we outline the first steps of data preparation and analysis ranging from the preparation of the test data to working on-site or via remote data access. The instructions are use cases based on the test data of the Sample of Integrated Labour Market Biographies (SIAB).

Please note that this instruction does not replace the guidelines for remote data access and on-site use. It is a complement to our guidelines and refers to them on several occasions. **Please do read** the guidelines for remote data processing and on-site use before you create programs for remote data access and on-site use.

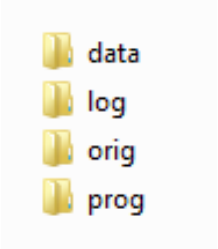
We are looking forward to any criticism or suggestions on how to improve these guidelines: iab.fdz@iab.de

I. Preparation of Remote Data Processing and On-Site Use

Set-up of test environment at your own workplace

- Create a directory <fdz*[your projectnumber]*> on your PC.
- Create the folders **orig**, **prog**, **data** and **log** within this directory.
- Do not create any subfolders.
- In case you use more than one data product for your project, please create a project directory for each data set as described above. Please note that users cannot transfer (copy or move) data between directories at the FDZ.

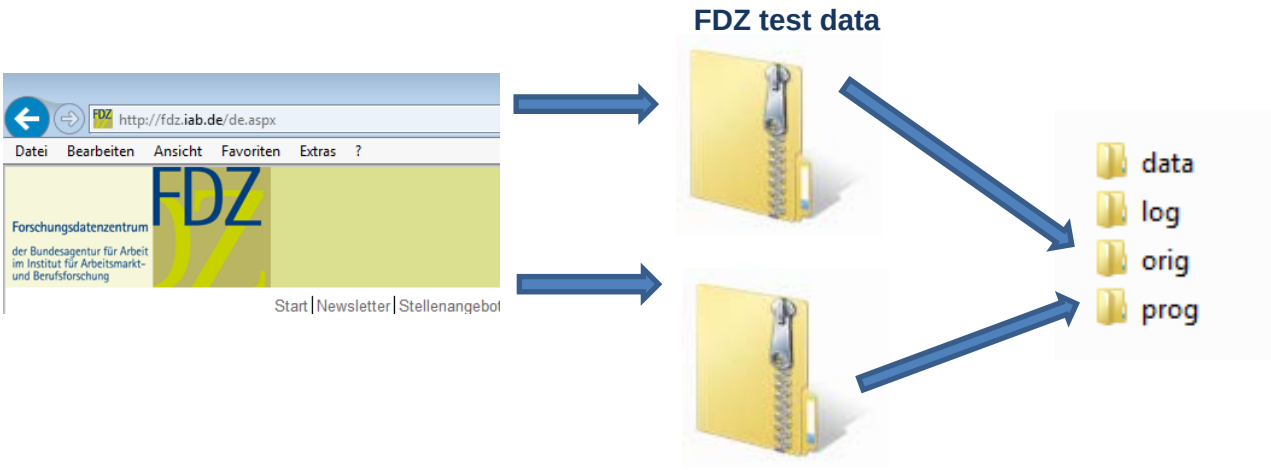
Fig. 1: Directory structure



Downloading and preparing the test data

- You can download the test data from the FDZ website under the respective data product. Please note that for certain data products no test data are available.
- Save the test data in the **orig** folder.
- Many data products of the FDZ contain sensitive variables which you have to request separately. The test data generally include all sensitive variables. Therefore, you need to modify the test data in advance, according to the sensitive variables that you have requested (see FDZ template **prepare_test_environment.do**).
- There are no test data for expansion modules. These additional data sets can be created following the descriptions of the data set in the data report. For this purpose, you may have to fill variables with random numbers.

Fig. 2: Downloading the test data and templates



Usage of template do-files

- The do-files can be downloaded from http://doku.iab.de/fdz/access/FDZtemplate.zip.
- Save the do-files in the **prog** folder.
- The do-files were developed with the SIAB test data. You can test the setup of your test environment and the FDZ templates by downloading the test data from our website (http://fdz.iab.de/en/FDZ_Individual_Data/integrated_labour_market_biographies.aspx). For other data sets you have to adjust the do-files accordingly.

Templates to prepare working with the test data at your own workplace:

- **profile.do**
  Do-files with this file name are automatically executed upon launching Stata if they are stored in one of the designated folders (for more details see the Stata help and the template). When launching Stata, this file automatically defines global macros for the paths of the folders stated above. Adapt this file to your working directory.

- **prepare_test_environment.do**
  Your requested data set usually does not contain all sensitive variables such as wo_kreis (place of residence: district) or nation (nationality), but instead contains the respective coarsened variables like wo_bula (place of residence: federal state) or nation_gr (nationality: aggregated). This do-file shows how to adjust the test data to the structure of your authorised research data set. Furthermore an example is given on how to generate your own test data for the requested expansion modules. This do-file only needs to be executed once.[1]

[1] Additional information for projects that use a specially anonymised data set at foreign sites: Please adjust your test data to the structure of your data set at the FDZ. Therefore, delete or anonymise variables corresponding to the anonymisation concept agreed upon with the FDZ (Example: Anonymisation of the variable nationality to a dummy variable with the values German and non-German citizenship).

Templates for evaluation programs:

In the following, we summarize the purpose of the individual template do-files. All evaluation programs are numbered consecutively in the order in which they should be executed. In the example initials are used in the file name to differentiate the programs in case of a project with multiple users. In this example, **JD** stands for the fictional user John Doe.

- **master.do**
  A file with the name **master.do** (with exactly this file name) is always necessary. With this do-file you execute the do-files for data preparation, description and analysis. The file serves as a program schedule and makes sure that all results are reproducible. Further notes can be found in Chapter 2 of the guidelines. Please note that do-files are only executed if they are called by *do ${prog}/do-file.do* in the master do-file.

- **jd01_dataprep.do**
  Prior to the actual data analysis, you usually undertake further steps of data preparation: You might delete observations which are not relevant for the analysis or generate new variables (e. g., a dummy variable for East and West Germany, or the employment duration). The edited data set can be saved under the path **$data** and loaded by other do-files. This template do-file includes several exemplary preparation steps.

- **jd02_describe.do**
  For descriptive analyses, the Stata commands **summarize** (sum) and **tabulate** (tab) are very handy for a quick output review. This do-file explains both commands. Further information about how to present descriptive statistics can be found in Chapter 2.2 of the guidelines.

- **jd03_analyses.do**
  This do-file exemplarily shows a simple least squares regression with the **regress command.** Please note that you need to give additional descriptions of the independent variables and covariates above the output of estimation results (see Chapter 2.2 for further information).

- **jd04_mergebhp.do**
  For some FDZ data products, certain establishment characteristics from the Establishment History Panel (BHP) can be requested. This do-file shows for the SIAB example how to merge the BHP basic establishment data. It also presents some simple evaluation examples.

- **jd05_graphs.do**
  This do-file presents some use cases on how to create graphs in accordance with the FDZ guidelines. Please follow the directions given in Chapter 5.4.

- **jd06_cleanup.do**
  For reasons of data parsimony and the memory restriction of 30 GB in the **data** folder, you should delete temporary files after you run a program as is shown in this do-file.

External aggregate data

- Merging aggregate data (such as official statistics) to FDZ data products is generally permitted. However, data sets need to be examined and approved in advance by the FDZ. Further notes can be found in Chapter 3.2 of the guidelines. To test merging and analysing external data in your test environment at home, please save the external data in the **orig** folder.

Ado-files

- If you want to use ado files, save them in the **prog** folder. Further information can be found in Chapter 3.1 of our guidelines.[2]

II. Working via Remote Data Access and On-Site Use

The weakly-anonymised data sets of the FDZ can be accessed via remote data access or on-site use. Contracts for on-site use automatically include the use of remote data access. Information about which mode of data access is available for your project can be found in your data use agreement. For both access modes, the FDZ uses the web-based application **JoSuA** (Job Submission Application):

- With respect to **remote data access**, JoSuA can be used to upload do-files, to inspect preliminary results and to download reviewed and released results.

- With respect to **on-site use,** do-files you created at your own workplace in advance have to be uploaded to JoSuA. After review by FDZ, you can download these files from the local network at the FDZ and then further edit them during on-site use. In addition, do-files that were created or changed during an on-site use can be uploaded to JoSuA again. After review, these files are available for remote data access and can be further edited at your own workplace. The same procedure applies if you need reviewed results after on-site use.

[2] Stata looks for ado-files in multiple directories by default. In order to avoid that ado-files are included during the test at home and then are not available during remote data access or on-site use, the search path in the test environment should be restricted to **$prog** (see template **profile.do**).

III. On-Site Use

On-site use prior to remote data access

We recommend to work on-site before using remote data access for the first time, especially when analysing complex data sets like SIAB, the Linked Employer-Employee data of the IAB (LIAB) or other linked data products. In doing so, you can become familiar with the data more quickly and work directly with your data set, which is sometimes very complex to process.

Making an appointment

For on-site use at the FDZ or an external site, please make an appointment at least two weeks in advance (contact details). Keep in mind that the capacity of guest workstations is limited and not every requested appointment can be approved.

Files for on-site use

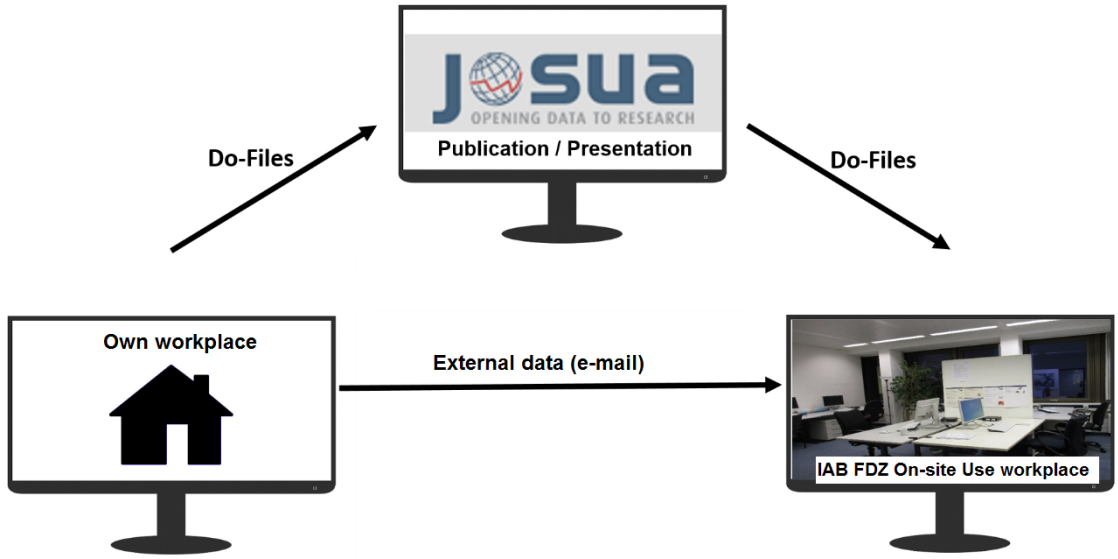
Data sets that were created via remote data access and stored in the folder **data** are available during on-site use.

- **Do-files:**
  To use evaluation programs on site that were prepared off site, please follow the descriptions in Chapter 4.4 of our guidelines. Please note that all do-files have to be uploaded to JoSuA in Presentation/Publication mode at least three business days prior to your on-site visit. The reviewed and released do-files can then be downloaded from JoSuA at your on-site workstation. You can save them in the **prog** folder of your project directory at the FDZ. To upload do-files via JoSuA, please make sure that you do not run any do-files via master.do. Please put comment signs around the lines calling other do-files or put the word “STOP“ before the execution of do-files so that the program stops. In doing so it can be prevented that programs are executed and unnecessarily consume server capacity.

- **Ado-files:**
  If you want to use ado-files on site, please read Chapter 3.1 of our guidelines.

- **External aggregate data:**
  External aggregate data must be sent to the FDZ via email at least three business days prior to your on-site visit. The data set must contain information on the number of observations underlying each variable. Please provide a detailed variable description for the data set and state the source of the data. We recommend to discuss the merging of aggregate data with the FDZ at an early stage. Further information can be found in Chapter 3.2 of the guidelines.

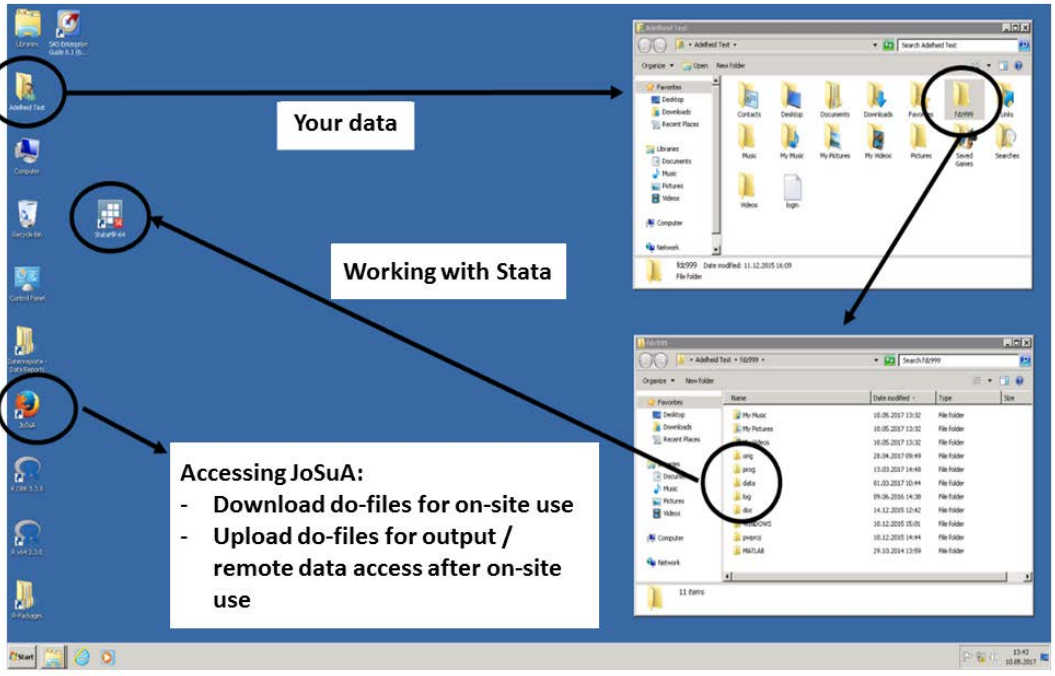
Fig. 3: Data transfer prior to on-site use



During on-site use

- Additional information on on-site use, the FDZ Datenreport series, and central papers of the FDZ Methodenreport series are available in a directory accessible to all guests. There you will also find this document and the templates presented above.
- Take the opportunity of information exchange with the staff of the FDZ at the site in Nuremberg.

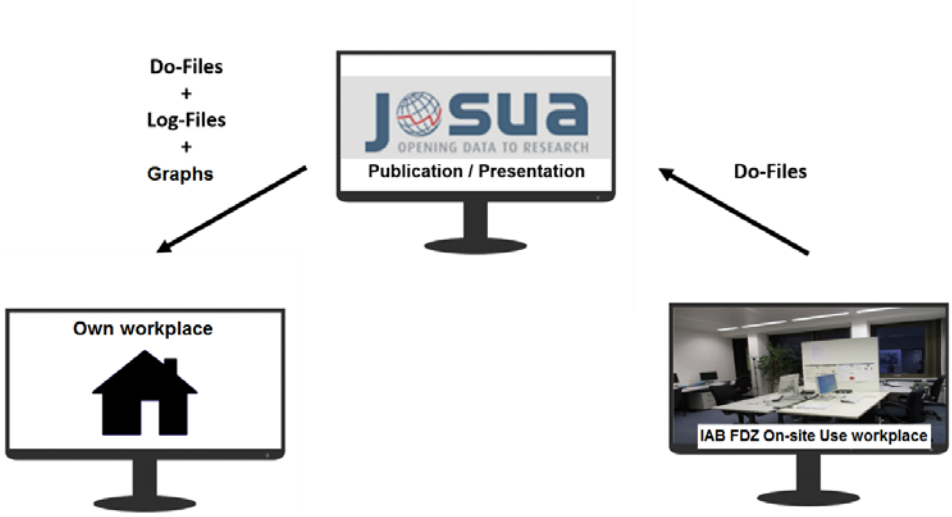
Fig. 4: Desktop at the on-site workstation



At the end of your on-site visit

- Please upload your do-files to JoSuA in Presentation/Publication mode at the end of your on-site visit. Only then the programs will be available for further editing at your own workplace or for remote data access.
- If you do not immediately need the results for a publication, but only need the do-files for further editing, please include a stop command in your master.do as described in Chapter 4.5.
- Please log out correctly from the server at the end of your stay.

Fig. 5: Transfer of data after on-site use



IV. Remote Data Access

Using JoSuA

A detailed instruction for using JoSuA for remote data access can be found in Chapter 8 of our guidelines. The following notes describe when and how to use JoSuA for your analyses and what you should bear in mind when creating do-files.

- Make sure that your do-files are executable without errors in your test environment before uploading them to JoSuA.
- In JoSuA the upload limit per job is **20 do-files**. The job limit per user of a project is **one job per day.** The daily job limit also persists after a job has aborted with an error.
- Make sure to specify paths with the following global macros: **$orig, $data, $prog, $log**.
- If necessary, upload ado-files to JoSuA. For installing user-written ado-files from community repositories, see Chapter 3.1 of our guidelines. Ado-files stored in the prog folder of your project directory at the FDZ are automatically recognized.
- Running time and data minimization:
    - The disk space of your **data** folder is limited to **30 GB**.
    - The **working memory** limit is **30 GB.** Your jobs are terminated automatically if they exceed the working memory limit.
    - In general, we urge you not to waste working memory. When the compute servers have a high workload, your job and the jobs of other users may terminate.
    - The file size of your test data set is a small fraction of the original data set. The original data sets and your processed data sets are larger and the jobs need considerably more processing time.
    - Please consider the notes on data minimization in Chapter 2.4 as well as the examples given in the template do-files.

Internal Use Mode

Please keep the following in mind when using the Internal Use mode:

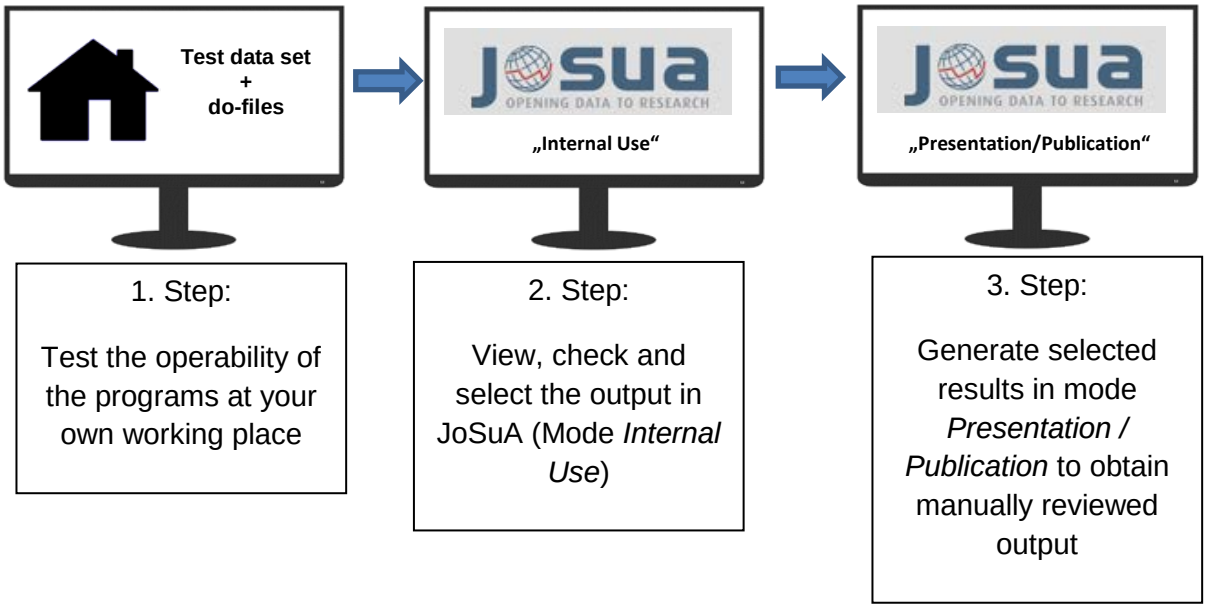
- The Internal Use mode can be used for data preparation, the preparation of on-site use, and for testing your programs. Due to data privacy regulations, the visible output will be automatically checked and censored. The output can be viewed within JoSuA's web interface but cannot be downloaded. The output must not be photographed, copied, transcribed, or printed. It is strictly forbidden to show the results to third persons. The output may only be used for further developing your programs. All other usage is a breach of contract (see Chapter 5.1 of the guidelines).
- The output is only visible until your next job in the Presentation / Publication mode.

Presentation / Publication Mode

This mode can be used to create downloadable results for publications or presentations.

- Please check the operability of your do-files in the Internal Use mode before using the Presentation / Publication mode.
- Restrict the output to those results that you actually need for a publication or presentation.
- Make sure to display your results in accordance with our guidelines. In our guidelines, you can find information on the output of tables, graphs (Chapter 5.5) and aggregated data sets (Chapter 5.5).
- Take a look at our template do-files for efficient programming and output display. This way, you can make sure that the programs are efficient and the output review by the FDZ can be made swiftly.

Fig. 6: Procedure remote data access with JoSuA



V. Working Tools

FAQ JoSuA
http://fdz.iab.de/de/FDZ_Data_Access/FDZ_Remote_Data_Access/faq.aspx

Test data
In general, the FDZ offers test data for the data sets that are available via on-site use and remote data access. Specific test data for each data set can be found on our website: http://fdz.iab.de/en.aspx

Data protection at the FDZ
Hochfellner, Daniela; Müller, Dana; Schmucker, Alexandra; Roß, Elisabeth (2012): Data protection at the Research Data Centre. FDZ-Methodenreport, 06/2012 (en), Nürnberg
http://doku.iab.de/fdz/reporte/2012/MR_06-12_EN.pdf

JoSuA
Eberle, Johanna; Müller, Dana; Heining, Jörg (2017): A modern job submission application to access IAB's confidential administrative and survey research data. FDZ-Methodenreport, 01/2017 (en), Nürnberg
http://doku.iab.de/fdz/reporte/2017/MR_01-17_EN.pdf

Working tools of the FDZ and publications on working with FDZ data
http://fdz.iab.de/en/FDZ_Overview_of_Data/working_tools.aspx

Guidelines for Remote Data Access and On-Site Use

1 Data Access at the Research Data Centre (FDZ)

1.1 Structure of the Project Directories

- Each research project receives its own directory (fdzXXXX) **per** FDZ data set. Different data sets cannot be used together.
- Each project directory contains the following subdirectories. Creating any other directories or subdirectories is not allowed.

orig:	This directory contains all the requested original data. The folder may also contain external aggregate data (see Chapter 3.2). You have read-only permission for this directory.
data:	All user-generated data sets are stored here. A **maximum** of **30 GB** is allowed per directory. If you exceed this maximum, you will not be able to submit new jobs until you have reduced your disk space.

The data in the directories orig and data can also be accessed by programs via JoSuA.

prog:	This folder contains all the evaluation programs for the on-site use (do-files) as well as further files (e.g. ado-files).
log:	This folder contains all the result files of the on-site use, including graphs.
doc:	This folder is for PDF documents that were previously submitted via email and for taking notes during on-site use. These documents are only available while on-site use and cannot be released.

Do- and Log-Files are not automatically synchronised between JoSuA and the project directory.

1.2 Job Submission Application (JoSuA)

- The FDZ uses the software application JoSuA which has been developed by the Institute of Labor Economics for
  - remote data execution,
  - the upload of do-files before on site-use and
  - provision of results and do-files after on site-use.
- Every user gets an account for JoSuA after a use agreement is concluded. It is forbidden to share the password with others.[3]
- Using JoSuA, data sets in the folders **orig** and **data** can be accessed. Please consider the programming advice given in Example 2.

[3] This user name and the initial password also apply to the on-site use. Any password changes in JoSuA will not be synchronised with the account for on-site use.

- JoSuA does not synchronize do- and log-files with the directories prog and log. Do- and log-files are stored elsewhere.
- Chapter 8 describes the JoSuA web interface in more detail.

2 Structure of the Program Files

2.1 The Master File

- All do-files used must be run from a master file. This file must always be named master.do. Its structure is illustrated in Example 1.
- Use the Stata global variables **$orig, $prog, $data, $log, $doc** to access your files. These variables are automatically made available for you during remote execution and on-site use and must not be overwritten.
- You cannot use *cd* commands or relative path definitions.
- Never use the *set max_memory* command.
- Certain commands are censored by the system. In these cases, you'll see the line commented out with the tag *** CENSORED ***.
- The master file must contain all evaluation programs used during the project. It is important to add a short description of the contents of the respective routines in the master file after every program start command.
- The evaluation programs must be numbered in the order in which they run. If several users are working separately on a project, their programs may be distinguished by adding abbreviations to the file names, e. g. *th01_PersDat.do*. This does **not** apply to the **master.do**.
- All files you generated (result files, analysis data sets, etc.) must be replicable based on your programs and the original data.
- Program steps that have already been executed and have not changed since the last submission must be commented out in the master file. Data sets stored in the folders orig and data can be accessed using JoSuA. Do- and log-files are stored on separate servers though.
- For JoSuA we recommend to set the linesize to 120 in Internal Use mode because that corresponds with the width of the output windows in JoSuA. Use the command *set linesize*.

Example 1: Contents of file master.do

```
* Master File of the project „Project name“
* Users: T. Hill (th) and W. Miller (wm)

version 14[4]
clear
set more off                  //so the program will scroll down all the way
*set linesize 120            //max. linesize Internal Use
*set linesize 255            //max. linesize Presentation / Publication

*do "$prog/th01_PersDat.do"   // preparation of individual-level data
*do "$prog/th02_BetrDat.do"   // preparation of establishment data
*do "$prog/th03_DatXY.do"     // merging
*do "$prog/th04_AnalyseA.do"  // examination of Hypothesis 1 (...)
*do "$prog/th05_AnalyseB.do"  // examination of Hypothesis 2 (...)
*do "$prog/wm01_Datagen.do"   // preparation of individual-level data
do "$prog/wm02_AnalyseX.do"   // examination of Hypothesis 5 (...)
do "$prog/wm03_AnalyseY.do"   // examination of Hypothesis 6 (...)


* Output of a file list in a .log file
capture log close
log using "$log/filelist.log", replace
dir $prog/
dir $log/
dir $data/
dir $orig/
log close
```

2.2 Program Files

- Result files (including graphs) must have the same file name as the program with which they were generated. For example, **01_PersDat.do** will generate the result file **01_PersDat.log**.
- The programs must be documented for the privacy review (see Example 2). Please take the following into account:
    - Open a log-file at the beginning of every do-file, before every other command and before the content-related description of the do-file,
    - Use interim headlines, such as “Preparation of the variables X” or “Establishment-level aggregation”,
    - Describe the content and function of loops,
    - Create variable labels and (if appropriate) value labels for all important variables that were newly generated,
    - Use meaningful variable names whenever possible, and

[4] We recommend to always state the most current Stata version that is available at the FDZ to avoid version conflicts and program terminations. Only use older versions if you need downward compatible program execution. The Stata version currently used at the FDZ can be found in JoSuA.

  - Provide a short description of newly generated variables immediately above descriptive statistics.

- Never copy results or IDs of establishments or individuals into your do-files.
- Any inclusion of data in do-files and scripts is strictly prohibited.[5] External aggregate data must be sent as data sets to the FDZ (see Chapter 3.2).

Example 2: Contents of a program file for data analysis

```
*Write log file:
capture log close
log using "$log/th04_AnalyseA.log", replace

* Program file of the project „Projektname“ - th04_AnalyseA.do
* Content: Examination of Hypothesis 1

* Aggregation level: industries by federal state
*                    aggregation in th03_DatXY.do
*                    min. 20 establishments per data line
*                    number of individuals per line is specified in n_pers

clear all

* Load original data:
use "$orig/Datensatzname", clear

* Analyses
[...]

* Save generated data:
save "$data/newname", replace
log close
```

2.2 Documenting the Number of Observations

- During our data privacy review (see Chapter 5.1), all values that are based on less than 20 observations (individuals and/or establishments) will be deleted. In order to prevent reverse calculation of deleted values, other values often have to be deleted or rounded as well.
- Therefore, for all descriptive or multivariate analyses, the number of cases they are based on must be specified. If the number of cases is missing, the corresponding results will be deleted.
- When displaying means, the minimum, maximum, and standard deviation must also be specified.
- When displaying ratios, the number of valid observations in each subgroup has to be displayed as well.
- When using weights in descriptive analyses, analogue unweighted results must always be specified as well. The weighted and the corresponding unweighted result tables must

[5] This does not apply to classifications and conversion tables (e.g., contribution assessment ceilings, DM-Euro translation, GDP deflators, CPIs, transcoding of industry codes, districts or employment agency districts).

always be listed one directly below the other. Please also make sure you program your loops accordingly.

- If the job is submitted in Presentation/Publication mode, a description of each variable used in a model must be included in front of multivariate analyses for disclosure review.
- For analyses of linked employer employee data, both the numbers of individuals and establishments have to be reported in descriptive tables.

2.3 Documenting Data Aggregations

- If the original micro data are aggregated (e.g. at establishment or region level, etc.), this must be described accordingly in the analysis program.
- If those aggregates are displayed later in the program, the underlying number of cases has to be computed:
    - For each aggregated variable there has to be an additional variable which contains the number of valid cases (individuals and/or establishments) per aggregate (e.g. **bysort aggregate: egen N_var1 = count(var1)**).
    - If the aggregate is a ratio, the number of valid cases has to be computed for each subgroup of the aggregate (e.g. number of men in state X and number of women in state X in addition to the ratio of women in state X).
    - If the **collapse** command is used, the number of not-missing values per variable has to be calculated: **collapse (count) N_var1 = var1 (count) N_var2 = var2.** It is not sufficient to simply count the number of lines via „collapse (count) persnr“.
- All result files generated after an aggregation must contain the following information at the beginning:
    - at which level the aggregation took place,
    - during which step in the program the aggregation took place (also specify in master.do),
    - the minimum number of individuals and establishments per data line (e.g.: cells containing <20 establishments were deleted), and
    - the name of the variable that contains the observation count per data line.
- If the corresponding references to the aggregation are missing from the beginning of the result file, it is possible that too much data gets deleted during the privacy review process.
- Taking account of the rule of data economy, please keep the amount of aggregated tables to a minimum.

2.4 Efficient Data Handling

- Use the storage space in your **data** directory sparingly. Once the maximum of **30 GB** is exceeded, no more jobs can be processed (except deletion programs, see Example 3) until the storage space is reduced.
- Temporary data sets should be saved to disk only if necessary for subsequent analyses. Data sets need to be reproducible with your do-files. Delete data sets that were generated in the meantime as soon as you no longer need them (see Example 3).
- Limit your data set to those variables and observations that you actually need for your analyses (see Example 4 and Example 5).
- When developing programs, it might be helpful to test with a subsample (using the command **sample** for instance).
- Use the **compress** command when saving data sets.
- The available RAM memory is limited to **30 GB per user and project**.

Example 3: Deletion program

```
* Program file to delete data that is no longer needed

capture log close
log using "$log/07_delete.log", replace

* Delete files:
capture erase "$data\test1.dta"  // erase "$data\test*" does not work
capture erase "$data\test2.dta"
[etc.]

log close
```

Example 4: Panel of BHP with year, establishment ID, and number of employees

```
use betnr az_ges using $orig\bhp_7514_m06_p50_v1_1975.dta, clear
gen int jahr = 1975   // Specify data type integer instead of float
forvalues j = 1976/2014 {
   append using $orig\bhp_7514_m06_p50_v1_`j'.dta, keep(betnr az_ges)
   replace jahr = `j' if missing(jahr)
}
```

Example 5: Merging employees from SIAB with basis establishment file

```
* Select employment spells:
use if quelle == 1 using $orig\siab_7514_v1.dta, clear

keep persnr betnr begepi tentgelt    // Select relevant variables
keep if year(begepi) >= 2000         // Limit to spells since 2000

* Merge variable w08_3 from basis establishment file:
gen int jahr = year(begepi)
merge m:1 betnr jahr ///
    using $orig\siab_7514_v1_bhp_basis_v1.dta, keepusing(w08_3)
keep if _merge ==3
drop _merge
```

3 Further Resources

3.1 Ado-files

- Ado-files cannot be downloaded from the internet at the on-site workstation at FDZ. The FDZ provides selected packages in an offline repository in its network (see the list at http://doku.iab.de/fdz/access/stata_ado.pdf)

- The repository can be used during on-site use and via remote data access. Please use the following command to install packages that are available at the FDZ. The command copies the corresponding files in your prog directory:

  ```
  net install packagename
  ```

- This command singularly copies all relevant data into the prog folder of the given project. This command should only be used once. Afterwards, the packages are available in the prog folder and are automatically recognized by Stata. All ado-files in the prog folder can be used when running jobs with JoSuA without further ado.

- The repository is updated at regular intervals. However, the files copied into the prog folder by using the net install command are not updated automatically. To copy a newer version of an ado package into the prog folder, use the following command:

  ```
  adoupdate
  ```

  (In some cases, only the command *net install [packagename], force* will work.)

- If you want to use other ado-files or own ado-files, please send them to the FDZ mailbox at least three days in advance of your on-site use. The packages cannot be provided at short notice. These ado-files are saved in the prog folder as well.

- If you intend to use ado-files exclusively for remote data processing, please upload your files to JoSuA. Ado-files need to be uploaded once in the *Projects* tab (see Chapter 8.4). Afterwards they can be used for remote data processing but not for on-site use.

3.2 External Aggregate Data

- External variables on an aggregated level (e.g. unemployment rates by districts) may be merged to the data if they comply with the FDZ privacy guidelines.
- Aggregated values must refer to at least 3 observations. It is not possible to merge individual observations. Variables have to be included in the data set giving the respective number of observations underlying each variable.
- The corresponding variables must be transferred as a data set to the FDZ together with a description of the data set (including variable descriptions, the aggregation level, and a source citation). It is not allowed to transmit aggregate data within a program file.
- After inspection and approval by the FDZ, the external data sets are available in the **orig** folder.

4 Data Usage

Please bear the detailed notes of First Steps as well as the templates for do-files in mind.

4.1 Testing Evaluation Programs

- Before submitting your programs for remote data access or before on-site use, please test their correct functioning by running master.do with the test data provided on the FDZ website.
- For JoSuA and on-site use, the path globals are automatically defined. For your own test environment you have to define the globals **$orig, $data, $prog, $log** and, if applicable, **adopath** before running master.do. Do **not** write the global definitions into master.do but into a file called **profile.do** (see First Steps). This file is automatically run by Stata and does not have to be called by master.do. You could for instance save the file in the current working directory (see also http://www.stata.com/help.cgi?profile). Do not upload the file to JoSuA.
- Run master.do in your test environment.
- After your tests have been successful, upload your programs **unchanged** to JoSuA.

4.2 Remote Data Access

- The programs you submit must be working. We do not modify any programs uploaded to JoSuA.
- We run your evaluation programs using the most recent Stata release available to us.
- Please do not send in a new program before you have received your previous results for this project.
- The timely provision of results depends on server utilisation and the complexity of computation and review. Please consider this in your planning.

4.3 Selecting the Job Mode in JoSuA

- Two usage modes are available in JoSuA:

1) Internal Use

  - The results are for “internal” purposes only, meaning that they can only be used to develop scripts.
  - Uploaded do-files will start automatically. The results are available via on-screen display.
  - There is a script-based disclosure review.
  - Results must not be copied, printed or extracted otherwise.
  - Results must not be shared with persons outside your project’s data use agreement. This also includes that third parties are not allowed to inspect results.
  - The use of results obtained from the Internal Use mode in presentations or publications is an infringement of the data privacy agreement.

- Please use the Internal Use mode until you need interim or final results for a presentation or a publication.

2) Presentation / Publication:

- In this mode, downloadable log-files are produced.
- This mode is to be selected whenever results are about to be published in a paper or a presentation. Check in advance if your programs run without any errors in the Internal Use Mode. Restrict the results to only the ones you need for publication.
- The FDZ will start your uploaded do-files and will perform manual disclosure review of the results.
- Please upload only do-files that are relevant for the presentation or publication. Put comment signs around any commands whose results you do not need or place the prefix "quietly" in front.

- It is not allowed to perform reverse calculations to identify any deleted values. Results of the Internal Use mode will therefore not be available once a Presentation / Publication job is submitted.
- Please keep in mind that some Stata-commands are forbidden in general with the **Internal Use** mode to protect data privacy.[6] Exceptions are not possible. In case your results cannot be produced with other Stata commands, you have the possibility to use the Presentation / Publication mode.
- Advice in case of job terminations can be found in Chapter 8.6.

4.4 Preparation of On-Site Use

- If you want to use any prepared evaluation programs on-site at the FDZ, you need to upload them via JoSuA in Presentation / Publication mode **at least three business days** prior to your first day of on-site. Please note in JoSuA's comment field that you need the do-files for on-site use.
- Please put the word **„STOP"** near the beginning of master.do or put comment signs around the lines calling other do-files, causing an error in the program run. This prevents the job from using computation time and producing output prior to your on-site use. The job is run by FDZ in order to make the do-files accessible in JoSuA.
- In case master.do was not modified in the way described above, all do-files that are called by master.do will automatically run. This may cause delays in the provision of do-files depending on processing time. Do-files will only be reviewed and approved once the program has finished.
- All do-files uploaded will be able to download from an internal server at the FDZ and saved in the **prog** folder after censoring.
- There is a shortcut to JoSuA on the desktop of your Stata server on-site.

[6] Currently the command list is censored in Internal Use mode. The FDZ may censor further commands in the future.

- If you intend to use any (aggregate) data or ado-files on-site at the FDZ, you need to send them to the FDZ mailbox **at least three business days** prior to the first day of on-site use. Please state the date of your on-site use in the email. If you have not received an acknowledgement of receipt two business days after sending your email, please let us know. Please see the Chapters covering ado-files (Chapter 3.1) and external aggregate data sets (Chapter 3.2).
- At your on-site workstation, you have access to a general directory (folder on the desktop), in which all relevant FDZ data reports, further working aids and central method reports (incl. corresponding program files) as well as related documents are available.

4.5 Return of the Reviewed Results after On-Site Use

- After on-site use, you need to start your modified programs in JoSuA. Results generated during your on-site stay will not be provided.
- At the end of your on-site stay, please upload all modified do-files to JoSuA and select the mode Presentation / Publication.
- Please make sure that master.do and the other program files are fully operable.
- There is a shortcut to JoSuA on the desktop of your Stata server.
- After a manual disclosure review, you receive your results and all uploaded do-files.
- You can run all do-files in case you immediately need the results for a presentation or publication. If you just need your do-files and do not want the program to run, you need to write “STOP” at the beginning of your master file or put comment signs around the commands calling other do-files. All uploaded jobs will always be started by FDZ.
- All uploaded do-files are downloadable in JoSuA after censoring. The FDZ does not transmit do-files from the project directory afterwards.

5 Results

5.1 Privacy Guidelines

- Results generated during on-site use or remote data access are reviewed by the FDZ team for compliance with data privacy guidelines and afterwards transmitted to the users. Disclosure review for analyses in Presentation / Publication mode normally takes up to 5 business days.
- Please read the information on how to document observation counts in Chapter 2.2 as well as aggregations in Chapter 2.3.
- Individual results will be deleted if they do not comply with our privacy regulations.
- Detailed information on the guidelines and examples can be found in the FDZ-Methodenreport (06/2012) by Hochfellner et al. (2012)[7].

[7] Hochfellner, Daniela; Müller, Dana; Schmucker, Alexandra; Roß, Elisabeth (2012): Data protection at the Research Data Centre. FDZ-Methodenreport, 06/2012 (en).

- The user pledges not to perform reverse calculations to identify any deleted values.
- In the event of an infringement, the user will be excluded from any further data usage for up to two years. Moreover, the penalties specified in the contract of use may be applied.
- Information on any breach of the obligations named above will be passed on to other research data centres.

5.2 Permitted Output Formats

- Stata result files must have the extension ".log" or ".txt".
- Results saved outside the Stata log-files must be written back to the log-file (see Example 6). For disclosure review, all results have to be printed in a readable format right in front of the included text.

Example 6: Including output in a log-file

```
* Including results which were returned outside the log files
[…]                             // Show results in a table or list
* Output content of a .csv or .txt file in the Stata window
type $log/ergebnisse.csv
type $log/ergebnisse.txt

* Output LaTeX code directly in Stata (example for a regression A and B)
[…]                             // Show results of the regression
esttab A B, c(b se), tex
```

5.3 Output Limits in JoSuA

Please consider the following output limits per job in JoSuA. If your job exceeds any of the limits, any surplus files are left out and files are cut off.

- Maximum number of output files (including graphs)[8]: 20
- Maximum size of a log-file: 2 MB
- Maximum size of all output files combined: 8 MB

5.4 Graphs

- Stata graphs (.gph) can only be transmitted if they are generated with the "asis" option (see Example 7). The formats wmf, eps and png are also allowed.
- In Internal Use mode, only the format **png** can be processed.
- Please generate graphs only if it is not possible to create them on the basis of tables with e.g. Excel on your own. The number of generated graphs should be as low as possible as the effort of disclosure control on graphs is very high.
- Please provide evidence on how many observations every depicted value is based on. This is possible by displaying the observations either right in the graphs or in tables placed directly before or after the generation of the graphs.

[8] In Internal Use Mode, the limit applies to the number of files prior to the conversion into preview graphs.

- Save the graphs in the folder addressed with the global **$log.**

Example 7: Output of graphs in valid format

```
* Produce graph in Stata format:
graph ... , ... saving($log\mygraph, replace asis)
* or
graph save $log\mygraph, replace asis
* Export of graphs in png-format (Internal Use Mode)
graph export $log\mygraph.png, replace
```

5.5 Transfer of User-Generated Aggregate Data Sets

- If you wish to receive aggregate data sets as result file, please consult with us about the procedure in advance.
- The generation of the aggregated data should take place during on-site use. After finishing the aggregated data, the data generating program has to be uploaded in JoSuA with the Presentation / Publication mode. Please write in the comment field that you want the aggregated data and the respective file name. Always save the data set in the **data** folder.
- The regulations for aggregated variables also apply to the documentation of the aggregated data sets (see Chapter 2.3).
- For each project, such a data set can be forwarded **only once**. The FDZ will transmit these data sets via email. This also applies if you want to have an aggregate data set transferred to a different FDZ project folder in order to merge it with other data sets.

6 Publication of Results

- The data and documentation must be properly cited in publications according to the specifications on the FDZ website (http://doku.iab.de/fdz/access/Zitierweisen_e.PDF).
- Please send a specimen copy to the FDZ.

7 Special Arrangements for Other Software

- By default, the FDZ offers analyses with Stata. If you need any other statistical programs for your evaluation purposes, please contact the FDZ early on. All jobs that rely on software such as Matlab or R therefore still need to be sent in via email. Please use this mode of remote data access only for those parts of your analyses that cannot be performed in Stata.
- All analysis and data preparation which are possible in Stata should be executed in Stata. Furthermore write a command to select the other software in the do-file.
- Only Stata is possible with JoSuA. If you use other software packages you need to send the do-files for **remote data execution** or **prior to on-site use** to iab.fdz@iab.de. Please consider the following details:
    - In order to process your orders, we need you to state the project specifications in the subject line of every email (see Example 8).

- Newly sent files will overwrite existing files if they have the same file name.
- Please already point out in the subject line under *Other* what must be taken into consideration when working with the programs (e.g. “Matlab”).
- Please note that we will send orders back unprocessed if any relevant information is missing from the subject line or the format requirements have not been complied with.
- Please do not send in a new program before you have received your previous results for this project. Please do not use JoSuA in the Presentation / Publication mode in this period.

- After on-site use:
  - After on-site use, the FDZ team will not return the results generated during the stay but instead run the corresponding programs again. Therefore, please make sure that master.do and the other program files are fully operable.
  - If you wish to receive program files and/or results after your on-site use period, please send an email to the FDZ mailbox (see Example 8). Add “after on-site use” to the subject line and specify in the email the date of the on-site use period and the program files you would like us to send you.
  - We will run master.do, and you will receive the reviewed results generated during this run of the program. By default, you will also receive the program files you have generated or modified during your on-site use period.

Example 8: Subject line of email for remote data processing using other software

Project number – required memory – expected running time – end date of contract – Other
fdz###, # GB, # hours/days, yyyy-mm-dd, [Other]

yyyy-mm-dd = date in year-month-day format

Examples:
fdz041, 3 gb, 14 h, 2016-01-31, Matlab
fdz041, 3 gb, 14 h, 2016-01-31, after on-site use on 2014-04-30, Matlab

8 JoSuA Manual

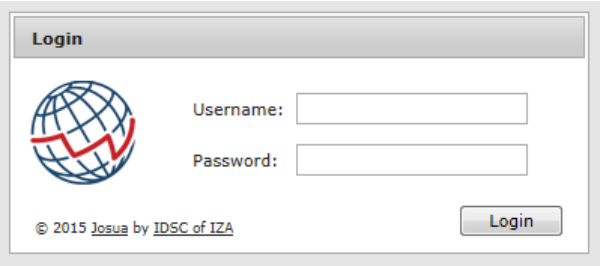
The following part describes how to navigate through the JoSuA web interface, send in remote jobs and subsequently access the results.

Please also see the FAQ to JoSuA on our website:
http://fdz.iab.de/en/FDZ_Data_Access/FDZ_Remote_Data_Access/faq.aspx

8.1 Login

- Go to www.josua.iab.de/gui/index.html.
- At the login window (see Figure 1), enter the username and the password that was provided to you by the FDZ.
- Your username and initial password is the same as the one you use during on-site use.[9]
- In case you have not received your login data yet or you forgot it, please contact the FDZ (iab.fdz@iab.de).
- Sharing the password with others is not allowed.[10]

Figure 1: Login



8.2 Start

After you have successfully logged in, the welcome page will appear. The main menu consists of several tabs: *My Jobs*, *Projects*, and *Job Submission*, which will be described in more detail below. The tab *My Jobs* is selected by default. In the right upper corner you can change your password, access the help, or log out.

8.3 My Jobs

The *My Jobs* tab (see Figure 2) provides a list of all your latest jobs and their status, i.e. whether they are in queue, running, or released. Click on the magnifier symbol to display details about the selected job.

[9] This user name and the initial password is the same as for on-site use. Any password changes in JoSuA will not be synchronised with the account for on-site use.
[10] In case your project includes other researchers, every user receives an account for his/her own.

Figure 2: My Jobs Tab

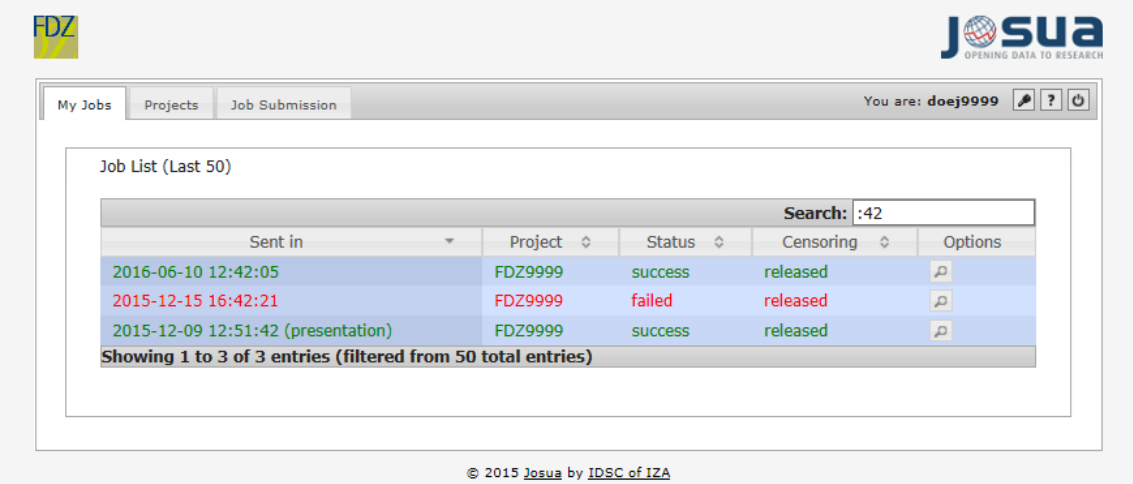
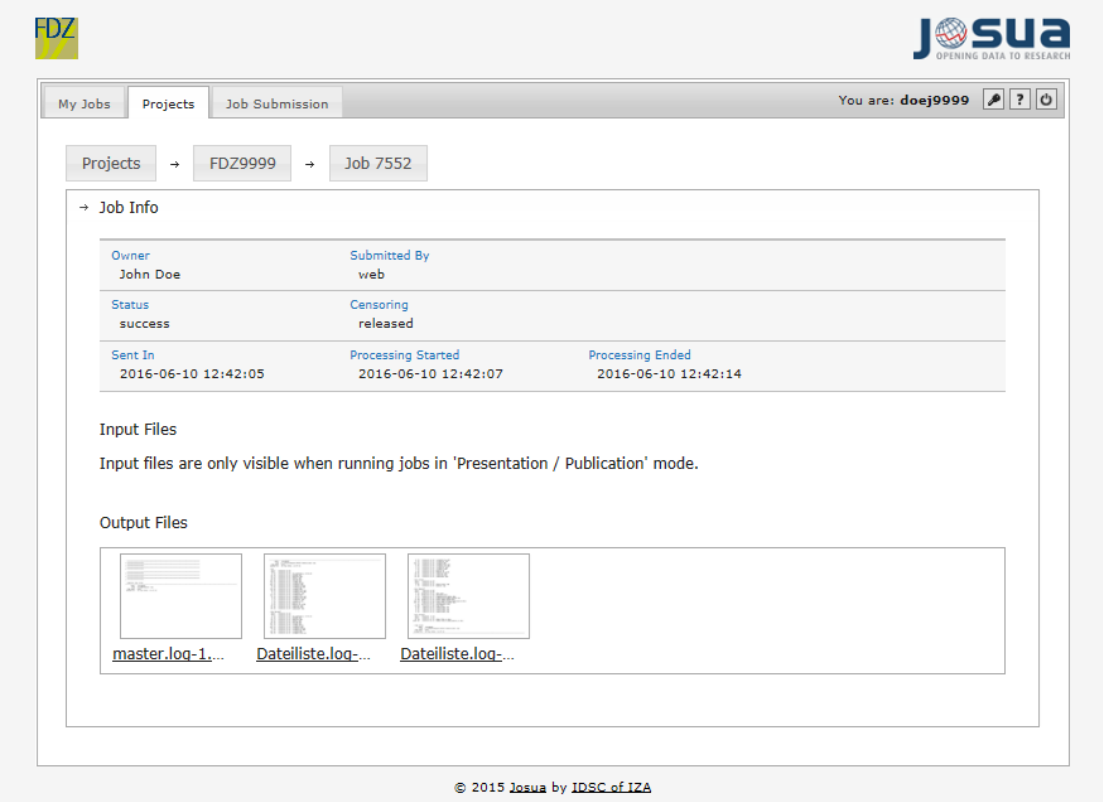


Table 1: Meaning of the fields *Status* and *Censoring* in the rider *My Jobs*

Status	Censoring	
queue *The job is currently in queue (mainly Presentation / Publication mode)*	**hold**	
running *The job is running. When the capacity rate of the servers is high, further delays can occur. When waiting for a free server, the status is also „running“. The true starting time is noted in the job details.*	**hold**	
failed *The job terminated due to an error.*	**hold** *Waiting for approval of results*	**released** *Results approved (until termination)*
success *The job was executed successfully.*	**hold** *Waiting for approval of results*	**released** *Results approved*

Figure 3: Job Details of Selected Job



Beside some general information about your job, you can find all submitted input and output files for this job in the Job Details (see Figure 3). At this time, a maximum of 20 result files can be created for each job. Log-files are automatically cut off if the size of 2 MB per file or 8 MB in total is exceeded. The presentation of the results depends on the mode in which the job was submitted.

- ***Input Files**:*
    - *Internal Use*: The download function is disabled in this mode.
    - *Presentation / Publication:* Gives a list of all do-files that were sent in for this job. You can download an input file by clicking on the file name or the download button. Please note that the link will appear only after completion of manual disclosure review.

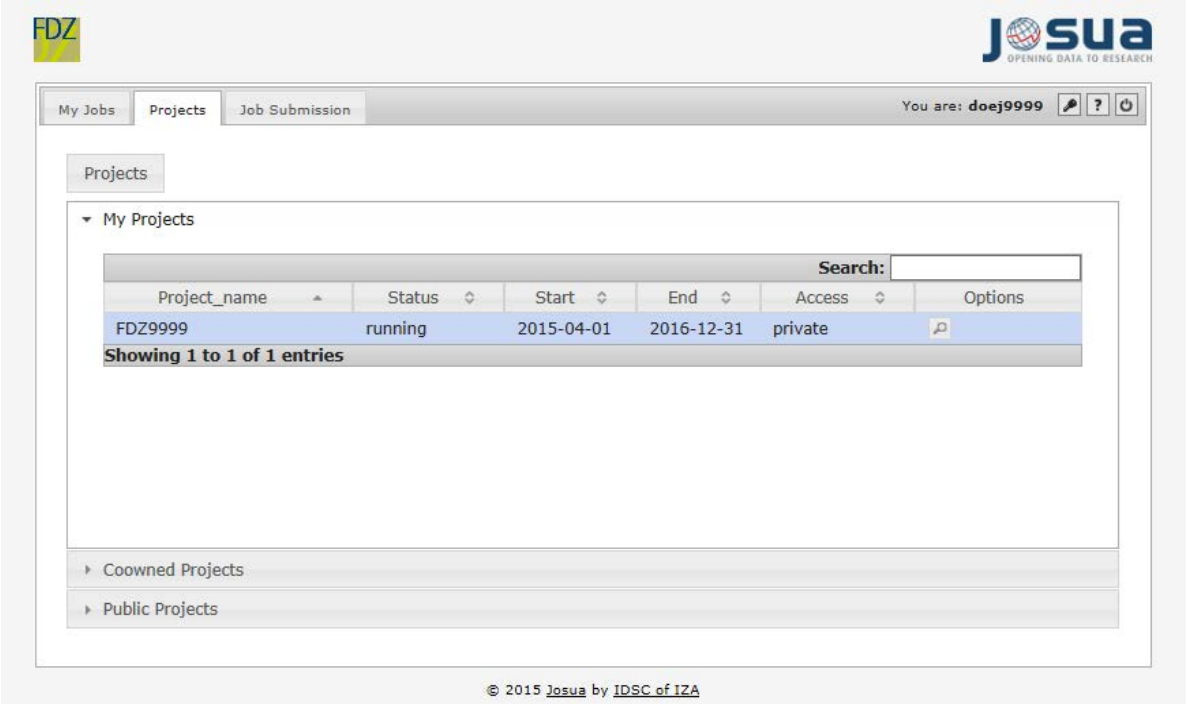
- ***Output Files***:
    - *Internal Use*: You can look at the results of your job by clicking on any of the thumbnails. Navigate through the preview window by using the arrow bars. Please note, that the results are only accessible as long as no new job in the Presentation / Publication mode is uploaded.
    - *Presentation / Publication***:** After your output has undergone manual data privacy control by the FDZ, all output files that have been accepted can be downloaded here. Please note that the link won't appear until the data privacy check is complete. To

download files click on the filename or on download and save the file on your computer.

8.4 Projects

In the *Projects* tab (see Figure 4), you can view basic information on your project, list jobs of other project members (if applicable) and upload ado-files. Clicking on *My Projects* or *Co-owned Projects*[11] will drop down the according project you are a member of. There are currently no public projects.

Figure 4: Projects Tab



A list similar to that in the picture above will drop down. Clicking on the magnifier symbol in the options column will open the Details for this project. Within this dialog, there are three dropdown menus:

- ***Project info:*** Displays all available information about the selected project (see Figure 5).
- ***Jobs:*** List of all jobs within the selected project. In addition to general information you can click on the magnifier which will provide the same options as you can find in *My Jobs*.
- ***ADO:*** Upload ado-files for your project (see Figure 6). Each user can add files that all authorized project members can use for their jobs. If there are already files that were uploaded by other project members, you can see them here. Click on the trash bin to delete files. JoSuA will prompt to confirm the deletion. To add a new file click *Add File* and browse the file on your hard disk or type the correct path, afterwards click *upload* to save your selection. Existing file names will be overwritten.

[11] My Projects are projects you are the first owner of. Co-owned projects are basically the same, but you are listed as co-owner. Both types of projects are equivalent.

Figure 5: Project Info

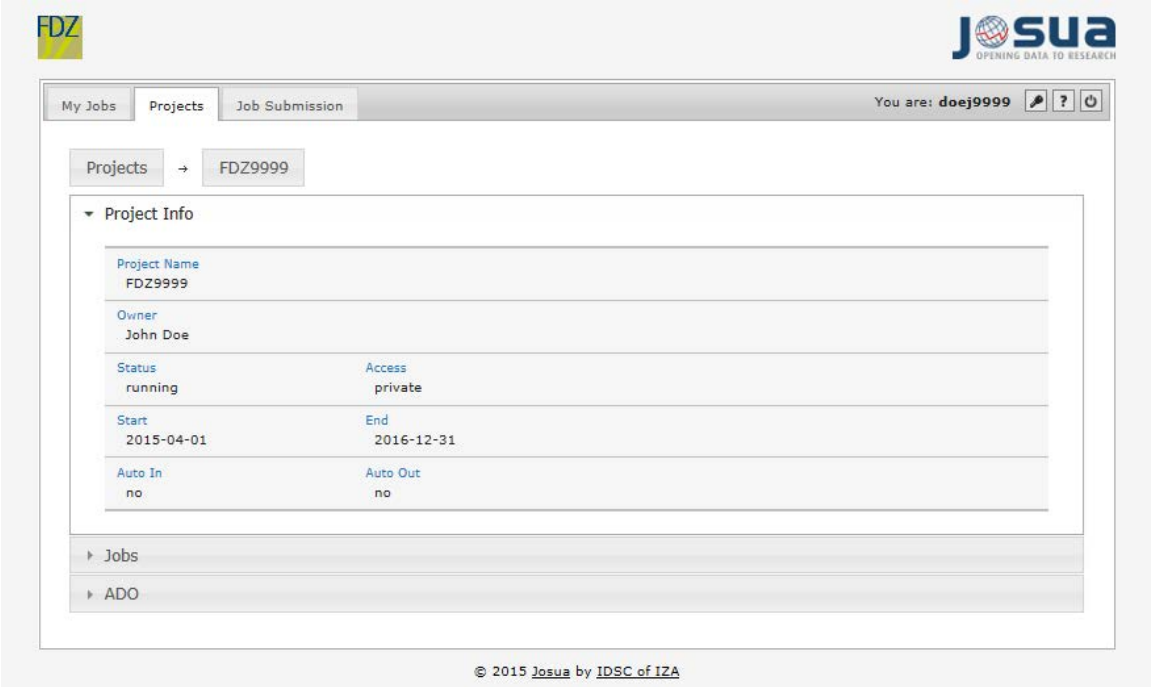
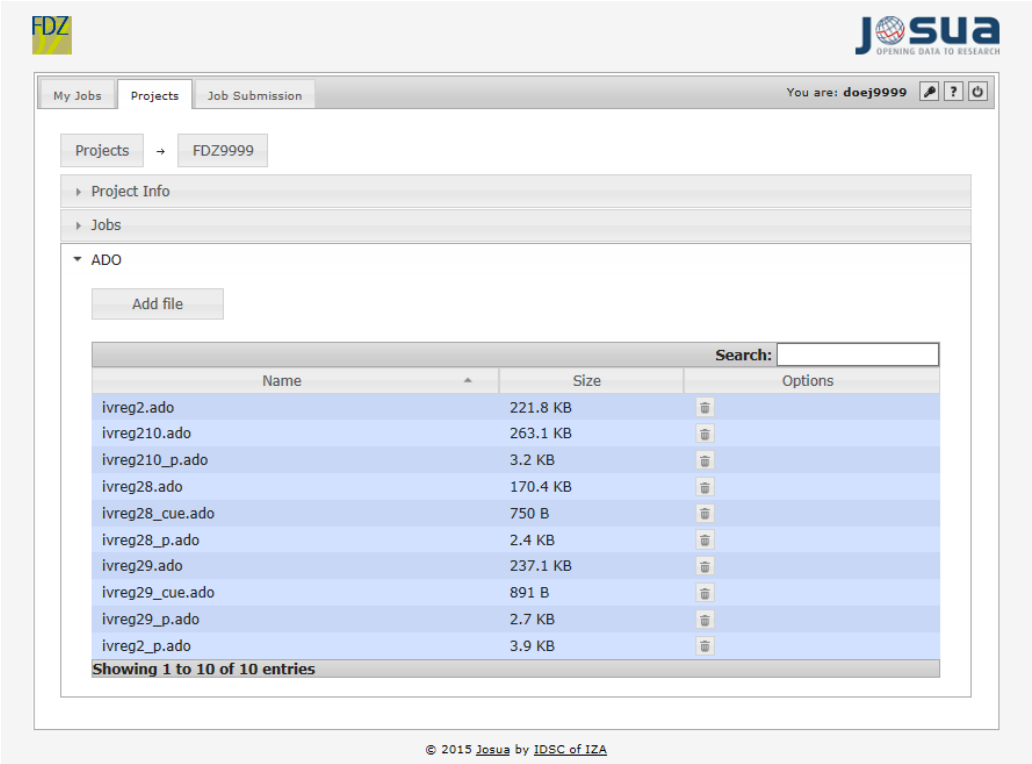
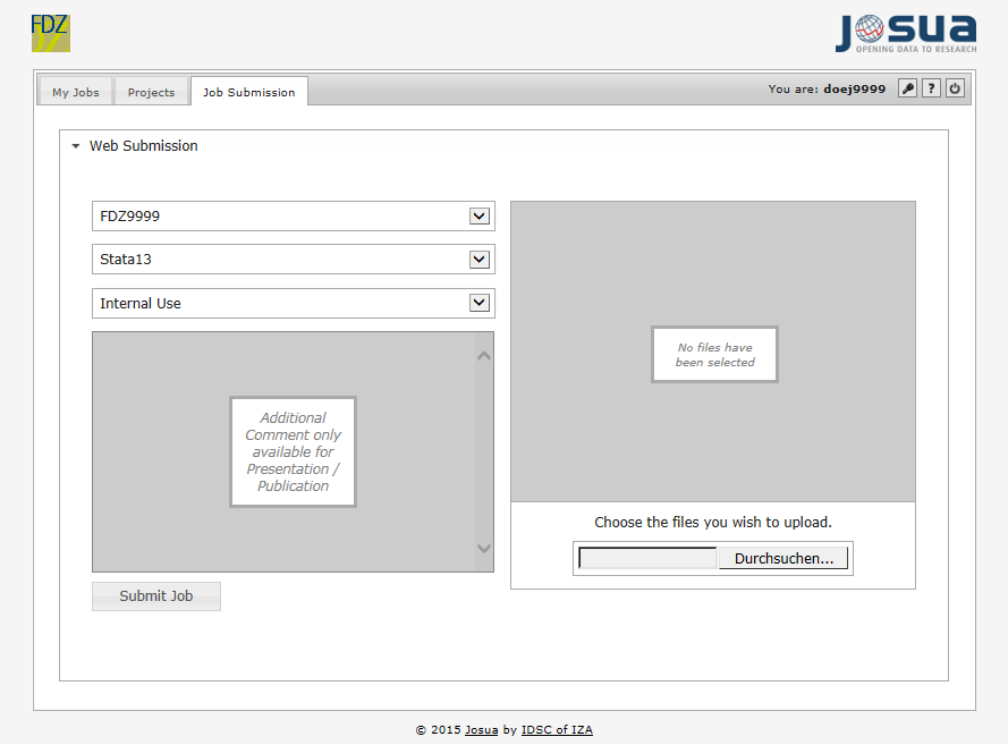


Figure 6: Ado-Files



8.5 Job Submission

Figure 7: Job Submission



You can submit jobs by uploading do-files to the *Job Submission* tab of the JoSuA web interface. All jobs you wish to submit must contain one file named **master.do**. All do-files that need to be executed have to be uploaded for each job.

The left side (see Figure 7) displays the project you currently work on (e. g., fdz9999) and the software package used. In the third drop-down menu in the left pane, you can choose whether your job will run in Internal Use or Presentation / Publication mode (see Chapter 4.3). If you use the latter, you can post an additional comment providing any relevant details or explanations. Please do not use special characters and limit your comment to 512 characters.

In the panel on the right-hand side, click the *Browse* button to search your disk for input files. Please do not upload more than 20 do-files per job. The uploaded files are listed above the Browse button, you can check the list or remove single files. Once you are finished uploading your do-files, click Submit job to send your job.

8.6 In case of job terminations with JoSuA

Should your Job terminate with the status failed, please independently check the following aspects:

- If a job terminates in Presentation / Publication mode, please wait until the job is released by the FDZ staff. Then try to figure out the problem by analysing the error message.
- Please make sure that your programs meet our Guidelines for Remote Data Access and On-Site Use, especially the path specifications.

- Check the FAQ regarding JoSuA for indications on the cause of the error and possible solutions.
- Check your Stata programs for programming errors.

If these actions do not yield a satisfactory solution, please contact the FDZ. When doing so, please provide the following information:

- Project number, job number in JoSuA
- Detailed description of the problem (e.g. exact wording of error message)

NEVER transmit results that were created in Internal Use mode or other outputs not approved by the FDZ staff. The transmission of non-approved results – even to the FDZ – is considered a breach of contract.

9 Quick Overview

Programs (do-files)

- Have to be adapted according to the guidelines of the FDZ (see Chapter 2)
- Do-files must not contain data
    - Never copy results or IDs of establishments or individuals into your do-files.
    - External aggregate data have to be sent as data sets to the FDZ (see Chapter 3.2).

Remote data processing jobs using JoSuA (Job Submission Application)

- 1 job per day
- Max. 20 do-files per job
- Prior to sending in a new job, please wait by all means until the results of any previous jobs in your project are released and / or the end of your on-site use.
- Selecting the job mode in JoSuA:

Internal Use	• Designed for all the preparatory steps of data preparation and analysis • Results must not be copied, written down, or published • Persons not covered by the project's data use agreement must not be given access to the results
Presentation / Publication	• If results are about to be published in a presentation or publication • Transfer of do-files before and after on-site use

On-site use

- Program and result files uploaded via JoSuA are stored on a different server and are not automatically transferred to the project shares.
- Do-files have to be uploaded to JoSuA in *Presentation / Publication* mode at least 3 business days prior to the beginning of your on-site use.

Data usage

- The amount of RAM memory is limited to **30 GB per user and project**. Please pay attention to your memory usage in order not to affect your current job or jobs of other users.
- The provision of external aggregate data sets in your project share will take up to 3 business days.

Results

- Disclosure review for analyses in Presentation / Publication mode may take up to 5 business days.
- Results of previous Internal Use jobs will not be accessible once a Job has been uploaded in *Presentation / Publication* mode.
- If you intend to create an aggregate data set for export, please contact us **in advance**.

If you have any questions, please contact the Research Data Centre (mailto:iab.fdz@iab.de). Please specify your project number (**fdzXXXX**) and, if applicable, the **JoSuA job ID**.